Khalid Rehman Hakeem
Hüseyin Tombuloğlu
Güzin Tombuloğlu *Editors*

Plant Omics: Trends and Applications

Springer

Plant Omics: Trends and Applications

Khalid Rehman Hakeem
Hüseyin Tombuloğlu
Güzin Tombuloğlu
Editors

Plant Omics: Trends and Applications

Springer

Editors
Khalid Rehman Hakeem
Faculty of Forestry
Universiti Putra Malaysia
Selangor, Malaysia

Hüseyin Tombuloğlu
Department of Biology
Fatih University
Buyukcekmece, Istanbul, Turkey

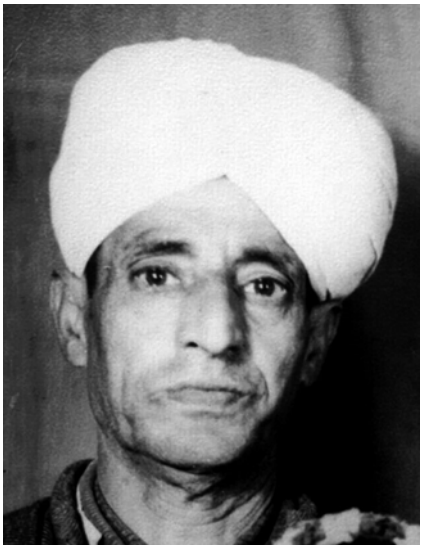
Güzin Tombuloğlu
Pathology Laboratory Techniques Program
Vocational School of Medical Sciences
Fatih University
Buyukcekmece, Istanbul, Turkey

ISBN 978-3-319-31701-4 ISBN 978-3-319-31703-8 (eBook)
DOI 10.1007/978-3-319-31703-8

Library of Congress Control Number: 2016949383

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland



(1920–2003)

To my Lovely late grandfather Hakeem Ali Muhammad (BABA) who has been my inspiration right from the beginning. May Almighty provide peace to his soul.

Khalid Rehman Hakeem

Foreword

Molecular markers revolutionized the study of living entities, being further enhanced by the in vitro amplification via polymerase chain reaction (PCR). In recent years, a new revolution has arisen, including genomics, transcriptomics, transposomics, proteomics, glycomics, lipidomics, metabolomics, and interactomics (known as -omics sciences). This has been mostly fueled by emerging new technologies, such as second- and third-generation nucleic-acid sequencing, as well as second-generation peptide-sequencing platforms, bioinformatics and statistical methodologies.

The book *Plant Omics: Trends and Applications* edited by Hakeem et al. (Springer) is an interesting and comprehensive revision about these topics. An overview of genomic analyses and resources in plants is presented by Aydin, Malik, and Afzal et al. in the Chapters 1, 2 and 11, respectively, highlighted by the so-called "next-generation" sequencing (NGS), like the second-generation nucleic-acid sequencing (SGS). The third-generation nucleic-acid sequencing (TGS) delivers even higher and faster throughput at much lower prices (the so-called $1000 and even $100 genome, referring to the cost of resequencing the human genome, which is boosting these developments for medical use). Specialized databases and bioinformatics tools to store and analyze the huge amounts of data generated by the different sequencing platforms are further described, allowing contig assembly, genome annotation, and gene prediction. These studies can be used to identify molecular markers, generate genomic maps, genotyping, evolutionary relationships, and thus generate phylogenic trees (dendrograms) in a fast and accurate way.

The current status, advantages and disadvantages, applications, and future perspectives of high-throughput sequencing via massively parallel platforms are described by Ari and Arikan in Chapter 5 and Afzal et al. in Chapter 11, including Roche 454, Applied Biosystems SOLiD, Illumina Solexa, and in situ RNA (cDNA) sequencing. The implications for plant breeding are reviewed, including development of molecular markers, high-resolution genetic maps and association mapping (AM), genome-wide association studies (GWAS), quantitative trait loci (QTL), and linkage disequilibrium (LD). Plant transcriptomics are further reviewed by Gurel et al. (barley response to drought and salinity), Candar-Cakir and Cakir (miRNA profiling), and Okay (identification of gene families using structural and functional

genomics) in Chapters 7, 8, and 9, respectively. Additionally, plant epigenetics and applications are described by Tarhan and Turgut-Karain in Chapter 10. They include DNA methylation, histone modification, and noncoding RNA (ncRNA).

Both traditional and modern QTL are reviewed by Jamil et al. in Chapter 3, including genotyping, phenotyping, mapping, and sequencing. This allows deciphering associations between genotypic and phenotypic variations in segregating populations, with the aid of molecular markers. Thus, the high-throughput sequencing (HTS) platforms allow performing genome-wide analyses with an unprecedented resolution, allowing to overcome the failures of previous approaches. These developments are a great contribution to marker-assisted and genomic-assisted breeding at an unprecedented resolution level. This way, it has been possible to improve previous biparental studies towards multiparental (population) analyses, with clear evolutionary and phylogenetic implications. Such analyses demand specialized bioinformatics and mathematical (statistical) models and tools like the Hidden Markov Model (HMM).

Gozukirmizi et al. review transposomics in plant genomes (Chapter 4). These mobile elements may take up significant amounts of plant genomes (e.g., 80 % in barley), being a keystone in plant-genome dynamics and evolution. They are involved in gene expression, being also responsible for chromosomal variations, including smaller mutations like insertions/deletions, as well as larger structural variations, such as duplications and overloading repetitions.

Molecular markers based on DNA and their applications are summarized by Karlik and Tombuloglu in Chapter 6. They include pre-PCR markers like restriction-fragment length polymorphisms (RFLP) as well as post-PCR ones like random-amplified polymorphic DNA (RAPD), simple-sequence repeats (SSR), amplified fragment length polymorphisms (AFLP), and single-nucleotide polymorphisms (SNP). Microarrays and RNA profiling (cDNA- or direct RNA-sequencing) are also considered.

Plant proteomics are reviewed in Chapters 12–15 by Shahzad et al. (overview including cell wall, cell membrane, chloroplast, mitochondrion, and nuclear proteomes), Noraida et al. (bamboo grass, including rapidly growing culms, fast-growing shoots, and sporadic flowering), Hu and Wang (abiotic-stress responses, including drought and heat stress in maize, rice, and wheat), and Xiong et al. (sex determination of dioecious plants, including a review of morphological and physiological methods, as well as the ones involving peptide and DNA markers, besides full-proteomic ones).

Chapters 16–18 deal with plant metabolomics, including the one by Imadi and Kazi (model plants like thale cress, as well as crops like cotton, barley, rice, sugarcane, *Solanum*, wheat, and maize), Turumtay et al. (methodological strategies and future prospects; combining spectrometry-based database technologies with multivariate statistical methodologies, including liquid chromatography/mass spectrometry (LC/MS), gas chromatography/mass spectrometry (GC/MS), and nuclear magnetic resonance (NMR)), and Sytar et al. (plant phenolics for food and medicinal use).

Plant glycomics are reviewed by Shahzad et al. in Chapter 19, including different analytical tools to study the cell wall, cell membrane, mitochondrion, and chloroplast. On the other hand, Afzal et al. describe plant lipidomics in Chapter 20, including the methodologies used in this scientific field and future perspectives. Finally, Shafique et al. deal with plant interactomics under salt and drought stress in rockcress, including different signaling transduction pathways responsible for the regulation of plant responses to stress and enhanced metabolism.

This work represents an updated, rigorously prepared and well-organized plant -omics revision. It is a valuable contribution for those aiming to remain updated in a wide range of -omics topics, including graduate-level students, instructors, and researchers. Furthermore, the integration of -omics technologies is a promising approach to bridge the gap between basic knowledge and applied approaches in plant research sciences.

Gabriel Dorado
Department Bioquímica y Biología Molecular
Universidad de Córdoba,
Córdoba, Spain

Turgay Unver
Biology Department
Faculty of Science
Cankiri Karatekin University, 18100
Cankiri, Turkey

Pilar Hernandez
Instituto de Agricultura Sostenible (IAS-CSIC)
Consejo Superior de Investigaciones Científicas,
Córdoba, Spain

Preface

To understand the organizational principle of cellular functions at different levels, an integrative approach with large-scale experiments, the so-called "omics" data, is needed. In recent years, Omical biotechnologies utilized in plant sciences include genomics, transcriptomics, transposomics, proteomics, glycomics, lipidomics, metabolomics, fluxosomics, and interactomics. These technologies have provided new insights into all the aspects of life sciences, including plant science. Omics is in fact providing a snapshot of the biological functioning of an organism. Plant Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of plants. Currently, omics is an essential tool to understand the molecular systems that underlie various plant functions. Furthermore, in several plant species, the development of omics resources has progressed to address particular biological properties of individual species. Integration of knowledge from omics-based research is an emerging issue as researchers seek to identify significance, gain biological insights, and promote translational research. From these perspectives, the current volume intends to provide the emerging aspects of plant systems research based on omics and bioinformatics analyses together with their associated resources and technological advances.

The present volume highlights the working solutions as well as open problems and future challenges in plant omics studies. Demonstrating the diversity of omics, we believe that this book will initiate and introduce readers to state-of-the-art developments and trends in omics-driven research.

This is our opportunity to thank the authors who have given their time unselfishly to meet the deadlines for each chapter. We greatly appreciate their commitment. We are also thankful to Prof. Gabriel Dorado (Spain), Prof. Turgay Unver

(Turkey), and Prof. Pilar Hernandez (Spain) for their suggestions and writing the foreword for this volume.

On behalf of the editorial team, I thank Springer-International team for their generous cooperation at every stage of the book production.

Selangor, Malaysia — Khalid Rehman Hakeem
Buyukcekmece, Istanbul, Turkey — Hüseyin Tombuloğlu
Buyukcekmece, Istanbul, Turkey — Güzin Tombuloğlu

Contents

About the Editors

Khalid Rehman Hakeem is working as a Fellow Researcher at the Faculty of Forestry, Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia, and also a Visiting Professor at Fatih University, Istanbul, Turkey. He has obtained his MSc. (Environmental Botany) as well as PhD (Botany) from Jamia Hamdard, New Delhi, India, in 2006 and 2011 respectively. He has conducted his postdoctoral research in the fields of forest dynamics and plant biotechnological studies from Universiti Putra Malaysia from 2012 to 2013. Dr. Hakeem has more than 9 years of teaching and research experience in Plant Eco-Physiology, Biotechnology and Molecular biology, Plant-Microbe-soil interactions as well as in Environmental sciences. Recipient of several fellowships at both national and international levels, Dr. Hakeem has so far edited and authored more than 20 books with international publishers. He has also to his credit more than 85 research publications in peer-reviewed international journals, including 30 book chapters with international publishers. He is also the editorial board member and reviewer of several high-impact international Journals. Dr. Hakeem is currently engaged in studying the plant processes at ecophysiological as well as proteomic levels.

Güzin Tombuloğlu is working as an Assistant Professor at the Pathology Laboratory Techniques Programme of Vocational School of Medical Sciences, Fatih University, Istanbul, Turkey. She has received her MSc. (Biology) degree in 2008 and PhD (Biotechnology) degree in 2014 from Fatih University. She has studied transcriptomics identification of barley boron tolerance mechanism during her PhD. She has conducted several projects on abiotic stress, plant stress responses, boron toxicity, and transcriptomics. She has 10 years experience in teaching molecular biology. She also worked as a Chairman in Pathology Laboratory Techniques Programme and Assistant Manager at Vocational School of Medical Sciences at Fatih University.

Hüseyin Tombuloğlu is working as an Assistant Professor at the Faculty of Science and Arts, Department of Biology, Fatih University, Istanbul, Turkey. He received his BSc. degree in 2007 from Istanbul University, Department of Molecular Biology and Genetics, Turkey, and also he studied as an exchange student in the University of Groningen, the Netherlands. He obtained his MSc. (Biology) degree in 2010 and PhD (Biotechnology) degree in 2014 from Fatih University. During this period, he worked on molecular biology of plants, specifically abiotic stresses and the long distance communication of plants via miRNAs. He has been awarded several projects supported by TUBITAK (The Scientific and Technological Research Council of Turkey) and BAP (Research Fund of Fatih University). Dr.Tombuloğlu has more than 8 years of teaching and research experience in Genetics, Molecular Genetics, Plant Physiology and Biotechnology, as well as Bioinformatics. His current research is focused on the organ-to-organ communication, boron stress, and proteomics.

Genome Analysis of Plants

Gülsüm Aydin

Contents

Abstract Genomics, emerged in the 1990s as a revolutionary approach, studies the structure and function of all the genes in an organism. Genome size studies, physical mapping, and genetic mapping applications were developed for characterizing and comparing genomes prior to the advent of high-throughput next-generation sequencing (NGS) technologies. Arrival of NGS techniques have redirected attention away from these older methods and made it possible to sequence, assemble, and analyze the genomes of many plant species. The release of the first plant genome sequence belonging to *Arabidopsis*, in 2000, brought new insights and perspectives into our

G. Aydin (✉)
Department of Biology, Faculty of Science, Selcuk University, 42130 Konya, Turkey
e-mail: gkalemtas@gmail.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_1

understanding of plant genomics. Rapid progress has since been made and not only model organisms but also a variety of species of ecological, agricultural, or economical importance has been sequenced generating a huge amount of data. These data are publicly available through web portals, e.g., the Ensembl Plants portal (http://plants.ensembl.org/index.html) and the NCBI genome portal (http://www.ncbi.nlm.nih.gov/genome/). However, the unprocessed sequence data are not very informative and they have to be annotated both at the structural level (identification of genes) and at the functional level (identification of gene function). Owing to the high cost and time required for manual genome annotation, genomes are generally annotated via automated gene prediction programs most of which are listed at the geneprediction.org web site (http://www.geneprediction.org/software.html). Annotation data obtained by this way may be utilized for both basic and applied research so that it helps to elucidate evolutionary relationships and develop better phylogenetic classification. Sequences of crop plants may aid in identification of economically important genes which in turn may help biologists to provide food, fiber, and fuel for the exponentially increasing population. As more whole genome sequences become available, it will increase the speed and lower the costs for studies regarding epigenomes, transcriptomes, and metabolomes.

Keywords Plant genomics • Next-generation sequencing • Genome annotation • Gene prediction • Evolutionary relationships • Phylogeny

1 Introduction

Organization of genes and genetic information within the genome, the methods utilized for collecting and analyzing this information and determination of the effect of this organization on biological functionality of the genes constitute fundamentals of genomics. The advent of high throughput NGS technologies have made it possible to sequence, assemble, and analyze the genomes of numerous plant species (Flagel and Blackman 2012). Enormous amount of sequence data collected at databases have necessitated annotation of genomes via automated gene prediction programs. Two basic steps in genome annotation are structural and functional annotation. Computational approaches to structural annotation (gene identification) can be broadly classified into three main categories: *ab initio* methods (intrinsic methods), homology-based/similarity-based methods (extrinsic methods), and integrated methods (Davuluri and Zhang 2003; Thibaud-Nissen et al. 2008; Goel et al. 2013). Once the gene is identified via one of these methods, the next step is assignment of a putative function to the predicted gene (functional annotation). Alignment of predicted protein sequence against a protein database is a common way of attributing a function to a protein. When there is no hit above a given threshold or no

well-characterized hit is determined in the database, looking for conserved domains lying in the gene models may help to assign a function to the predicted protein as well (Ouyang et al. 2009). Although there are a number of tools that perfectly assign gene structures and functions to these genes, it is just a prediction and still subject to a degree of uncertainty. Therefore, a predicted gene or predicted protein function needs to be supported with direct experimental data to reduce the risk of disagreement between biological function and annotation (Thibaud-Nissen et al. 2008; Dale et al. 2012).

The availability of whole genome sequence data provides a deep understanding of the molecular and cellular function of genes. It can also be utilized for gene-targeted mutational forward genetics, sequence-based marker development, and microarray platform design for gene expression studies (Springer et al. 2009; Flagel and Blackman 2012). These tools may be utilized for molecular breeding and identification of economically important genes. Providing food, fiber, and fuel for the exponentially increasing population is a challenge for plant biologists in the twenty-first century. Therefore, the use of sequence data for molecular breeding and identification of economically important genes is an essential step towards the solution of this global issue. Genome sequence data can also provide insight into evolutionary relationships among organisms or genes (Snel et al. 2005). Comparative evolutionary genomics emerged as a powerful tool to study evolutionary changes among organisms and to identify the genes that are conserved among species. Elucidation of evolutionary dynamics of genes and genomes is also helpful in understanding disease susceptibility (Das and Hirano 2012).

In the present chapter, I attempt to take a practical look at the computational tools utilized for analysis of whole genome sequence data. I also address how generation of NGS technologies switched the molecular analysis of plants from a single gene to the whole genome. The new generation of comparative genomics as a consequence of rapid accumulation of sequence data and how it offers a powerful aid to study evolutionary relationships among organisms are also discussed.

2 The Genetic Structure of Plant Genomes

Genomics is a discipline in genetics that studies the organization of genes and genetic information within the genome, the methods utilized for collecting and analyzing this information and determination of the effect of this organization on biological functionality of the genes. Genome size, gene content, extent of repetitive sequences, and polyploidy/duplication events are the most remarkable features of plant genomes. Plants carry mitochondrial and chloroplast genomes besides nuclear genome which is the largest and most complex (Campos-De Quiroz 2002). The size of nuclear genome varies over nearly 2000-fold, from 63 Mbp for *Genlisea margaretae* (Greilhuber et al. 2006) to 125 Gbp for *Fritillaria assyriaca* (Bennett and Smith 1991). Table 1 reveals genome sizes of a number of important plant species (Arumuganathan and Earle 1991). Although genome size is not closely associated with organism complexity (*C*-value paradox), the

Table 1 Genome sizes of selected plants

Scientific name	Common name	Haploid size (Mb)
Arabidopsis thaliana	Thale cress	125
Oryza sativa	Rice	424
Vitis vinifera	Grapevine	483
Sorghum bicolor	Sorghum	748
Lycopersicum esculentum	Tomato	907
Glycine max	Soybean	1115
Brassica napus	Rapeseed	1200
Zea mays	Corn	2292
Hordeum vulgare	Barley	4873
Triticum aestivum	Wheat	16,000

genomes of more complex organisms tend to be larger compared to the genomes of less complex organisms (Vinogradov 2004). Most of the time, the variation in genome size is not related with differences in gene size or gene number. Research has shown that plants exhibit extensive conservation of both gene content and gene order and that different plant species generally use homologous genes for identical functions (Bennetzen 2000; Bennetzen et al. 2005). Differences in genome size can mainly be attributed to the repeated DNA content and the ploidy level. Polyploidy is a rapid event that can double genome size in a single generation, and most plants are either current polyploids or have a polyploidy origin. However, plant geneticists have shown that the most significant contributor to genome size is repetitive DNA sequences. These sequences may be organized in tandem arrays or they may show a dispersed distribution in the genome. Retrotransposons with long terminal repeats (LTRs) are involved in the latter category and comprise most of the repetitive sequences in plant genomes. They constitute only 10 % of the small *Arabidopsis* genome, whereas they account for at least 60–80 % of the 20-fold larger maize genome (Schmidt 2002). LTR-retrotransposons are often related with the large heterochromatic regions flanking functional centromeres. In plant species with large genomes such as maize and barley, many of the LTR-retrotransposons are intermixed with genes, usually as nested structures. On the other hand, in plant species with small genomes such as *Arabidopsis*, rice, and sorghum the genic regions frequently have only single LTR-retrotransposons inserted in or near genes (Bennetzen et al. 2005; Lee and Kim 2014).

2.1 Genetic Maps

A genetic map (linkage map) shows the order of molecular markers throughout chromosomes as well as the genetic distances, usually expressed in terms of centiMorgans (cM), existing between neighboring molecular markers. Genetic maps help to understand the organization of plant genomes and once in hand, they aid in the development of plant breeding applications such as the identification of

Quantitative Trait Loci (QTL) and Marker Assisted Selection (MAS) (Campos-De Quiroz 2002). QTL analysis enables identification of the loci responsible for variation in complex, quantitative traits. Determination of the genes regulating these traits and revealing the function of these genes is often the actual goals of QTL analysis. For example, identifying loci responsible for improvement of crop yield or quality and then assembling the favorable alleles in elite lines comprise the basis of breeding projects (Borevitz and Chory 2004). The most prominent feature of MAS is that it facilitates indirect selection for an allele responsible for a certain phenotype, once a molecular marker genetically linked to the expression of that allele has been detected. Thus presence of the molecular marker will always be related with the existence of the allele of interest. Genetic maps also aid in establishment of the extent of duplication and genome colinearity between different species (Campos-De Quiroz 2002). Moreover genetic maps may be used for plant gene isolation through positional cloning, once the genetic position of any mutation is developed (Campos-De Quiroz et al. 2000). Eventually, advances in DNA sequencing facilitated direct sequence-based genetic mapping. The single-nucleotide polymorphism (SNP) markers are much more numerous compared to other markers enabling generation of extremely dense genetic maps. For this reason SNP has become the molecular marker of choice and SNPs have ensured depth sufficient for high-quality mapping of QTL and association mapping studies (Duran et al. 2010).

2.2 Physical Maps

Genetic maps provide markers along chromosomes. However, there are often vast spaces between markers to provide an entry point into genes. The kilobases per centiMorgan (kb/cM) ratio is large even in model plants. For example, it is 120–250 kb/cM in *Arabidopsis* and 500–1500 kb/cM in corn. Accordingly, a 1 cM interval may harbor ~30–100 or even more genes. Physical maps are utilized to bridge such gaps representing the entire DNA fragment located between neighboring molecular markers. Physical maps can be defined as a set of relatively large pieces of partially overlapping DNA encompassing a given chromosome (Campos-De Quiroz 2002). Although first-generation physical maps were based on yeast artificial chromosomes (YACs), chimerism and stability issues led to introduction of bacterial artificial chromosomes (BACs) as alternatives to YACs (Shizuya et al. 1992). Despite YACs can carry pieces of insert DNA up to 3 Mb, approximately ten times longer compared to BAC inserts (up to 350 kb), lack of chimerism and the simplicity of BAC manipulation have made BACs the vector of choice for physical mapping (Peterson et al. 2000). Physical mapping was assumed a convenient way of assembling a genome in a way that would enable eventual complete sequencing. The first eukaryotic genomes were sequenced using a physical mapping approach (Peterson 2014).

Investigation methods such as genome size studies, physical mapping, and genetic mapping were developed for characterizing and comparing genomes and

they were utilized in the validation, correction, and exploitation of DNA sequence data prior to the advent of high throughput NGS technologies (Peterson 2014). Arrival of these postgenomics techniques have redirected attention away from these older methods and made it possible to sequence, assemble, and analyze the genomes of many plant species. Therefore these technologies have switched the molecular analysis of plants from a single gene to the whole genome (Flagel and Blackman 2012). The information gathered through analysis of whole genome sequence data can be applied to determine gene function and regulation, which will obtain access to all genes of an organism. It can also be utilized to analyze evolutionary relationships among organisms and will enable a systematic understanding of genome organization and plant biology (Soneji et al. 2010).

3 Plant Genome Annotation

Genome sequence information allows a better understanding of the way genes are organized within the genome and the way they influence each other to identify biological functions. Analysis of this information for the whole genome constitutes the basis of genome analysis. The improvement in genome analysis aided by automation and various software tools has expedited the whole genome sequencing in all organisms as well as plants. Genome sequences for a high number of plant species, especially those with small genomes and well-defined genetic resources such as *Arabidopsis*, *Poplar*, *Sorghum*, rice, and grape are available and sequencing for many species is in progress or planned in the near future (Thibaud-Nissen et al. 2008; Parida and Mohapatra 2010). Recently completed plant genome projects include; sugar beet (*Beta vulgaris*) (Dohm et al. 2014), tomato (*Solanum lycopersicum*) (The Tomato Genome Consortium 2012), eggplant (*Solanum melongena* L.) (Hirakawa et al. 2014), coffee (*Coffea canephora*) (Denoeud et al. 2014), peach (*Prunus persica*) (The International Peach Genome Initiative 2013), chickpea (*Cicer arietinum*) (Varshney et al. 2013), common bean (*Phaseolus vulgaris*) (Schmutz et al. 2014), cotton (*Gossypium raimondii*) (Li et al. 2015), sweet orange (*Citrus sinensis*) (Wu et al. 2014), orchid (*Phalaenopsis equestris*) (Cai et al. 2015), banana (*Musa acuminata*) (D'Hont et al. 2012), barley (*Hordeum vulgare*) (The International Barley Genome Sequencing Consortium 2012), Norway spruce (*Picea abies*) (Nystedt et al. 2013), and loblolly pine (*Pinus taeda* L.) (Neale et al. 2014). Obtaining the basic information of crop genomes is significant for accelerating breeding pipelines and for a better understanding of the molecular basis of agronomically important traits, such as yield and tolerance to abiotic and biotic stresses. Wheat (*Triticum aestivum*), the staple food for 30 % of the human population, is a hexaploid species ($6x = 2n = 42$, AABBDD) that originates from multiple hybridizations between three different progenitor species (comprising the subgenomes: A, B, and D). The hybridization events resulted in a large and highly redundant genome and complicated the generation of a complete and properly ordered reference genome sequence for bread wheat (Eversole et al. 2014). The International Wheat

Genome Sequencing Consortium (IWGSC) adopted a chromosome by chromosome strategy to circumvent this complexity. On 18 July 2014, the IWGSC published a draft sequence of the bread wheat genome in a special issue of the international journal Science (The International Wheat Genome Sequencing Consortium 2014). In this special issue, three other research articles were published presenting major advances toward obtaining a reference sequence and providing new insight into the structure, organization, and evolution of the bread wheat (Choulet et al. 2014; Marcussen et al. 2014; Pfeifer et al. 2014).

3.1 Plant Genome Databases

With the rapid development of NGS technologies, enormous amount of sequence and annotation data has been generated and collected in the genome databases. These data are publicly available through web portals, such as: the Ensembl Plants portal (http://plants.ensembl.org/index.html) and the NCBI genome portal (http://www.ncbi.nlm.nih.gov/genome/). As genome browsers integrate genome sequence data with annotation data, they provide an exclusive platform for molecular biologists to search, browse, retrieve, and analyze the genomic data effectively and conveniently. The graphical interface of genome browsers help researchers to extract and summarize information from vast amount of raw data. Two types of web-based genome browsers are available: (1) the multiple-species genome browsers and (2) the species-specific genome browsers. Table 2 lists several major web-based plant genome browsers accessed by a large number of users worldwide. The multiple-species genome browsers integrate sequence and annotation data for many organisms and support cross-species comparative analysis. Most of these browsers provide annotations, regarding gene model, expression profiles, transcript evidence,

Table 2 List of major web-based plant genome browsers

Resource	URL
Multiple-species genome browsers	
NCBI Map Viewer	http://www.ncbi.nlm.nih.gov/mapview/
Ensembl Plants	http://plants.ensembl.org/index.html
Phytozome	http://www.phytozome.net/
VISTA	http://pipeline.lbl.gov/cgi-bin/gateway2
PlantGDB	http://www.plantgdb.org/prj/GenomeBrowser/
Species-specific genome browsers	
TAIR	http://www.arabidopsis.org
Gramene	http://www.gramene.org
SGN	http://solgenomics.net/genomes
Rice Genome	http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/
MaizeDB	http://www.maizegdb.org/

regulatory data, etc. On the other hand, the species-specific genome browsers (Table 2) generally focus on one model organism and may provide more annotation data for a particular species (Wang et al. 2013). The *Arabidopsis* Information Resource (TAIR) (http://www.arabidopsis.org) is one of the most widely used species-specific database that provides genetic and molecular biology data for *Arabidopsis thaliana*. Being the first plant to be completely sequenced (The *Arabidopsis* Genome Initiative 2000) it served as a model organism in the last 40 years for gene discovery studies and accepted as a reference point for investigation of other species' genomes (Katam et al. 2010).

3.2 Repeat Masking

In general, repeat identification and masking is the first step in genome annotation. Plant genomes can be very repeat rich; for example, 90 % of the wheat genome is thought to consist of repeats (Gill et al. 2004), and they account for ~60–80 % of the maize genome (Schmidt 2002). Repetitive sequences (SINEs, LINEs, etc.) and low-complexity sequences such as homopolymeric runs of nucleotides complicate genome annotation. These sequences need to be masked before a sequence similarity search to exclude statistically significant but biologically uninteresting matches. The process of 'masking' involves transforming every nucleotide identified as a repeat to an 'N' or to a lower case a, t, g, or c. This step constitutes a signal for downstream sequence alignment and gene prediction tools that these DNA segments are repeats. Prior to masking, the repeated sequences should be accurately identified. However, identification of repeats is complicated by the poor conservation of these sequences and accurate repeat detection usually requires generation of a repeat library for the genome of interest (Yandell and Ence 2012). Either homology-based tools (Buisine et al. 2008; Han and Wessler 2010) or de novo tools (Price et al. 2005; Morgulis et al. 2006) can be utilized to create these libraries. Highly conserved protein-coding genes, such as tubulins and histones may be identified by de novo tools, as well as transposon sequences. Therefore it is important for the users to carefully post-process the outputs of these tools and to remove protein-coding sequences (Yandell and Ence 2012).

After it has been generated, a repeat library can be utilized in conjunction with a tool such as RepeatMasker (http://www.repeatmasker.org). RepeatMasker, an efficient tool in masking both low complexity and interspersed repeats, makes use of custom libraries of repeats and supports several eukaryotic repeat databases from Repbase (Jurka et al. 2005). Failure to mask genome sequences may give rise to millions of spurious BLAST (Basic Local Alignment Search Tool) alignments which will create false evidence for gene annotations. Another issue when repeats are left unmasked is insertion of segments of transposon open reading frames (ORFs) as additional exons to gene predictions due to the fact that many transposon ORFs look like true host genes to gene predictors. Such an error would completely

corrupt the final gene annotations. Therefore good repeat masking is an important issue for the accurate annotation of protein-coding genes (Yandell and Ence 2012).

The release of the first plant genome sequence belonging to *Arabidopsis*, in 2000, marked the beginning of the plant genomics era (The *Arabidopsis* Genome Initiative 2000). Since then there has been striking progress in the area of plant genomics. Huge amount of data is generated via NGS technologies. However, it is not very informative and has to be interpreted through annotation of the functional elements of the genome. Annotation means obtaining biological information from raw sequence data and it can be divided into structural and functional annotation. Structural annotation is identification of the genes and determination of their structure and it is highly dependent on specific computational programs and availability of transcribed sequences. Functional annotation is determination of the physiological, biochemical, and biological role of the protein/RNA encoded by a gene, and it is reliant on sequence similarity to other known genes or proteins (Thibaud-Nissen et al. 2008).

3.3 Structural Annotation

The ultimate aim of gene prediction is determination of protein-coding genes, non-protein coding genes (RNA genes) and regulatory regions in genomic DNA. Although identification of RNA genes and regulatory regions (promoters) are of great importance due to their functional roles in plant genomes, I will concentrate on protein-coding genes owing to the scope of this chapter. Prior to NGS technologies, experiments were carried out at the bench on single DNA clones for identification of individual genes. Nowadays the rapid rate at which sequence data accumulates has necessitated the use of bioinformatics tools for gene identification (Goel et al. 2013). A great number of gene prediction programs are available for prokaryotic and eukaryotic organisms some of which are listed at the geneprediction.org web site (http://www.geneprediction.org/software.html). Eukaryotic genomes are generally larger than that of the prokaryotes and the gene density is usually lower. In eukaryotes, genes consist of coding segments (exons) which are interrupted by long noncoding segments (introns) (Sleator 2010). Moreover, the coding sequences are subject to alternative-splicing which is a process of joining exons in different ways during RNA splicing (Schellenberg et al. 2008). These common features of eukaryotic genomes render gene prediction in plant genomes rather difficult compared to prokaryotic genomes (Primrose and Twyman 2003; Wang et al. 2004). Although prokaryote gene prediction can be complicated by overlapping regions which make determination of translation start sites difficult (Palleja et al. 2008), it is relatively straightforward due to the absence of introns and higher gene density (Wang et al. 2004). There are two distinct aspects of current gene prediction programs: the first is the type of information utilized by the program and the second is the algorithm that is employed by these programs to combine that information into an accurate prediction (Sleator 2010). Computational approaches to gene identification can be

broadly classified into three main categories: *ab initio* methods (intrinsic methods), homology-based/similarity-based methods (extrinsic methods), and integrated methods (Davuluri and Zhang 2003; Thibaud-Nissen et al. 2008; Goel et al. 2013).

3.3.1 *Ab Initio* Methods

As gene finders first became available in the 1990s, they improved genome analyses since they enabled rapid identification of genes in assembled DNA sequences. These tools are generally called *ab initio* gene predictors because they utilize computational methods rather than external evidence (such as EST and protein alignments) to determine gene location and structure (Yandell and Ence 2012). *Ab initio* gene prediction rely on statistical and computational methods to determine gene-specific features such as core promoters (e.g., TATA-box), splice sites, polyadenylation sites, start and stop codons, exons and introns (Ouyang et al. 2009). These functional sites are called signals and methods utilized to identify them are signal sensors. The variation in base composition between coding and noncoding DNA plays a significant role in gene prediction as well as the feature-dependent methods. The type of sensors which exploit innate characteristics of the DNA sequence itself to determine whether the sequence is coding or noncoding, is called intrinsic content sensors. Although there are a high number of base composition parameters in coding and noncoding DNA, hexamer base composition (hexamer usage) gives the best discrimination. In addition to hexamer usage; nucleotide composition, codon usage, GC content, and base occurrence periodicity are useful intrinsic content sensors (Mathe et al. 2002; Goel et al. 2013). A great number of *ab initio* gene predictors consist of several different specific sensors that are usually integrated together by Hidden Markov Models (HMM). HMM is a statistical technique that has been invaluable in determination of protein-coding sequences, and in identification of intron–exon boundaries. A Markov model, defines the probability of appearance of a given base (A, T, G, or C) at a given position, when this probability depends on the appearance of one or more of the previous nucleotides (Mathe et al. 2002; Dale et al. 2012). *Ab initio* gene prediction programs are extensively used in automated genome annotation due to their speed and requirement of little computational effort. On the other hand they have limitations: specificity and sensitivity of some gene finders are over 90 % at the nucleotide level, but it is much lower at the gene level. Moreover most gene predictors are not feasible for complicated gene structures and nonconventional biological signals such as (1) long introns, (2) noncanonical introns, (3) alternative splicing, (4) overlapping genes, (5) nested genes, (6) frameshift errors, and (7) introns in untranslated regions (Ouyang et al. 2009). Another issue is training; *ab initio* gene finders utilize organism-specific genomic traits, namely codon frequencies and dispersion of intron–exon lengths, to separate genes from intergenic segments and to identify intron–exon structures. Most gene predictors are provided together with precalculated parameter files which include such information for a number of widely studied genomes, such as *A. thaliana* and *O. sativa*. Even closely related organisms can vary in terms of intron lengths, codon

usage, and GC content. Therefore the gene predictor needs to be trained for the genome of interest unless it is intimately related to an organism for which precompiled parameter files are available (Korf 2004). Some popular gene predictors can be trained by aligning ESTs, RNA sequences, and protein sequences to a genome even when pre-existing reference gene models are not available. However, it generally requires the user to have some basic programming skills (Yandell and Ence 2012).

GenemarkHMM (Lukashin and Borodovsky 1998), GlimmerHMM (Majoros et al. 2004), and Augustus (Stanke and Waack 2003) are *ab initio* gene prediction programs that are widely used for plants.

3.3.2 Homology-Based Methods

Homology-based methods have usually been called extrinsic in opposition to others that rely on some intrinsic properties (compositional bias, GC content, codon usage, etc.) of the coding/noncoding sequences. Experimentally derived transcripts (in the form of ESTs and full-length cDNAs) are important and comprehensive sources of evidence for structural annotation of gene models. Utilization of homology searching programs to compare genomic sequence data to gene, cDNA, EST, and protein sequences already present in databases is a simple way of identifying a gene within a genome (Mathe et al. 2002; Primrose and Twyman 2003). The numbers of ESTs and cDNAs vary significantly depending on the species. For maize there are over 1.7 million ESTs and there are ~1 million for wheat. Since ESTs and cDNAs are single-pass sequences their accuracy is low and they are highly redundant. Although these features of ESTs and cDNAs limit their use, it can be resolved through minimization of these sequence sets into a set of assemblies that represent all of the transcripts and in which sequencing errors are reduced by production of consensus sequences (Ouyang et al. 2009). Moreover ESTs are originated from the 3′ends of poly(A)$^+$ transcripts and contain 3′ untranslated sequences. Therefore they cannot be expected to determine all coding exons. In some cases ESTs can be originated from processed pseudogenes or unprocessed intronic sequences and they are not reliable indicators of a gene or a mature mRNA (Primrose and Twyman 2003).

The most widely used programs for determination of similar nucleotide sequences in the databases to the query sequence are the BLAST family (Davuluri and Zhang 2003). BLASTN algorithm searches a nucleotide database using a nucleotide sequence, BLASTX translates a nucleotide query into all six frames (three possible reading frames on each strand of a DNA molecule) and searches a protein database, and BLASTP searches a protein database using a protein sequence. MegaBLAST is a better choice for identifying the input query and searching with large genomic query (ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf). BLASTN is generally utilized to find out similar sequences from the database, and usually it is hard to identify the exon boundaries. After finding a cDNA or EST match to the query sequence, spliced alignment programs can be used to efficiently align an EST or cDNA with the genomic sequence (Davuluri and Zhang 2003).

Earlier alignment tools such as AAT (Huang et al. 1997) and EST_GENOME (Mott 1997) were too slow and compute intensive for the size and scope of most plant genomes. Later on, faster and more accurate alignment tools including sim4 (Florea et al. 1998), BLAT (Kent 2002), GeneSeqer (Usuka et al. 2000) and GMAP (Wu and Watanabe 2005) were developed. Although these tools has improved the quality of spliced alignments, issues remain relating to errors in EST sequences, correctly aligning small exons, incorporating nonconsensus splice sites and discriminating paralogous alignments (Thibaud-Nissen et al. 2008).

In addition to cDNA and ESTs, protein sequences present in databases may be compared to genomic sequences for identification of probable protein coding regions. Getting information from protein alignments is especially important for genes in which the number of available ESTs or cDNAs is low. Protein searches enable comparison against diverged species due to the fact that sequence conservation is higher at the protein than at the nucleotide level. Although this method may give information regarding gene location, it is unlikely to exhibit gene structure as intron–exon boundaries may vary between species (Ouyang et al. 2009). Therefore, alignment of genomic sequence with protein sequence database by programs, such as BLASTX, is usually followed by utilization of spliced alignment programs such as Genewise (Birney and Durbin 2000) or GeneSeqer (Usuka et al. 2000) to identify the gene structure by comparing the genomic DNA sequence to the target protein sequences (Davuluri and Zhang 2003).

3.3.3 Integrated Methods

In general, integrated methods combine homology-based approaches with *ab initio* approaches and thus make more accurate gene predictions (Allen et al. 2004; Yandell and Ence 2012; Goel et al. 2013). *Ab initio* predictions may be combined with homology-based data within a single program such as EUGENE'HOM (Foissac et al. 2003), AUGUSTUS (Stanke et al. 2006), GenomeScan (Yeh et al. 2001), Jigsaw (Allen and Salzberg 2005) and EvidenceModeler (http://evidence-modeler.github.io) or via an annotation pipeline with a set of consecutive processes. TIGR rice genome annotation was performed via the latter approach. Initial gene models were generated by the program Fgenesh (http://www.softberry.com) and the gene models were refined by the program PASA (Haas et al. 2003).

Automated gene prediction is a sort of artificial intelligence which perfectly assigns gene structures, but it is still subject to a degree of uncertainty in the absence of experimental evidence and need to be refined as new genome sequences or relevant experimental data become available (Thibaud-Nissen et al. 2008; Dale et al. 2012). For example, the analysis of the genomic sequence of *Arabidopsis* was initially reported in the year 2000 by the consortium of sequencing centers (The *Arabidopsis* Genome Initiative 2000), reannotated by TIGR over a period of 5 years (Haas et al. 2005), and is nowadays maintained by the *Arabidopsis* Information Resource (Rhee et al. 2003). The annotation data has changed dramatically since 2000 and improvements are still being made. Since automated gene prediction may

easily fail to identify certain aberrant gene structures such as noncanonical introns, polycistronic genes, and short genes, researchers should consider browsing the gene predictions together with any available evidence through an annotation viewer/editor, or even manually annotate genomes when necessary (Thibaud-Nissen et al. 2008). Sophisticated genome editors such as Apollo (Lee et al. 2009) and Artemis (Berriman and Rutherford 2003) enable users to go beyond passive viewing to interactively modifying and refining precise locations and structures of genes within genomes (Lee et al. 2013).

3.4 Functional Annotation

Once the structure of a gene is identified and the nucleic acid sequence is converted into a protein sequence, a putative function may be assigned to the predicted protein. Alignment of predicted sequences against a protein database is a common way of attributing a function to a protein. Sequence comparisons also can be utilized to determine particular motifs in a protein (e.g., ATP-binding, DNA-binding) and these may give information about function as well. Protein alignments against protein databases are usually performed with BLASTP. The number of protein hits and the quality of the results depend mostly on the parameters used for BLASTP. Expectation value (*E*-value), identity and coverage cut-offs are set empirically dependent largely on personal experience and representation of related sequences in the databases (Thibaud-Nissen et al. 2008; Ouyang et al. 2009). The UniProt Knowledgebase (UniProtKB) is the universal resource for extensive curated protein information, including classification, function, and cross-reference. It is composed of two sections: UniProtKB/Swiss-Prot which is manually annotated and reviewed and UniProtKB/TrEMBL which is automatically annotated and is not reviewed (Bairoch et al. 2005). The quality of the data in UniProtKB/Swiss-Prot is very high because the protein sequences are extensively annotated with information including function and biological role of the protein, protein family assignments, and bibliographical references. On the other hand, the less robust UniProtKB/TrEMBL database provides higher likelihood of finding a similar protein since it contains all of the protein sequences translated from EMBL/GenBank/DDBJ nucleotide sequence databases in addition to those in UniProtKB/Swiss-Prot. However, these entries require manual annotation unlike those in UniProtKB/Swiss-Prot (The UniProt Consortium 2011).

3.4.1 Domain Search

Although sequence comparison is a very powerful method for identification of gene function, its power largely depends on the volume of data available in the databases. The success of this method increases as more data accumulates in the databases, but it is still an important bottleneck to functional annotation. Significant matches of a

gene under question to another sequence may predict the biochemical and physiological function of the novel gene. In some cases, matches may occur to a gene from another organism whose function in that organism is unknown (Primrose and Twyman 2003). If there is no hit above a given threshold or no well-characterized hit is determined in the database, looking for conserved domains lying in the gene models may help to assign a function to the predicted protein (Ouyang et al. 2009). Databases of such conserved domains have been built based on analytical methods like domain profiles, motif recognition, fingerprints of collections of motifs and hidden Markov models. The InterPro database (http://www.ebi.ac.uk/interpro/) is a resource that enables functional analysis of protein sequences by categorizing them into families and predicting the presence of important domains and sites. PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, and PANTHER are the major databases that make up the InterPro consortium. InterPro combines the individual strength of each of these databases to generate a single resource for the scientists to access extensive information including protein families, domains, and functional sites. The InterPro database groups predictive models, known as signatures, supplemented by several different databases and generates additional functional annotations, including Gene Ontology (GO) terms wherever possible (Jones et al. 2014; Mitchell et al. 2015).

3.4.2 Gene Ontology

There are a high number of databases contributing functional analysis of large gene lists. Different kinds of information gathered from these databases should be integrated to make the best use of these sources. The Gene Ontology project (http://www.geneontology.org/) is a result of this integration effort and it provides a set of controlled, structured vocabularies, known as ontologies, to describe three features of gene products: biological process, location within cellular component and molecular function (Gene Ontology Consortium 2004; Zheng and Wang 2008). GO enables protein function to be associated with gross cellular or whole organism functions such as biosynthetic processes, growth, cell cycle, and nucleic acid replication (Primrose and Twyman 2003). Theoretically, GO terms should be assigned manually depending on experimental evidence. However, they can also be assigned based on sequence and structural similarity and phylogeny when there is no experimental data (Haas et al. 2005). Manual assignment of GO terms is time-consuming and consequently numerous tools have been developed for computational assignment of GO terms in large scale. These methods take the advantage of mapping (linking of various classification systems to GO terms) the gene products to proteins with identified GO terms, and the GO annotation is transferred to the query protein (Ouyang et al. 2009). Online GO annotation tools, such as GOA (Huntley et al. 2015), GOEAST (Zheng and Wang 2008), and GoFigure (Khan et al. 2003) are available for large-scale GO annotation.

High-throughput sequencing technologies caused generation of huge amount of data and a number of tools, described above, have been developed to predict all the

genes and assign functions to them. However, a predicted gene or predicted protein function is just a prediction and needs to be supported with direct experimental data to reduce the risk of disagreement between biological function and annotation. The controversy between biological function and annotation may be originated from inaccurate sequencing or inaccurate annotation and it may be eliminated through mutation analysis that involves inactivating the gene product by gene disruption (Primrose and Twyman 2006; Flagel and Blackman 2012). Next-generation sequencing technologies have generated an information reservoir sufficient to stimulate decades of follow-up research. Therefore effort and new resources might be more optimally canalized toward a deeper analysis of existing genomes. However, investment in plant genome sequencing continues due to the fact that sequencing is both a tool for new discovery and a means of identifying the function of known genes.

4 Molecular Phylogenetics

The fundamental use of DNA and protein-sequence data is comprehension of the cellular function. However, it can also be utilized to examine the evolution of genes and their protein products which is known as molecular phylogeny (Primrose and Twyman 2003). Conventionally, phenotypic characteristics were used to construct phylogenies and still these characteristics have an important role in the analysis of data such as fossils. With the advent of sequencing technologies, a combination of molecular and statistical techniques has become available to figure out evolutionary relationships among organisms or genes (Snel et al. 2005). All living organisms share fundamental molecular mechanisms and biological functions, reflecting that species descended from a common ancestor. Molecular phylogenetics study the structure and function of molecules and the way they evolve in time, to deduce these evolutionary relationships. Although this field of study emerged in the early twentieth century, it started to be effectively applied only after the advent of molecular biology techniques in 1960s, and has been reenergized as whole genome sequencing for complex organisms has become faster and less expensive (Li 1997; Lio and Goldman 1998; Hall 2004). Recently, molecular phylogenetics has become a tool used for genome comparisons: classification of metagenomic sequences, interpretation of genomes and identification of genes, regulatory elements, and noncoding RNAs in newly sequenced genomes. Molecular phylogenetics continues to grow and find new applications as the quantity of publicly available genomic data increases (Yang and Rannala 2012).

The main steps in any phylogenetic analysis are: assembly and alignment of a dataset, construction of phylogenetic trees from sequences utilizing computational methods and stochastic models, and statistical testing and evaluation of the constructed trees. The nature and scope of phylogenetic analysis may show significant variation and necessitate different datasets and computational methods. However, the main steps in any phylogenetic analysis remain the same (Lio and Goldman

1998; Linder and Warnow 2005). There are two methods of representation for evolutionary trees which graphically show relationships among species or genes over time; phylogenetic trees and dendograms. In phylogenetic trees, evolutionary distance is measured with respect to horizontal branch length, whereas it is measured along the length of the segments in dendograms. Although the results of each method seem different they show exactly the same relationships (Primrose and Twyman 2003).

There are two approaches used for the phylogenetic analysis of genome-scale data (Yang and Rannala 2012). The supertree approach relies on separate analysis of each gene and subsequent assembly of the subtrees for individual genes into a supertree for all species which is managed by the use of heuristic algorithms. The separate analysis enables analyzing the differences in the reconstructed subtrees or the extensiveness of horizontal gene transfer. On the other hand, it is ineffective for estimating a common phylogeny that underlies all genes (Bininda-Emonds 2004). In the supermatrix approach, a data supermatrix is created using interconnected sequences for multiple genes, in which lacking data are represented with question marks, and the supermatrix is then utilized for tree reconstruction. Differences in evolutionary dynamics among the genes are ignored by most of the supermatrix analyses (de Queiroz and Gatesy 2007).

There are numerous online tools and databases that apply popular methods to perform phylogenetic analysis. These include TNT, HYPHY, PAML, PHYLIP, MEGA, and BEAST (Table 3). A phylogenetic tree constructed by one of these tools may differ from the tree generated by any of the other tools depending on the algorithms and sources for sequence information utilized by the database. Therefore it is important to try different methods, compare the results, and then identify the database which works best for the type of dataset in question. Researchers may access a comprehensive list of phylogeny packages and free web servers at http://evolution.genetics.washington.edu/phylip/software.html.

Table 3 Several commonly used phylogenetic programs

Name	URL
TNT (Tree analysis using new technology)	http://www.lillo.org.ar/phylogeny/tnt/
HYPHY (Hypothesis testing using phylogenies)	http://www.hyphy.org
PAML (Phylogenetic analysis by maximum likelihood)	http://abacus.gene.ucl.ac.uk/software/paml.html
PHYLIP (the PHYLogeny Inference Package)	http://evolution.gs.washington.edu/phylip.html
MEGA (Molecular evolutionary genetic analysis)	http://www.megasoftware.net/
BEAST (Bayesian evolutionary analysis sampling trees)	http://beast.bio.ed.ac.uk/

5 Comparative Plant Genomics

The rapid rate at which genome sequence data accumulates has given rise to an era of comparative plant genomics. Comparative genomics analyzes the differences and similarities in genome structure and organization in different organisms. The desire for having a much more detailed understanding of the process of evolution and the requirement of translating DNA sequence data into proteins of known function are the two drivers for comparative genetics (Primrose and Twyman 2006). An application of comparative genomics is described above in Sect. 3.3 which defines utilization of homology based methods for identification of genes in newly sequenced genomes. DNA tends to have its primary sequence evolutionarily conserved and this enables prediction of biological function of target DNA by comparison with a related DNA sequence (Lyons and Freeling 2008). Comparison of the structure and contents of various species in a phylogenetic manner helps to understand the evolutionary forces that have shaped modern plant genomes (Flagel and Blackman 2012). These evolutionary forces can be listed as follows: (1) vertical descent (speciation) with modification; (2) gene duplication; (3) horizontal gene transfer (HGT); (4) gene loss; and (5) fission, fusion, and other rearrangements of genes (Koonin 2005). Being regarded as the primary events of genome evolution, vertical descent and duplication events have been well defined in the pregenomic era. On the other hand; gene loss, HGT, and gene rearrangements were considered among significant, fundamental generalizations of the emerging evolutionary genomics (Doolittle 2000; Koonin and Galperin 2003; Lawrence and Hendrickson 2003). Along with these essential evolutionary events, key concepts of evolutionary biology such as orthologs and paralogs should be well defined to allow a deeper understanding of evolution of genes, gene ensembles and ultimately, complete gene repertoires of organisms.

5.1 Orthologs and Paralogs

With the advent of complete genome sequencing, a new language has been developed to identify the relationships between genomes meaningfully, i.e., evolutionary genomics. Within this context orthology and paralogy are the two keystone definitions and a clear distinction between orthologs and paralogs is crucial for the development of a robust evolutionary classification of genes. Orthologs are homologous genes in different organisms that are related via speciation (vertical descent), whereas paralogs are homologous genes within an organism that are related via duplication. Orthologs evolve by gradual accumulation of mutations and they encode proteins with the same function. On the other hand, paralogs encode proteins with related but nonidentical functions and they arise by gene duplication followed by mutation accumulation (Primrose and Twyman 2006; Lyons and Freeling 2008). In some cases orthologous relationships are

inextricably intertwined with paralogous relationships due to the combination of speciation and duplication events, along with HGT, gene loss and gene rearrangements. Correct usage of the terms, ortholog and paralog is not only important because they provide accuracy to the descriptions of genome evolution but also because they have distinct and significant evolutionary and functional connotations. The key point in discrimination of orthologs is that they are derived from a single ancestral gene (Koonin 2005).

5.2 Synteny, Duplication, and Polyploidy

The comparative approach to plant genomics has led to emergence of many interesting findings such as conserved gene order, ancestral polyploidy, and conserved gene content. The first hints regarding conserved gene order came from comparative mapping efforts which showed that marker order was often conserved even when the species being compared no longer shared orthologous chromosomes due to rearrangements (Moore et al. 1995; Gale and Devos 1998). This shared gene order, known as synteny, is useful for assigning orthology and paralogy which are the key steps in defining evolutionary relationships between genomic regions. The first genomic search with regard to synteny came as the second plant genome sequence, rice (*Oryza sativa* spp.), was published (Goff et al. 2002; Yu et al. 2002). Rice genome sequence was compared to *A. thaliana* and it was revealed that these two species were diverged almost 200 million years ago and some pre-genomic analysis indicated that little synteny would remain (Devos et al. 1999). Subsequently the pre-genomic analysis was supported with the rice genome sequence which exerted that only modest stretches of synteny existed between rice and *A. thaliana* (Goff et al. 2002). In order to increase the power and sophistication of the synteny analyses, the comparisons between species should be typically updated as new genome sequences are released. By this way, we can improve our understanding of gene order and its evolutionary dynamics across plants.

Gene order conservation in vertebrates is much more evident (Mouse Genome Sequencing Consortium 2002; Smith et al. 2002) compared to the two major branches of the angiosperms (eudicots and monocots) (Grant et al. 2000; Rossberg et al. 2001; Vandepoele et al. 2002; Simillion et al. 2004). The rapid structural evolution of angiosperms appears to be due largely to whole-genome duplications and subsequent gene loss (Coghlan et al. 2005), fractionating ancestral gene linkages across multiple chromosomes (Paterson et al. 2004, 2005). Duplications may be local (tandem), segmental, whole chromosome, or whole genome (polyploidy) (Freeling 2008). Additionally, the reshuffling of short DNA sequences by transposable elements dramatically reduces the extent of large-scale colinearity in heterochromatic regions. These duplication and transposition events lead to a considerable variation in genome size and arrangement even within close relatives and complicate comparisons of gene arrangements in angiosperms (Bowers et al. 2005).

Recent and ancient polyploidy is quite common among angiosperms. Although recent polyploidy may be detected via chromosome counts (Adams and Wendel 2005), detection of ancient polyploidy requires an almost complete genome sequence. DNA duplication and subsequent gene loss leading to fractionation prevent detection of ancient polyploidies via chromosome counts and generally return a polyploid to a chromosomal number and gene count (Lyons and Freeling 2008). Recent analyses of ESTs (Pfeil et al. 2005; Cui et al. 2006) and genome sequences (Tuskan et al. 2006; Jaillon et al. 2007) suggest that almost all angiosperms are paleopolyploids. This paleopolyploid architecture of angiosperms necessitates the use of considerably different computational approaches for their genome comparisons with respect to those utilized for mammals and other organisms (Tang et al. 2008).

Many genes and gene families have essential functions and are conserved among living organisms. Although this holds true for plants, gene duplication and loss generate an extensive variation in the size of conserved genes families among plant species (Velasco et al. 2007). Early diverging plant lineages such as bryophytes, lycophytes, and the green algae show great gene family conservation compared to angiosperms. Comparisons of gene family sizes between some early diverging plants and angiosperms have shown that the early diverging plant lineages, the green algae *Chlamydomonas reinhardtii* and the lycophyte *Selaginella moellendorffii*, usually have lower number of genes per shared gene family compared to angiosperm species. On the other hand, the moss *Physcomitrella patens*, has many conserved gene families that are larger than those found in the angiosperms (Flagel and Blackman 2012). These patterns suggest that contribution of gene family expansion and contraction to evolution and diversification of plants could be greater than the production of novel genes (Flagel and Wendel 2009). From this point of view, understanding the evolutionary history of amplification and reduction in gene family size in plants can be seen as a tool for enlightening functional evolution. Analysis of the types of genes that have expanded or contracted within each lineage indicates the selective pressures that plant genomes face. For example, there is strong evidence pointing biased retention of transcription factors in *A. thaliana* and maize lineages after recent genome duplications (Blanc and Wolfe 2004; Seoighe and Gehring 2004). Moreover, a comparison of *A. thaliana*, moss, poplar, and rice showed that genes which respond to abiotic and biotic stress factors are preferentially conserved over evolutionary time when paralogs arise by tandem duplication rather than polyploidy (Hanada et al. 2008).

Although each species experiences essential gains and losses of genetic material, the aspects of genomic evolution described above have revealed that plants possess a conserved gene content. This may occur in the form of recent and ancient genome duplication or lineage-specific amplification and reduction in gene family size. New findings will help gaining a deeper understanding of these unique aspects of genomic evolution in plants.

5.3 Web Resources for Comparative Genomics

Comparative genomics has proven itself invaluable for understanding patterns and processes of genome evolution and also in illuminating aspects of gene function. The availability of complete genome sequences has notably changed our conception on the complexity of genome evolution, genome organization, gene function, and regulation in plants. However, the rapid increase in the number of available genomes complicates large-scale analyses for nonexperts, whereas the computational requirements to extract biological information grow rapidly. Evolutionary analyses are further complicated due to biological variation between species and differences in sequence quality. Therefore, databases for comparative genomics that may overcome some of these challenges are invaluable resources for experimental biologists. Table 4 summarizes major databases that plant biologists can get comparative genomics data.

6 Conclusion

Full genome sequences are available for numerous plant species. However, every genome assembly contains missing or inaccurate information so that the genome descriptions remain in some respects incomplete. Annotation of genomes is an imprecise, challenging, and ever-changing task as it is highly dependent on algorithms or homology-based evidence. Therefore annotation data need to be supported with direct experimental data to reduce the risk of disagreement between biological function and annotation. However, even the genome sequences of model organisms contain thousands of annotated genes lacking experimentally determined function. Thus, it could be argued that we already have enormous amount of genome sequence data at our fingertips. Therefore, effort and new resources might be more optimally canalized towards a deeper analysis of existing genomes. However, investment in plant genome sequencing continues due to the fact that sequencing is both a tool for discovery of new genes and a means of achieving information about gene function.

Analysis of genomes via sequence data is not only important for providing a deep understanding of the molecular and cellular function of genes but also for providing insight into evolutionary relationships among organisms or genes. Comparison of the structure and contents of various species in a phylogenetic manner helps to understand the evolutionary forces that have shaped modern plant genomes. Sequence data is also invaluable for identification of economically important genes and improvement of breeding programs. As more whole genome sequences become available it will increase the speed and lower the costs for studies regarding epigenomes, transcriptomes, and metabolomes.

Table 4 List of main web-based databases for plant comparative genomics

Name	URL	Description
PLAZA	http://bioinformatics.psb.ugent.be/plaza/	Offers comparative genomics data for 37 plant species and allows users to browse the annotated genomes, gene families, and phylogenetic trees.
CoGe	https://genomevolution.org/CoGe/GEvo.pl	GEvo, CoGe's Genome Evolution Analysis tool, compares multiple genomic regions from any number of organisms to identify patterns of genome evolution.
Phytozome	http://www.phytozome.net/	Provides a view of the evolutionary history of every plant gene at the level of sequence, gene structure, gene family, and genome organization and also provides access to the sequences and functional annotations of a growing number of complete plant genomes.
Gramene	http://www.gramene.org/	Applies a phylogenetic framework for genome comparisons and uses ontologies to integrate structural and functional annotation data. Whole-genome alignments complemented by phylogenetic gene family trees help infer syntenic and orthologous relationships.
PGDD	http://chibba.agtec.uga.edu/duplication/	Used to identify synteny information among plant genomes and is mainly focused on unraveling genome duplication events during the history of angiosperm evolution.
MEGA	http://www.megasoftware.net/	Utilized for reconstructing the evolutionary histories of species and inferring the extent and nature of the selective forces shaping the evolution of genes and species.
Ensembl Plants	http://plants.ensembl.org/index.html	The Ensembl Gene Tree pipeline is used to construct gene trees that reveal evolutionary history of gene families, including identification of candidate gene duplication and speciation events, derived from the multiple sequence alignments.

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8(2):135–141

Allen JE, Salzberg SL (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21(18):3596–3603

Allen JE, Pertea M, Salzberg SL (2004) Computational gene prediction using multiple sources of evidence. Genome Res 14(1):142–148

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9(3):208–218

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33(Database issue):D154–D159

Bennett MD, Smith JB (1991) Nuclear-DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 334(1271):309–345

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12:1021–1029

Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95(1):127–132

Berriman M, Rutherford K (2003) Viewing and annotating sequence data with Artemis. Brief Bioinform 4(2):124–132

Bininda-Emonds ORP (2004) Phylogenetic supertrees: combining information to reveal the tree of life. Kluwer, Dordrecht

Birney E, Durbin R (2000) Using GeneWise in the Drosophila annotation experiment. Genome Res 10(4):547–548

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16(7):1679–1691

Borevitz JO, Chory J (2004) Genomics tools for QTL analysis and gene discovery. Curr Opin Plant Biol 7(2):132–136

Bowers JE, Arias MA, Asher R, Avise JA, Ball RT et al (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc Natl Acad Sci U S A 102(37):13206–13211

Buisine N, Quesneville H, Colot V (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics 91(5):467–475

Cai J, Liu X, Vanneste K, Proost S, Tsai WC et al (2015) The genome sequence of the orchid Phalaenopsis equestris. Nat Genet 47:65–72

Campos-De Quiroz H (2002) Plant genomics: an overview. Biol Res 35(3–4):385–399

Campos-De Quiroz H, Magrath R, McCallum D, Kroymann J, Scnabelrauch D, Mitchell-Olds T, Mithen R (2000) Alpha-keto acid elongation and glucosinolate biosynthesis in *Arabidopsis thaliana*. Theor Appl Genet 101(3):429–437

Choulet F, Alberti A, Theil S, Glover N, Barbe V et al (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet 21(12):673–682

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16(6):738–749

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F et al (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature 488:213–217

Dale JW, Schantz MV, Plant N (2012) From genes to genomes: concepts and applications of DNA technology, 3rd edn. Wiley-Blackwell, Oxford

Das S, Hirano M (2012) Comparative genomics and genome evolution. Curr Genomics 13(2):85

Davuluri RV, Zhang MQ (2003) Computer software to find genes in plant genomic DNA. In: Grotewold E (ed) Plant functional genomics. Humana, Totowa, NJ, pp 87–107

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends Ecol Evol 22(1):34–41

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345: 1181–1184

Devos KM, Beales J, Nagamura Y, Sasaki T (1999) *Arabidopsis*-rice: will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res 9(9):825–829

Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F et al (2014) The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature 505:546–549

Doolittle WF (2000) Uprooting the tree of life. Sci Am 282(2):90–95

Duran C, Eales D, Marshall D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Appleby N, Batley J, Basford K, Edwards D (2010) Future tools for association mapping in crop plants. Genome 53(11):1017–1023

Eversole K, Feuillet C, Mayer KFX, Rogers J (2014) Slicing the wheat genome. Science 345:285–287

Flagel LE, Blackman BK (2012) The first ten years of plant genome sequencing and prospects for the next decade. In: Wendel JF, Greilhuber J, Doležel J, Leitch IJ (eds) Plant genome diversity, vol 1, Plant genomes, their residents, and their evolutionary dynamics. Springer, New York, pp 1–15

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytol 183(3):557–564

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8(9):967–974

Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T (2003) EUGENE'HOM: a generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res 31(13):3742–3745

Freeling M (2008) The evolutionary position of subfunctionalization, downgraded. In: Volff J-N (ed) Plant genomes, vol 4. Karger, Basel, pp 25–40

Gale MD, Devos KM (1998) Plant comparative genetics after 10 years. Science 282(5389): 656–659

Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue):D258–D261

Gill BS, Appels R, Botha-Oberholster AM, Buell CR, Bennetzen JL et al (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics 168(2):1087–1096

Goel N, Singh S, Aseri TC (2013) A comparative analysis of soft computing techniques for gene prediction. Anal Biochem 438(1):14–21

Goff SA, Ricke D, Lan TH, Presting G, Wang RL et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp japonica). Science 296(5565):92–100

Grant D, Cregan P, Shoemaker RC (2000) Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc Natl Acad Sci U S A 97(8):4168–4173

Greilhuber J, Borsch T, Muller K, Worberg A, Porembski S, Barthlott W (2006) Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. Plant Biol 8(6):770–777

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31(19):5654–5666

Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RK Jr, Maiti R, Chan AP, Yu C, Farzad M, Wu D, White O, Town CD (2005) Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. BMC Biol 3:7

Hall BG (2004) Phylogenetic trees made easy: a how-to manual, 2nd edn. Sinauer Associates, Sunderland, MA

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38(22):e199

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148(2):993–1003

Hirakawa H, Shirasawa K, Miyatake K, Nunome T, Negoro S, Ohyama A, Yamaguchi H, Sato S, Isobe S, Tabata S, Fukuoka H (2014) Draft genome sequence of eggplant (Solanum melongena L.): the representative Solanum Species INDIGENOUS TO THE OLD WORLD. DNA Res 21:649–660

Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. Genomics 46(1):37–45

Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Res 43(Database issue):D1057–D1063

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161):463–467

Jones P, Binns D, Chang HY, Fraser M, Li W et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110(1–4): 462–467

Katam R, Panthee D, Basenko E, Bandopadhyay R, Basha SM, Eswaran K, Kole C (2010) *Arabidopsis* genome initiative. In: Kole C, Abbott AG (eds) Principles and practices of plant genomics, vol 3, Advanced genomics. CRC Press, New York, pp 175–204

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12(4):656–664

Khan S, Situ G, Decker K, Schmidt CJ (2003) GoFigure: automated Gene Ontology annotation. Bioinformatics 19(18):2484–2485

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39: 309–338

Koonin EV, Galperin MY (2003) Sequence–evolution–function: computational approaches in comparative genomics. Kluwer, Boston

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59

Lawrence JG, Hendrickson H (2003) Lateral gene transfer: when will adolescence end? Mol Microbiol 50(3):739–749

Lee S-I, Kim N-S (2014) Transposable elements and genome size variations in plants. Genomics Inform 12(3):87–97

Lee E, Harris N, Gibson M, Chetty R, Lewis S (2009) Apollo: a community resource for genome annotation editing. Bioinformatics 25(14):1836–1837

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE (2013) Web Apollo: a web-based genomic annotation editing platform. Genome Biol 14(8):R93

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Li FG, Fan GY, Lu CR, Xiao GH, Zou CS et al (2015) Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol 33:524–530

Linder CR, Warnow T (2005) An overview of phylogeny reconstruction. In: Aluru S (ed) Handbook of computational molecular biology. Chapman & Hall, New York, pp 19-1–19-39

Lio P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Res 8(12):1233–1244

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26(4):1107–1115

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53(4):661–673

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20(16):2878–2879
Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M et al (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092
Mathe C, Sagot MF, Schiex T, Rouze P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30(19):4103–4117
Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S et al (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(Database issue):D213–D221
Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution—grasses, line up and form a circle. Curr Biol 5(7):737–739
Morgulis A, Gertz EM, Schaffer AA, Agarwala R (2006) WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22(2):134–141
Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13(4):477–478
Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915):520–562
Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D et al (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59
Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584
Ouyang S, Thibaud-Nissen F, Childs KL, Zhu W, Buell CR (2009) Plant genome annotation methods. In: Somers DJ, Langridge P, Gustafson JP (eds) Plant genomics: methods and protocols. Humana, New York, pp 263–282
Palleja A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics 9:335
Parida SK, Mohapatra T (2010) Whole genome sequencing. In: Kole C, Abbott AG (eds) Principles and practices of plant genomics, vol 3, Advanced genomics. CRC Press, New York, pp 120–174
Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A 101(26):9903–9908
Paterson AH, Freeling M, Sasaki T (2005) Grains of knowledge: genomics of model cereals. Genome Res 15(12):1643–1650
Peterson DG (2014) Evolution of plant genome analysis. In: Paterson AH (ed) Genomes of herbaceous land plants, vol 69, Advances in botanical research. Elsevier, Paris, pp 13–46
Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. J Agric Genom 5. www.ncgr.org/research/jag
Pfeifer M, Kugler KG, Sandve SR, Zhan BJ, Rudi H et al (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091
Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol 54(3):441–454
Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):i351–i358
Primrose SB, Twyman RM (2003) Principles of genome analysis and genomics, 3rd edn. Blackwell, Berlin
Primrose SB, Twyman RM (2006) Principles of gene manipulation and genomics, 7th edn. Blackwell, Oxford
Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D et al (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res 31(1):224–228

Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and Capsella genomes. Plant Cell 13(4):979–988

Schellenberg MJ, Ritchie DB, MacMillan AM (2008) Pre-mRNA splicing: a complex picture in higher definition. Trends Biochem Sci 33(6):243–246

Schmidt R (2002) Plant genome evolution: lessons from comparative genomics at the DNA level. Plant Mol Biol 48(1–2):21–37

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet 20(10):461–464

Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci U S A 89:8794–8797

Simillion C, Vandepoele K, Saeys Y, Van de Peer Y (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res 14(6):1095–1106

Sleator RD (2010) An overview of the current status of eukaryote gene prediction strategies. Gene 461(1–2):1–4

Smith SF, Snell P, Gruetzner F, Bench AJ, Haaf T, Metcalfe JA, Green AR, Elgar G (2002) Analyses of the extent of shared synteny and conserved gene orders between the genome of Fugu rubripes and human 20q. Genome Res 12(5):776–784

Snel B, Huynen MA, Dutilh BE (2005) Genome trees and the nature of genome evolution. Annu Rev Microbiol 59:191–209

Soneji JR, Rao MN, Sudarshana P, Panigrahi J, Kole C (2010) Current status of on-going genome initiatives. In: Kole C, Abbott AG (eds) Principles and practices of plant genomics, vol 3, Advanced genomics. CRC Press, New York, pp 305–353

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5(11):e1000734

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225

Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol 7(Suppl 1):S11.1–18

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320(5875):486–488

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815

The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

The International Peach Genome Initiative (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45:487–494

The International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345:1251788

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39:D214–D219

Thibaud-Nissen F, Wortman J, Buell CR, Zhu W (2008) Structural, functional, and comparative annotation of plant genomes. In: Kahl G, Meksem K (eds) The handbook of plant functional genomics: concepts and protocols. Wiley-VCH, Weinheim, pp 373–395

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313(5793):1596–1604
Usuka J, Zhu W, Brendel V (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. Bioinformatics 16(3):203–211
Vandepoele K, Simillion C, Van de Peer Y (2002) Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. Trends Genet 18(12):606–608
Varshney RK, Song C, Saxena RK, Azam S, Yu S et al (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol 31:240–246
Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2(12):e1326
Vinogradov AE (2004) Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? Curr Opin Genet Dev 14(6):620–626
Wang Z, Chen Y, Li Y (2004) A brief review of computational gene prediction methods. Genomics Proteomics Bioinformatics 2(4):216–221
Wang J, Kong L, Gao G, Luo J (2013) A brief introduction to web-based genome browsers. Brief Bioinform 14(2):131–143
Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875
Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U et al (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol 32:656–662
Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13(5):329–342
Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13(5):303–314
Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res 11(5):803–816
Yu J, Hu SN, Wang J, Wong GKS, Li SG et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp indica). Science 296(5565):79–92
Zheng Q, Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36(Web Server issue):W358–W363

Genomics Resources for Plants

Adeel Malik

Contents

Abstract Recently, genomics research in plants has offered considerable quantity of complex data which in turn has enhanced the development of specialized databases and analysis tools. The explosion in the growth of focused plant databases and tools is attributed to the advancement in the next-generation sequencing technology which has quickly generated enormous volumes of data at a reasonably low cost. These specialized plant databases cover varied types of information such as physical and genomic sequence maps, QTLs, loci and their several alleles, analysis tools, phenotypes, stocks, sequence information, and molecular markers. However, the main challenge to decipher this overflowing data for the improvement of crops is ever growing with an increase in the size as well as type of data. To address this issue, many new databases are constantly being developed with numerous analysis tools. The main purpose of these plant resources is to provide a single reserve for all

A. Malik (✉)
Perdana University Centre for Bioinformatics (PU-CBi), MARDI Complex,
Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia
e-mail: adeel@procarb.org

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_2

the information that may be important for an individual or a group of species. Here, we review several of the extensively used genomics resources for plants, and we hope that this chapter will be of some help by serving as an initial point for people with research interests towards plant science.

Keywords QTL • Bioinformatics • Genomics • Molecular markers • Sequence maps • Genome projects

1 Introduction

The DNA sequence and expressed gene sequence data that is being produced in the recent times mainly comes from the next-generation sequencing (NGS) technologies, which generate massive volumes of data at a reasonably low price and short time (Lai et al. 2012). NGS technology has been applied to sequence genomes of several plant species of agricultural significance (International Rice Genome Sequencing Project 2005; Paterson et al. 2009; Schnable et al. 2009). The main challenge that is faced by the research community is how to decode this plenteous data for the improvement of crops. There still remains a gap from the genomic data generation to crop improvement. However, many companies and individual investigators are attempting to bridge this gap by developing methods to mine the genomic data with emphasis on crop improvement (Lai et al. 2012). For example, modern genetic technologies such as marker-assisted selection (MAS) and genetic engineering (GE) have previously steered the development of new crop varieties. Additionally, novel technologies like genome editing, have shown great potential for crop improvement (Ronald 2014).

Over the last few years with the advancement in technology, the plant genomics research has provided substantial amount of diverse data which in turn boosted the development of generic DNA sequence as well as specialized databases (Lai et al. 2012). These plant databases encompass enormous information on plants including physical and genomic sequence maps, QTLs, loci and their several alleles, analysis tools, phenotypes, stocks, sequence information, and molecular markers. (Bal and Reddy 2014). This data has been growing continuously over the past few years equally in terms of type and volume. The objective of these databases is to offer a single resource for all the information that may be significant to an individual or a group of species. Some of the earlier developed databases (Erwin et al. 2007; Love et al. 2005; Stein and Thierry-Mieg 1998) may have limited functionality; however, the newer ones are continuously evolving with several embedded analysis tools such as GBrowse (Arnaoudova et al. 2009; Donlin 2007) and EnsEMBL system (Flicek et al. 2011). With the advancement in genomic technology, an increasing number of crop genomes are becoming available with a rapid increase in the number of specialized databases. In this chapter, we summarize some of the widely used resources for plant genomics and also provide their web links wherever possible.

We hope this chapter will be of some help and serve as a starting point to researchers interested in plant science.

2 1000 Plants (oneKP or 1KP)

Availability: https://sites.google.com/a/ualberta.ca/onekp/

The 1000 plants also known as oneKP or 1KP, is an initiative that has produced large-scale gene sequencing data for over 1000 plant species. The 1KP venture is a worldwide multidisciplinary consortium with main supporters being Alberta Ministry of Innovation and Advanced Education, Musea Ventures (Somekh Family Foundation), Beijing Genomics Institute in Shenzhen (BGI-Shenzhen), China National GeneBank (CNGB), iPlant Tree-of-Life (iPToL) Grand Challenge, Compute Canada (Westgrid), Alberta Innovates Technology Futures (AITF-iCORE Strategic Chair). Initially, the selection of samples was centered on a series of overlapping subprojects with scientific aims that might be addressed by sequencing of numerous plant species. With additional groups joining 1KP, however, the goals progressed and are now epitomized by the varied collection of papers (https://pods.iplantcollaborative.org/wiki/display/iptol/OneKP+companion+papers).

On an average, 1KP dataset consists of about 2Gb of RNA-seq data per sample on the Illumina sequencing platform (GA2 or HiSeq). In general, one tissue sample per species was sequenced, except in cases where the scientific goals required more. SOAPdenovo-trans (http://soap.genomics.org.cn/SOAPdenovo-Trans.html), a de novo transcriptome assembler was used to assemble paired-end data. Typically, each sample generated 10 k scaffolds with lengths of more than 1 kb. All the resultant data which consists of raw unassembled reads, assembled transcriptomes, and as an estimate of gene expression levels, averaged read depths computed across each scaffold is available at Westgrid (http://onekp.westgrid.ca/1kp-data) and TACC (http://web.corral.tacc.utexas.edu/OneKP) databases. A major highlight of 1KP is that the samples were selected to represent all known species across the plant kingdom. Attempts were made to sequence at least a representative of approximately all of the 415 known angiosperm families, and almost one fifth of the sequenced species are algae. Such large-scale gene sequencing was never attempted on most of these species (https://pods.iplantcollaborative.org/wiki/display/iptol/OneKP+Capstone+Wiki).

3 *Amborella* Genome Database

Availability: http://www.amborella.org/

Amborella genome database provide access to all the data that is generated as a result of *Amborella* Genome Project. The project has three main goals:

(a) To produce a draft genome sequence and gene annotation of superior quality for *Amborella* trichopoda using a whole genome shotgun approach.
(b) By using fluorescent in situ hybridization (FISH) and single molecule restriction mapping (i.e. optical mapping), extend and improve the large-scale physical map of *Amborella*.
(c) Develop a freely available database with bioinformatics analysis tools that supports comparative analyses and data mining for researchers interested in plant science.

Amborella Genome Project is funded by the National Science Foundation Plant Genome Research Program (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5338&org=BIO) with the main participants being the scientists from Penn State University, University at Buffalo, University of Georgia, University of Florida, University of California Riverside, and Indiana University. In addition to these, several other contributors from universities in the United States, Canada, Mexico, China, Korea, Taiwan, Singapore, Germany, Italy, Denmark, France, and New Caledonia also contribute to the *Amborella* genome project (http://amborella.huck.psu.edu/project).

Amborella trichopoda is believed to be a lone living species of the sister lineage to all other existing flowering plants thereby offering an exclusive reference for deducing the genome content and structure of the most recent common ancestor (MRCA) of living angiosperms. The sequencing of *Amborella* genome led to the identification of an ancient genome duplication preceding angiosperm diversification, without any proof of subsequent, lineage-specific genome duplications (*Amborella* Genome Project 2013). *Amborella* has the most widespread genomic resources amongst all basal angiosperms and already has high-quality genomic libraries, a physical map, and a huge transcriptome database. Availability of these resources, along with its essential phylogenetic position and reasonable genome size (870 million base pairs), makes *Amborella* the excellent choice for the first basal angiosperm to be completely sequenced (http://www.amborella.org/).

4 BAR

Availability: http://bar.utoronto.ca/

The Botany Array Resource (BAR) offers a means to obtain and store microarray data for Arabidopsis thaliana as well as provides variety of user friendly tools such as Expression Browser (to carry out "electronic Northerns"), Expression Angler (for detecting genes that are co-regulated with a gene of interest), and Promomer (for finding possible cis-elements in the promoters of individual or co-regulated genes) for viewing and mining gene expression data (Toufighi et al. 2005). The BAR web portal can be divided into three sections based on the types of available analytical tools. These sections are summarized as follows:

4.1. *Gene Expression and Protein Tools:* This section is dedicated to tools that allow users to view expression patterns as electronic fluorescent pictographs or heatmaps, explore promoters, and identify protein–protein interactions. The section is further subdivided and includes tools as described below:

4.1.1. *ePlant:* This tool is linked to numerous freely available databases so that a user can download the most recent genome, interactome, and transcriptome datasets. The data can be displayed with visualization tools with a user interface that can be zoomed.

4.1.2. *Expression Anglers:* Expression Angler aims to identify coexpressed, anti-correlated or condition/tissue-specific genes in 5 of the data sets available at the AtGenExpress Consortium, an in-house database of BAR, or from NASCArrays. The other types of tools that are available under this section include:

4.1.2.1. *Poplar Expression Angle:* finds coexpressed, anti-correlated, or condition/tissue-specific genes in two poplar data sets specifically from the Campbell Laboratory.

4.1.2.2. *Sampler Angle:* identifies samples that may have similar expression profiles and might be valuable for chemical genomics or typing mutant tissues.

4.1.2.3. *e-Northerns w. Expression Browser:* performs electronic Northerns by means of the Expression Browser, and can be used to ask questions such as how up to 125 genes of interest are being expressed, with the gene expression data sets amassed up to now in the BAR database or other datasets like AtGenExpress Consortium. A variant of this tool is Poplar Expression Browser that performs northerns using the gene expression data from Campbell Lab's PopGenExpress developmental or drought stress data sets.

4.1.3. *Arabidopsis eFP Browsers:* These set of tools help in generating "electronic fluorescent pictographic" illustrations for the gene of interest's expression patterns based on the Map of Arabidopsis Development (Schmid et al. 2005), the AtGenExpress Consortium data (Abiotic Stress—Kilian et al. 2007, Biotic Stress, and the Chemical and Hormone Series), cell-type or seed-specific data and other data. In addition to these Arabidopsis specific eFP Browsers, there are few eFP Browsers dedicated to various dicots (Poplar, Medicago, Soyabean, Potato, Tomato, and *E. salsugineum*), some monocots (Maize, Rice, Barley, and Triticale) as well as for a limited number of animals (mouse and human).

4.2. *Other Gene Expression & Protein Tools*

4.2.1. *Expressolog Tree Viewer:* This tool can be used to find expressologs for the gene of interest in other species, specifically, homologous genes that show analogous expression patterns in same tissues in other species.

4.2.2. *Promomer:* Recognizes overrepresented n-mer "words" in the promoter of a gene of interest, or in promoters of coexpressed genes.

4.2.3. *Cistome:* Predicts novel cis-elements in the promoters of coexpressed genes by using various cis-element prediction programs, or an updated version of Promomer known as Promomer 2.

4.2.4. *Arabidopsis Interactions Viewer (AIV):* Allows viewing of predicted as well as experimentally determined protein–protein interactions in Arabidopsis (http://bar.utoronto.ca/affydb/BAR_instructions.html#AIV). AIV searches a database of 70944 predicted and 36329 experimentally verified interacting proteins in Arabidopsis. The predicted interactions also known as *interologs* were previously constructed to better understand the overall signaling in Arabidopsis (Geisler-Lee et al. 2007). The experimentally verified Arabidopsis protein–protein interactions (PPIs) are derived from databases such as BIND (Bader et al. 2003), high-density Arabidopsis protein microarrays (Popescu et al. 2007; Popescu et al. 2009), Arabidopsis Interactome Mapping Consortium (Braun et al. 2011), MIND (https://associomics.dpb.carnegiescience.edu/Associomics/Home.html), and various other literature sources (http://bar.utoronto.ca/interactions/cgi-bin/arabidopsis_interactions_viewer.cgi).

4.2.5. *Rice Interactions Viewer:* Queries a database of 37472 predicted and 430 verified rice interacting proteins (http://bar.utoronto.ca/interactions/cgi-bin/rice_interactions_viewer.cgi).

4.2.6. *Gene Slider:* Can be used to create long sequence logos or automatically highlight motifs of interest. A navigation slider at the end of the display page specifies which area of the sequence is currently being shown (http://bar.utoronto.ca/geneslider/).

4.3. Molecular Markers and Mapping Tools

These set of tools help in performing Next-Generation Mapping, or generate your own markers using our molecular marker tools.

4.3.1. *Next-Generation Mapping (NGM):* Allows for the fast localization of recessive EMS-induced mutations within an F2 mapping population that has been combined and sequenced en masse using next-generation sequencing platforms (http://bar.utoronto.ca/ngm/).

4.3.2. *Marker Tracker:* A storehouse for genetic markers available at the University of Toronto. Presently the database consists of all potential markers for Arabidopsis thaliana that were generated by scripts using biopython library. A significant feature of Marker Tracker is the virtual gel representations which show just how the CAPS markers would run for individual accession (http://bar.utoronto.ca/markertracker/).

4.3.3. *Blast Digester:* It will analyze a nucleotide BLAST output for restriction enzymes that cleave the aligned sequences differentially. It is valuable for making CAPS or PCR-RFLP markers for mapping (Ilic et al. 2004) (http://bar.utoronto.ca/ntools/cgi-bin/ntools_blast_digester.cgi).

Other Genomic Tools and Widgets

This section consists of a large variety of additional tools that can be used for carrying out diverse tasks such as remove duplicates from lists, perform multidimensional Venn analyses, or generate random lists of identifiers. The complete list of these tools is provided in Table 1.

5 Cacao Genome Database (CGD)

Availability: http://www.cacaogenomedb.org/

Theobroma cacao (cacao tree) is a short, tropical tree grown in several countries including Côte d'Ivoire, Ghana, Indonesia, Nigeria, Brazil, Cameroon, Ecuador, Colombia, Mexico, and Papua New Guinea and is a source of cocoa powder, a key component of chocolate (Saski et al. 2011). As per a report, the world cocoa production was projected to be three million tons in 2010 with an annual expected average growth rate of 2.2 % from 1998 to 2010 (http://www.worldcocoafoundation.org/learn-about-cocoa/cocoa-facts-and-figures.html). The production of Cacao is presently under risk from numerous sources including an increase in the frequency of fungal diseases such as black pod, frosty pod, and witches' broom (Evans 2007; Hebbar 2007). The announcement of the complete cacao genome sequence has offered scientists with access to the state-of-the-art genomic tools, allowing added efficient investigation and accelerating the breeding practice, thus advancing the release of higher cacao cultivars. A huge volume of sequence data, transcript data, physical map data, and single-nucleotide polymorphism (SNP) data is being generated as a consequence of the cacao genome sequencing project.

The sequenced genotype, Matina 1–6, is a typical of the genetic background usually found in the cacao producing countries, facilitating outcomes to be applied instantly and largely to existing commercial cultivars. Matina 1–6 is extremely homozygous which significantly decreases the complication of the sequence assembly process. The current release of the sequencing data is a preliminary one and already covers about 92 % of the genome, with an estimate 35,000 genes. The data is accessible via Cacao Genome Database (http://www.cacaogenomedb.org/main). The database can be searched by one of the five different ways:

5.1. *Gene OR Sequence:* search genes or unigene contigs, markers, and supercontigs by providing names in the search form.
5.2. *Homology:* Searches genes that are annotated with explicit function assigned by homology.
5.3. *InterPro:* Search genes that are annotated with particular protein domains.
5.4. *Gene Ontology:* Find genes that are annotated with specific GO terms or search all GO terms that are annotated to one or more genes.
5.5. *KEGG:* Search genes that are annotated with specific KEGG terms or discover all KEGG terms linked to one or more genes.

Table 1 Additional genomics-related tools and widgets that are available at BAR

S. No.	Tool	Description	Link
1.	Arabidopsis Citation Network Viewer	Shows 54,033 publications since 1965 to March 2015, and draws lines to represent citations among the papers.	http://bar.utoronto.ca/~jwaese/50YearsOfArabidopsis/
2.	ClustalW with MView Output	A web-based version of the multiple sequence alignment that runs on BAR	http://bar.utoronto.ca/ntools/cgi-bin/ntools_multiplealign_w_mview.cgi
3.	DataMetaFormatter	A tool for reformatting Arabidopsis expression data and adding numerous quantities of meta-information, such as protein–protein interactions, functional classification etc.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_treeview_word.cgi
4.	HeatMapper Plus	A tool for applying third dimension of information through color-coding to a 2D table of data and a thumbnail graphic in addition to the table.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_heatmapper_plus.cgi
5.	Duplicate Remover	Removes duplicates in lists.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_duplicate_remover.cgi
6.	Venn Selector	Shows identifiers in common and unique to two sets of sequences.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_venn_selector.cgi
7.	Venn SuperSelector	Allows you to input multiple lists of genes or term with associated values.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_venn_superselector_geneview_values.cgi
8.	The Random ID List Generator	Generate n sets of y genes containing z number of randomly generated Arabidopsis AGI IDs.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_random_list_generator.cgi
9.	AGURR	a tool for identifying and visualizing outliers in an expression data set across a multitude of experimental conditions	http://bar.utoronto.ca/agurr/cgi-bin/agurr_main_driver.cgi
10.	Classification SuperViewer	Generates an overview of functional classification of a list of AGI IDs based on the GO database.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_classification_superviewer.cgi

11.	_at to AGI converter	Converts 25 k Affy GeneChip IDs into AGI IDs and vice versa.	http://bar.utoronto.ca/ntools/cgi-bin/ntools_agi_converter.cgi
12.	MASTA	Tool for probing differentially expressed genes against a microarray database for in silico suppressor/enhancer and inhibitor/activator screens.	http://bar.utoronto.ca/masta/masta.html
13.	GeneMANIA	GeneMANIA finds other genes that are related to a set of input genes using wide array of data such as protein and genetic interactions, pathways, coexpression, colocalization and protein domain similarity.	http://genemania.org/
14.	Topographic Phylomaps	Displays phylogenetic relationships.	http://bar.utoronto.ca/~jwaese/TopographicPhyloMap/

CGD also offers some broad data mining tools for sequence and annotations from the *T. cacao* cv. Matina 1–6 cacao genome sequencing project. These tools are summarized as follows:

5.6. *CacaoCyc:* A Pathway Genome Database (PGDB) for Theobroma cacao cv. Matina 1–6 created by using pathway tools software from SRI International. CacaoCyc was generated from the initially annotated release (v0.9) of T. cacao cv. Matina 1–6 genome sequence and consists of 437 predicted pathways and 2570 predicted reactions comprising about 8119 enzymes (http://cacaocyc.cacaogenomedb.org/). The database can be searched for genes, proteins, compounds, RNAs, reactions, pathways, operons, and GO terms.

5.7. *GBrowse:* Cacao GBrowse can be used to visualize gene models, transcripts, the results of genome-wide SNP association studies, as well as linkage studies (http://www.cacaogenomedb.org/tools/gbrowse).

5.8. *CMAP:* The CMap Comparative Map Viewer lets a user to visualize and compare genetic maps between and among species (http://www.cacaogenomedb.org/tools/cmap).

5.9. *GBrowse_syn:* It can be used to show several genomes, with a central reference species compared to two or more additional species (http://www.cacaogenomedb.org/tools/gbrowse_syn).

5.10. *WebFPC:* Used to view the Cacao physical map (http://www.cacaogenomedb.org/tools/webfpc).

5.11. *BLAST:* CGD BLAST allows users to conduct homology searches between their sequences against the publicly available Cacao EST sequences (http://www.cacaogenomedb.org/tools/blast).

6 CerealsDB

Availability: http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php

CerealsDB is a web-based resource for wheat (*Triticum aestivum)* encompassing a variety of genomic datasets that may help researchers and plant breeders so that the most suitable markers for marker-assisted selection can be selected. The CerealsDB database was constructed by Functional Genomics Group at the University of Bristol and now contains more than 100,000 varietal SNPs. Many of these SNPs have been experimentally verified. In addition, CerealsDB also comprises of databases for DArT markers and EST sequences, as well as links to a draft genome sequence for the Chinese Spring wheat variety (Wilkinson et al. 2012). The CerealsDB also allows users to look for diversity array technology (DArT) markers bin-mapped to chromosomes arms. The Wheat Ideograms on the DArT page allows access to a list of bins comprising DArT markers which can be opened by clicking on the suitable chromosome image (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/dart_index.php).

The information about SNP markers is divided based on platform as follows:

6.1. *Axiom® 820 K and 35 K SNP Arrays:* A collaborative effort between researchers at Affymetrix and Functional Genomics Group at the University of Bristol, resulted in the development of an "820,000" and a "35,000" feature wheat SNP array (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/axiom_download.php).

6.2. *iSelect Array:* Encompasses more than 80,000 SNP loci out of which about 44,000 have been mapped (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/iselect_mapped_snps.php).

6.3. *KASP probes:* Competitive Allele-Specific polymerase chain reaction (KASP) assay technology has facilitated high-throughput array-based discovery of sequence polymorhpisms (Allen et al. 2011). KASP platform consists of about 100,000 varietal SNPs in SNP database. There are 8700 markers for which assays have been developed and 7228 marker assays have been validated. The number of markers that have been mapped is 5033 (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/kasp_mapped_snps.php). The options to search KASP SNP database include:

6.3.1. *Select primers*—Provides details on the primers for individual SNPs or for all the SNPs on a specific chromosome. On the other hand, all SNPs found on an individual chromosome can be downloaded as an Excel worksheet.

6.3.2. *Obtain haplotypes*—The ideogram provides access to haplotype information for all the wheat varieties tested at Bristol. By clicking the appropriate chromosome image of the ideogram, a user will have access to the details on that particular chromosome. Haplotypes can also be selected for specific varieties on particular subgenomes and chromosomes.

6.3.3. *Contig information*—Expedites the search for information about the contig on which a SNP is observed by providing a SNP ID, or the contig name to see whether it contains any SNPs. If a SNP is found, the output shows the contig sequence, SNP position, mapping details, and any associated BLAST annotation.

6.3.4. *Align to Brachy*—By using a known SNP_id, this option can be used to discover additional SNPs containing sequences that align to the similar region in *Brachypodium*. As of 22nd Sep, 2015, out of approximately 38,000 wheat contigs in the CerealsDB database merely 5000, 13,000, and 19,000 align to *Brachypodium* at e-Values of 100, 45, and 10 respectively.

6.4. *TaqMan® probes:* A partnership between CerealsDB and Life Technologies (http://www.thermofisher.com/my/en/home.html) to design a pool of 4800 TaqMan® (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/taqman.php) SNP Assays for varietal markers within the wheat genome (http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/taqman_mapped_snps.php).

Table 2 Different BLAST search options available at CerealsDB

S. No.	Description of BLAST	Link
1.	BLAST search against CerealsDB	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/blast_kaspar.php
2.	BLAST search against the Whole Genome Shotgun assembly of the wheat line "Synthetic W7984" available from the European Nucleotide Archive (Accession PRJEB7074).	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/blast_WGS.php
3.	BLAST search the Chinese Spring draft genome assembly or raw sequence reads	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/search_reads.php
4.	BLAST search the Rye exome assembly or raw sequence reads	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/search_rye_reads.php

CerealsDB also has a diverse set of BLAST search tools that can be accessed from http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/DOC_BLAST.php and are summarized in Table 2.

7 Gramene

Availability: http://archive.gramene.org/

Gramene is an online open database for plant comparative genomics with initial focus on RiceGenes project (McCouch and Paul 1993; Ware et al. 2002) but now in addition to numerous species of rice, it has developed into a major resource for diverse plants such as *Arabidopsis*, *Brachypodium*, maize, sorghum, poplar, and grape. The current release [Release Notes 43 (December 2014)] of Gramene hosts about 39 complete and some partial genomes (http://gramene.org/release-notes-43) that can be visualized using Ensemble (http://ensembl.gramene.org/genome_browser/index.html). The annotations present in Gramene comprise of *ab initio*, evidence-based and community-generated gene predictions, repeat regions, and homology in addition to cross-references to sequences in public databases, locations of quantitative trait loci (QTLs), locations of microarray probes, cross-references to sequences in public databases and genome variation such as SNPs and indels (Youens-Clark et al. 2011). Besides having large number of complete and partial genomes, Gramene also has variation data incorporated into the genomes that may help in understanding the significance of variation (Youens-Clark et al. 2011). The various different components of Gramene are summarized in Table 3.

In order to explore the whole genome alignments (WGA), Gramene offers pre-computed whole genome and gene–gene alignments centered on LastZ (Harris 2007) or its originator BlastZ (Schwartz et al. 2003) and translated BLAT (tBLAT) (Kent 2002). Gramene also utilizes standard Ensembl GeneTree approaches (15) to

Table 3 Showing different sections of Gramene database

S. No.	Module	Description	Link
1.	Genomes	Provides data integration and visualization tools for genome annotations and comparisons.	http://archive.gramene.org/genome_browser/index.html
2.	Markers	Displays basic information about the various markers used for mapping. The precise information shown depends on the type of marker, however all markers will show the marker name, synonyms, source species, and a listing of map positions.	http://archive.gramene.org/markers/
3.	Genetic Diversity	This section focuses in storing genotypes, phenotypes and their environments, germplasm, and association data.	http://archive.gramene.org/db/diversity/diversity_view
4.	Pathways	Offers access to pathway databases for rice, maize, Bracypodium, and sorghum, respectively. Also provides mirrors of pathway databases from Arabidopsis, tomato, potato, pepper, coffee, Medicago, *E. coli*, and the MetaCyc and PlantCyc reference databases.	http://archive.gramene.org/pathway/
5.	Proteins	Provides details on Swissprot-Trembl protein entries from family Poaceae (Grasses).	http://archive.gramene.org/protein/
6.	Genes database, AKA Gene and Allele database	Contains descriptions of genes and alleles associated with morphological, developmental, and agronomically essential phenotypes, variants of physiological characters, biochemical functions and isozymes.	http://archive.gramene.org/rice_mutant/
7.	Ontologies	Provides information for structured controlled vocabularies (Ontologies) for the subsequent knowledge domains and their links to various objects such as QTL, phenotype gene, proteins, and Ensembl rice genes.	http://archive.gramene.org/plant_ontology/#to
8.	Comparative Maps (CMap)	A part that allows to view genetic, physical, sequence, and QTL maps for many species of cereal crops.	http://archive.gramene.org/cmap/
9.	Quantitative Trait Loci (QTL)	Comprises of QTL recognized for several agronomic traits in rice, maize, barley, oat, sorghum, pearl millet, foxtail millet, and wild rice.	http://archive.gramene.org/qtl/
10.	Species Pages	Allows access to various grass species that are now available in Gramene database	http://archive.gramene.org/species/

create gene trees and predict ortholog and paralog relationships among species. The most updated (as of Aug 10, 2015) database consists of about 56,387 GeneTree families that were constructed encompassing 1,280,368 individual genes (1,432,406 input proteins) from 39 plant genomes (and 5 nonplant outgroups) (http://www.gramene.org/release-notes-46). A synteny analysis pipeline is also implemented in Gramene that employs gene ortholog assignments from Compara GeneTree output as additional parameter to confirm homology. Synteny analysis allows investigators to deduce ancestral locations of genes, and the finding of conserved synteny offers a degree of assurance that genes are true orthologs (Youens-Clark et al. 2011).

Additionally, Gramene hosts metabolic pathway databases for ten plant species which include rice, sorghum, maize, brachy, arabidopsis, medicago, popular, coffea, tomato, and potato (Table 4). Three reference databases viz, EcoCyc, MetaCyc, and PlantCyc, are also available. By using these databases users can perform their own search or browse for genes, enzymes, metabolites, and metabolic pathways, to make cross-species comparisons, and to conduct analysis on their own specific data sets (http://www.gramene.org/pathways). A plant metabolic and regulatory pathways database known as Plant Reactome is also introduced by Gramene, which offers a Systems Biology Graphical Notation (SBGN)-based online user interface derived from the (Human) Reactome database model. These pathways, reactions, and gene entries present in Plant Reactome are cross-referenced to various well-known bioinformatics databases, such as UniProt, ChEBI, PubChem, PubMed, Gramene, and Plant Ensembl genomes, and Gene Ontology (GO) (http://www.gramene.org/pathways).

8 Kazusa Tomato Genomics Database (KaTomicsDB)

Availability: http://www.kazusa.or.jp/tomato/

KaTomicsDB is a portal website for tomato genomics and consists of the following three main databases:

8.1. *Tomato Marker Database* (http://marker.kazusa.or.jp/tomato/): Provides information on more than 8000 SNP and about 21,100 SSR markers, i.e., primer sequences and DNA fragments together with marker loci, genetic linkage maps of the DNA markers and genotyping data of the SNPs for 42 lines (Hirakawa et al. 2013, Shirasawa et al. 2010a, b). Additionally, by using similarity search methods, most of the markers have been mapped on the tomato genome, and ordered with the predicted genes (Shirasawa and Hirakawa 2013).

8.2. *Tomato Functional SNP Database* (http://plant1.kazusa.or.jp/tomato/): This database contains genotype data of more than 7000 SNPs loci in 40 tomato lines. The SNPs were grouped into six sets based on their locations in the genes predicted on the tomato genome sequence (Hirakawa et al. 2013). The genes with SNPs were annotated by similarity searches against the KOG (Tatusov et al. 2003), KEGG (Ogata et al. 1999), NR in NCBI (http://www.ncbi.nlm.nih.gov), TAIR10 (Garcia-Hernandez et al. 2002), and PDB (Berman et al. 2000) databases.

Table 4 Metabolic pathway databases for ten plant species

	Class						
Pathway	Pathways	Enzymatic reactions	Transport reactions	Polypeptides	Enzymes	Transporters	Compounds
RiceCyc	306	2103	87	47894	6040	603	1543
BrachyCyc	320	2057	87	26,633	7,723	950	1641
SorghumCyc	292	1838	9	36,347	10,636	269	1356
MaizeCyc	424	2132	106	39,655	8887	305	1453
AraCyc[a]	549	3463	45	9871	9923	328	2661
MedicCyc[a]	217	1498	1	4054	3426	33	1215
PoplarCyc[a]	440	2961	34	20,823	20,800	655	2247
PotatoCyc[a]	199	1079	1	20,743	1317	0	849
CoffeaCyc[a]	183	984	1	8168	541	0	781
LycoCyc[a]	456	2616	26	34,729	8033	344	1867
EcoCyc[a]	322	1693	392	4538	1518	263	2497
MetaCyc[a]	2150	11,671	550	11,570	9699	313	11,227
PlantCyc[a]	999	5458	86	189,505	189,397	6380	4663

Pathways in bold and italics represent reference pathways
[a]Mirror database. Not curated by the Gramene database

Furthermore, the locations of SNPs on the three-dimensional structures constructed by means of homology modeling can also be explored by the visiting researchers (Shirasawa and Hirakawa 2013).

8.3. *Tomato SBM DataBase* (http://www.kazusa.or.jp/tomato_sbm/): *s*elected *B*AC clone *m*ixture database denoted as SBM, represents two sets of BAC clone pools (SBM-I and SBM-II) that were generated by using the BAC end sequences from SOL Genomics Network (SGN), an online resource for the plants of Solanaceae family, which includes potato, eggplant, pepper, and tomato (Mueller et al. 2005). The total number of reads that were generated from the sequencing of these two sets is 4,248,000. The assembly of these sequences resulted in nonredundant sequences of 540,588,968 bp that consists of about 100,783 contigs (http://www.kazusa.or.jp/tomato_sbm/about.html). The sequence information of these SBM contigs can be retrieved through the free nucleotide databases such as DDBJ, Genbank, or EMBL under the accession numbers BABP01000001-BABP01100783.

In addition to the above mentioned databases, KaTomicsDB web portal also houses various downloadable datasets:

8.4. *KDRItomatoMutants2015:* A Variant Call Format (VCF) file that includes 5145 mutations of four ethyl methanesulfonate (EMS) and three gamma irradiation in the tomato Micro-Tom mutants (Shirasawa et al. 2015).
Availability: http://www.kazusa.or.jp/tomato/download/KDRItomatoMutants2015.vcf.gz

8.5. *KDRItomatoMicro-Tom2015:* A Variant Call Format (VCF) file that includes 1,140,687 spontaneous single nucleotide polymorphisms and indel polymorphisms in wild-type Micro-Tom lines (Shirasawa et al. 2015).
Availability: http://www.kazusa.or.jp/tomato/download/KDRItomatoMicro-Tom2015.vcf.gz

8.6. *KDRItomatoSNP2013:* This is a HapMap file that consists of 1247 SNPs genotyped with the Illumina GoldenGate array across 663 tomato accessions (Shirasawa et al. 2013).
Availability: ttp://www.kazusa.or.jp/tomato/download/KDRItomatoSNP2013.hmp.txt.gz

8.7. *KDRItomatoSNP2013:* A Variant Call Format (VCF) file that includes 1,473,798 SNP sites from resequencing data for six tomato lines (Shirasawa et al. 2013).
Availability: http://www.kazusa.or.jp/tomato/download/KDRItomatoSNP2013.vcf.gz

9 Parasitic Plant Genome Project (PPGP)

Availability: http://ppgp.huck.psu.edu/

The main objective of PPGP is to perform the comparative functional genomic analysis of parasitic plants so that the genome-wide variations which steered this parasitic lifestyle can be discovered as well as the modifications that stemmed as a

result of adoption of the parasitic life-style. In this project the transcriptomes of parasitic genera from Orobanchaceae (Triphysaria, Striga, and Orobanche) and two closely linked nonparasites (Lindenbergia philippensis, Orobancheae, and Mimulus; an independent sequencing project—http://www.mimulusevolution.org) will be compared. In spite of being evolutionarily related, these plant species are nevertheless span the range of parasitic capability from free-living to absolutely heterotrophic parasites. PPGP aims to sequence the developmental stage-specific cDNAs from Triphysaria versicolor, Striga hermonthica, and Orobanche aegyptiaca to obtain an in-depth sampling of EST sequences from each of these parasite species. The ability to develop haustoria, a distinct organ that forms the physical and physiological link among host and parasite, is a key feature of parasitism (Bandaranayake and Yoder 2013). It is because of the principal role of haustorium in parasitism, the sequencing efforts will focus on developmental stages from haustorial initiation to complete establishment of parasite on the host (http://ppgp.huck.psu.edu/).

PPGP project is a collaborative effort between the members from Virginia Tech, Penn State University, University of Virginia, and University of California, Davis among others. The project website (http://ppgp.huck.psu.edu/) is supported by the National Science Foundation (NSF: http://nsf.gov/) and hosted by Penn State University (http://www.psu.edu/). The project database can be searched by using a "keyword" or "GO Classification" from the search page (http://ppgp.huck.psu.edu/search.php). Additionally, the users can also blast search a query sequence against the generated libraries for the given species (http://ppgp.huck.psu.edu/blast.php). The raw sequencing data, assemblies, and legacy builds can be downloaded from http://ppgp.huck.psu.edu/download.php. All libraries are expected to be transcriptome sequences except wherever indicated.

10 Plant Genome DataBase Japan (PGDBj)

Availability: http://pgdbj.jp/?ln=en

PGDBj is a gateway website whose purpose is to incorporate plant genome-related information from various databases (DBs) as well as literature. The PGDBj portal comprises of three component databases and a cross search system, which offers a unified search over the contents of the databases (Asamizu et al. 2014). These three databases are:

10.1. *Ortholog DB:* this database contains information about orthologous genes, which was constructed on the basis of their corresponding amino acid sequence similarity. More than 500,000 amino acid sequences of 20 different Viridiplantae species were BLAST searched and clustered. Potential users may hunt for orthologs in diverse subsets of organisms by querying the database by means of either amino acid information or keywords (http://pgdbj.jp/en/ortholog-db.html). All the data from the current version (version 1.57.0) as of 28th September, 2015, in the form of FASTA sequences, tab separated orthologous information and cluster annotations is freely available (http://pgdbj.jp/en/ortholog-db/data-download.html).

10.2. *Plant Resource DB:* Integrates the SABRE (Systematic consolidation of Arabidopsis and other Botanical REsources) database (Fukami-Kobayashi et al. 2014), which provides cDNA and genome sequence resources amassed and maintained in the RIKEN BioResource Center and National BioResource Projects (http://pgdbj.jp/en/plant-resources.html).

10.3. *DNA Marker DB:* Provides manual or automatic curated information of DNA markers, quantitative trait loci and related linkage maps, from the literature and external databases (http://pgdbj.jp/en/dna-marker-linkage-map.html).

A distinct characteristic of the PGDBj is that the visitors have access to several plant genome databases through the Ortholog DB, serving as the main hub. The gene cluster information is valuable to speculate about gene families and evolutionary relations between genes across diverse species, thereby leading to the detection of novel genes and their function elucidation. An added feature of PGDBj is that it provides DNA marker and QTL data of important agronomic traits which were manually curated from the literature. The assimilation of such information will inspire the use of the PGDBj by investigators in the field, and the application of this data may speed up the harvest improvement method (Asamizu et al. 2014).

11 Plant Genome Duplication Database (PGDD)

Availability: http://chibba.agtec.uga.edu/duplication/

PGDD is an open access database that aims to detect and list plant genes in terms of intragenome or cross genome syntenic relationships. The present focus of PGDD is on flowering plants for which the whole genome sequences exist (Lee et al. 2013). As of now, PGDD contains genomic data for 47 plants comprising bryophytes, chlorophyta, and angiosperms. Since the last few years, PGDD has been providing data for syntenic relationships on the basis of colinear blocks among plants and has played a substantial role is diverse research areas such as evolution of gene families (Li et al. 2009), annotations (Watanabe et al. 2008), and polyploidy events (Barker et al. 2009).

The main page of PGDD displays a table covering basic information about all plants in the most up-to-date version that consists of plant name, genome version used, number of genes, original web link to download the data as well as main reference for the genome. Moreover, the table also provides related URLs, such as taxonomy information at NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi), so that users can effortlessly browse associated information for individual plants (Lee et al. 2013). To show gene collinearity, PGDD uses three important tools as follows:

11.1. *Dot-plot:* The colinear blocks between two plant genomes are shown as Dot-plots, so that the investigators can visualize the global view of all blocks. Each point on a dot-plot denotes a matched gene pair. Many options are available to adjust the plot by filtering subsets of gene pairs e.g. to display simply

a thin range of synonymous substitutions (Ks values) of gene pairs as a substitution to segregate the gene pairs by means of age (http://chibba.pgml.uga.edu/duplication/index/dotplot).

11.2. *Locus Search:* A colinear block consisting of an explicit locus can be found by using Locus-search, which can also be used to display the structure of this colinear block. Locus-search outputs the search results in the form of an alignment image and a tabulated list of genes. In the former case, PGDD shows genes in colinear blocks, so that any changes at the gene-level (e.g. insertion and deletion) can be determined without any difficulty. In the list of genes the inferred function of each gene along with the Ks and Ka values of the gene pair are shown, making it easier for the user to determine evolutionary distance and potential variations in function between genes (http://chibba.pgml.uga.edu/duplication/index/locus).

11.3. *Map-View:* By performing a BLAST search and then visualizing the hits, the Map-View program will attempt to hunt for regions where proximal (syntenic) hits can be observed, thereby, enabling the comparative mapping through different taxa. The database contains the predicted proteins from numerous plant genome projects that are listed on the PGDD main page. The program accepts FASTA formatted sequences as an input to carry out the BLAST search (http://chibba.pgml.uga.edu/duplication/index/blast).

In addition to the above mentioned tools, PGDD users can also download the different datasets for the blocks available in this database. The block files between two plants can be downloaded as a compressed (gunzip format) comma-separated file (CSV) by selecting any two plants in the provided combo box and clicking the "download" button. The downloaded file contains gene pairs in colinear blocks as well as Ka and Ks values of the pairs (http://chibba.pgml.uga.edu/duplication/index/downloads). Additionally, the DNA and protein sequences, and chromosomal locations for predicted gene models can also be retrieved (http://chibba.pgml.uga.edu/duplication/index/files).

12 Plant Genome and Systems Biology (PGSB)

Availability: http://pgsb.helmholtz-muenchen.de/plant/

The PGSB (http://pgsb.helmholtz-muenchen.de/plant/index.jsp) is web portal dedicated to the analysis of plant genomes by providing resources for bioinformatics analyses. The backbone of PGSB is PlantsDB (Nussbaumer et al. 2013), a database to store and manage the data for as many as 12 plant species (Table 5). Additionally, PlantsDB also offers a platform for integrative and comparative genomics research in plants. PGSB also provides several tools (summarized below) to query, explore, and analyze plant genome data in context of PlantsDB (http://pgsb.helmholtz-muenchen.de/plant/tools.jsp).

Table 5 List of plant species available at PGSB for genomics analysis

S. No.	Species (common name)	URL
1.	Solanum lycopersicum (tomato)	http://pgsb.helmholtz-muenchen.de/plant/tomato/index.jsp
2.	Medicago truncatula (barrel medic)	http://pgsb.helmholtz-muenchen.de/plant/medi3/index.jsp
3.	Arabidopsis thaliana (thale cress)	http://pgsb.helmholtz-muenchen.de/plant/athal/index.jsp
4.	Phoenix dactylifera (date palm)	http://pgsb.helmholtz-muenchen.de/plant/pdact/index.jsp
5.	Sorghum bicolor (sorghum)	http://pgsb.helmholtz-muenchen.de/plant/sorghum/index.jsp
6.	Zea mays (maize)	http://pgsb.helmholtz-muenchen.de/plant/maize/index.jsp
7.	Oryza sativa (rice)	http://pgsb.helmholtz-muenchen.de/plant/rice/index.jsp
8.	Brachypodium distachyon (purple false brome)	http://pgsb.helmholtz-muenchen.de/plant/brachypodium/index.jsp
9.	Lolium perenne (perennial ryegrass)	http://pgsb.helmholtz-muenchen.de/plant/lolium/index.jsp
10.	Hordeum vulgare (barley)	http://pgsb.helmholtz-muenchen.de/plant/barley/index.jsp
11.	Triticum aestivum (wheat)	http://pgsb.helmholtz-muenchen.de/plant/wheat/index.jsp
12.	Secale cereale (rye)	http://pgsb.helmholtz-muenchen.de/plant/rye/index.jsp

12.1. *The TransPLANT Genome Resources Registry:* transPLANT is a European-Union sponsored e-infrastructure to support genomic data analysis for plants and aims to fund coordination and research activities, and offer free access to computational resources for model plants and agriculturally important plant species (http://www.transplantdb.eu/). An archive of essential sequence-based resources for species of agricultural and commercial significance is maintained at the transPLANT Genome Resources Registry (http://pgsb.helmholtz-muenchen.de/plant/transplant/genomeResources.jsp).

12.2. *CrowsNest:* A Comparative Map Viewer at PGSB that uses a dynamic graphical interface to visualize and explore genome-wide chromosome organization as well as synteny between two or more plant genomes. CrowsNest can be used to make comparisons from the macro to the micro scale and allows discovering chromosome breakage and duplications. It can also be used to compare gene order, loss, deletion, or inversion among related species (http://pgsb.helmholtz-muenchen.de/plant/crowsNest/index.jsp).

12.3. *RNASeqExpressionBrowser:* An online tool used to search and visualize RNA-seq expression data (Nussbaumer et al. 2014a). It can create comprehensive reports for selected genes containing expression data as well as associated annotations. The current version also provides searching for differentially expressed and coexpressed genes (http://pgsb.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/index.jsp).

12.4. *chromoWIZ:* A web-based tool allowing visualization of genomic positions of important genes and comparing these records between various plant genomes (http://pgsb.helmholtz-muenchen.de/cgi-bin/db2/chromowiz/index.cgi?ID=chromoWIZ_MLFJQ6ZIZPBZ3LW). chromoWIZ can be searched by using gene identifiers, functional annotations, or sequence homology in four grass species (*Triticum aestivum*, *Hordeum vulgare*, *Brachypodium distachyon*, *Oryza sativa*) (Nussbaumer et al. 2014b).

12.5. *PGSB Repeat Element Database (PGSB-REdat) and Catalog (PGSB-REcat):* PGSB-REdat version (PGSB-REdat_v9.3p) currently contains about 62,000 sequences, with an emphasis on de novo identified LTR-retrotransposons (McCarthy and McDonald 2003) from grass (~37,000) and other (~9000) genomes hosted in PlantsDB. The repeat elements are categorized by PGSB-REcat, a hierarchical repeats grouping catalog, which expedites data extraction at different levels of detail.

13 Phytozome

Availability: http://phytozome.jgi.doe.gov/pz/portal.html

Phytozome is a comparative hub for plant genome and gene family data and analysis from the Department of Energy's Joint Genome Institute (JGI) (Goodstein et al. 2012). Phytozome offers a view of the evolutionary history of each plant gene at the sequence level, gene structure, and family as well as genome organization. The current version of Phytozome (v10.3.1) has 61 sequenced and annotated green plant genomes, 47 of which have been clustered into gene families at 12 evolutionarily significant nodes (http://phytozome.jgi.doe.gov/pz/portal.html). These families allow quick access to clade-specific orthology/paralogy relationships in addition to better understandings of clade-specific novelties and expansions. Attempts have been made to annotate each gene in the Phytozome with PFAM (Finn et al. 2014), KOG (Koonin et al. 2004), KEGG (Kanehisa and Goto 2000), PANTHER (Mi and Thomas 2009) and GO (Ashburner et al. 2000) assignments wherever applicable.

Data can be accessed by using Phytozome's PhytoMine (an InterMine interface to data from Phytozome: http://phytozome.jgi.doe.gov/phytomine/begin.do) and several popular open source components such as GBrowse (Stein et al. 2002), Jalview (Waterhouse et al. 2009), BioMart (Smedley et al. 2009), mView (Brown et al. 1998) with custom visualization code for gene family search, inspection and evaluation. Genes, and gene families can also be retrieved from Phytozome by using both keyword and sequence similarity-based search methods. BLAST and BLAT searches of organism genomes, proteomes, and gene family consensus sequences, can be carried out to discover the genomic regions, gene transcripts, peptides, and gene families matching the initial query sequence. Whether the genes and gene families are found through keyword or sequence similarity based searches, they can be observed separately or pooled dynamically to generate composite families, before being viewed and analyzed (Goodstein et al. 2012).

14 Plant Repeat Databases

Availability: http://plantrepeats.plantbiology.msu.edu/index.html

In plants, a considerable percentage of the genome constitutes the repetitive sequences which in turn can hamper the sequencing and genome annotation efforts. The Plant Repeat Databases (http://plantrepeats.plantbiology.msu.edu/index.html) were constructed to help in the accumulation and discovery of repeat sequences in plant genomes. Currently, the databases consists repetitive sequences from more than 10 plant genera that includes *Arabidopsis*, *Brassica*, *Glycine*, *Hordeum*, *Lotus*, *Medicago*, *Oryza*, *Solanum*, *Sorghum*, *Triticum,* and *Zea* (http://plantrepeats.plantbiology.msu.edu/composition.html). These repetitive sequences were grouped as super-classes, classes, and subclasses on the basis of structure and sequence composition (Ouyang and Buell 2004). The super-class transposable elements (TEs) comprises retrotransposons, transposons, and miniature inverted-repeat transposable elements (MITEs) (Feschotte et al. 2002). Plant centromeres are composed of regions of tandemly repeated sequences (satellite repeats), which are intermingled with other repetitive sequences, for instance centromeric- and pericentromeric-specific retrotransposons (Nagaki et al. 2003; Copenhaver et al. 1999). Telomeric sequences are represented by telomere repeat sequence and telomere-associated sequences (McKnight et al. 1997). Additionally, the final super-class of repetitive sequences is classified as rDNAs that encode the structural RNA elements of ribosomes (Shishido et al. 2000).

The repeat database was constructed by querying the repetitive DNA sequences of selected plant genera from GenBank and additional available records. Any duplicate and vector sequences were removed and the repeats were grouped into five super-classes (Transposable elements, Centromere-related, Telomere-related, rDNA, and Unclassified) which were further broken down into major classes of repeats. The assembled repetitive sequences contained by the similar plant family were pooled into a repeat database for the plant family (Ouyang and Buell 2004). Table 6 summarizes the most updated statistics of the Plant Repeat Databases. Users can query any of the genera-specific databases that are available at the portal by using a BLAST search (http://plantrepeats.plantbiology.msu.edu/search.html). Additionally, all these repeat sequence databases can also be downloaded as a whole or a subset based on a repeat class or repeat source as flat files (http://plantrepeats.plantbiology.msu.edu/downloads.html)

Table 6 Statistics of the repetitive sequences within each plant repeat databases

Family	Number	Total length (kb)
Gramineae (Hordeum, Oryza, Sorghum, Triticum, and Zea)	5261	8637.9
Fabaceae (Glycine, Lotus, and Medicago)	308	292.6
Brassicaceae (Arabidopsis, Brassica)	775	449.4
Solanaceae	371	215.9
Total	6715	

15 PLAZA

Availability: http://plaza.psb.ugent.be/

The main goal of PLAZA 3.0 (Proost et al. 2015) is to allow comparative genomics data accessible for plants through a user-friendly online interface. At PLAZA 3.0, exhaustive information about genome organization can effortlessly be queried and visualized in addition to details about structural and functional annotation, gene families, protein domains, and phylogenetic trees. In comparison to the first release of PLAZA (Proost et al. 2009) which included nine organisms, the current version now consists of 47 (31 dicots and 16 monocots) plant species (Proost et al. 2015). Table 7 summarizes the various statistics of PLAZA 3.0. The presence of these new species provides an extensive phylogenetic variety along with an additional thorough sampling of specific clades. The functional annotation has been upgraded with the inclusion of data from sources such as Gene Ontology, MapMan, UniProtKB, PlnTFDB, and PlantTFDB. Moreover, with the improved algorithms, functional annotation from well-characterized plant genomes can be transferred to other species. These advancements in the PLAZA 3.0 make it a resourceful and comprehensible reserve for users who aim to explore genomic data to investigate diverse characteristics of plant biology. PLAZA 3.0 also offers a large number of analysis tools divided into four sections as follows:

15.1. *Gene Families*

15.1.1. *Expansion Plot:* It is used to discover the copy-number gene family variation between two groups of species. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/gene_families/expansion_plot).

15.1.2. *Gene Family Finder:* This tool allows identifying extended gene families explicit to one or more species. The Gene Family Finder tool can be employed to carry out two distinct search operations: to identify species/clade-specific gene families and to find species/clade-expanded gene families. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/gene_families/findtool).

15.2. *Colinearity*

15.2.1. *Ks-graphs:* This tool can be used to plot histograms from Ks values for all colinear gene pairs within one or from two different species. Multiple graphs can be superimposed on each other by using Ks-graphs. The Ks values can be compared from duplication events in

Table 7 Summary of Dicots and Monocots genomic data present in the PLAZA 3.0

Type	No. of species	Genes	Coding	Multigene gene families	Phylogenetic trees
Dicots	31	1,087,713	1,012,693	26,192	17,232
Monocots	16	537,114	503,231	19,612	13,632

different organisms or Ks values from speciation events. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/ks)

15.2.2. *Skyline plot:* The Skyline plot provides a summary of the colinear regions that occur within a set of selected species, i.e. regions between selected species having conserved gene content and order. This region is selected by means of a reference gene that must always be provided. The plot finds all multiplicons that match the reference chromosome, and also checks for each organism on the number of segments that can be found together with the reference segments in one multiplicon. Therefore, Skyline plot can be used to investigate the quantity of recognizable duplications in the evolutionary history of organisms.
(http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/collinearity/index)

15.2.3. *Synteny plot:* These plots report the local gene organization for homologous genes within a family. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/synteny/index)

15.2.4. *WGDotplot:* Reports all colinear regions between and/or within species using a pairwise approach. WGDotplot comes in two different versions: the standard version (which uses an HTML clickable map and supports the display of colinear regions between two species) and the Java Applet version (the interactive Applet version that has no limitations on the number of species). (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/dotplot/index)

15.3. *Localization*

15.3.1. *Functional Clusters:* This tool shows the identified clusters of functionally related genes on a per chromosome basis. These clusters are detected by using the CHunter (Yi et al. 2007) tool. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/functionalcluster/overview)

15.3.2. *WGMapping tool:* The Whole Genome Mapping (WGMapping) tool can be used to display the organization of a set of genes on all chromosomes of a particular species. The WGMapping tool has two different approaches of operation. In the first such approach, all genes are selected and presented on the chromosomes of the species, and then the gene type is displayed. In the second mode the user outlines a gene set (by using GO label, InterPro domain, Reactome pathway, gene family or workbench), and then these genes are shown on their corresponding positions on the chromosomes.
(http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/genome_mapping/index)

15.4. *Other*

15.4.1. *BLAST:* Users can also perform a BLAST search against the PLAZA database or an individual plant species database available at PLAZA. (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v3_dicots/blast/index)

15.4.2. *Workbench:* The workbench provides a portal to analyze multiple genes in batch through user-defined gene sets. All analysis carried out using workbench is private and requires password protected user registration. (http://bioinformatics.psb.ugent.be/knowledge/wiki-plaza/workbench)

16 Plant Expression Database (PLEXdb)

Availability: http://www.plexdb.org/index.php

PLEXdb is a combined gene expression database for plants and plant pathogens bridging between genotype to phenotype through transcript profiling (Dash et al. 2012). PLEXdb assimilates several data sets from a widespread range of plant and plant pathogen microarrays and offers a single resource to access, analyze, and disseminate expression data for broad comparative functional genomics investigations (Shen et al. 2005; Wise et al. 2007). The main objectives of PLEXdb is to make this data easily available to aid users address biological problems in question, and provide integration of data and tools that are presently available only from different resources. These incorporated databases and tools of PLEXdb allow researchers to use commonalities in plant biology for a comparative attitude to functional genomics by using extensive expression profiling data sets (Dash et al. 2012). PLEXdb consists of the following main sections:

16.1. *Plants (Plant Arrays):* An expression resource for plant microarrays, annotations and datasets. Contains MIAME (Minimum Information About a Microarray Experiment)/Plant-compliant and Plant Ontology enhanced expression database for both monocots (5 species) along with dicots (8 species) (http://www.plexdb.org//modules/PD_general/plants_list.php).

16.2. *Pathogens (PathoPLEX):* Arrays and resources for plant pathogens and includes MIAME/Plant-compliant and Plant Ontology enhanced database for plant pathogens as well as symbionts. PathoPLEX offers arrays for nine phytopathogenic fungi (Fusarium species), oomycetes, nematodes, and symbiotic bacteria (http://www.plexdb.org//modules/PD_general/pathogens_list.php).

16.3. *Expression Atlases:* Can be used to explore the gene expression in various organs, tissues, or developmental stages. Expression Atlases is divided into two sections—dicots and monocots (http://www.plexdb.org/modules/PD_general/atlas.php).

16.4. *Gene List Suite:* A gene list is a set of probe set names that could be imported directly or created by a range of analyses on microarrays or microarray experiments. For instance: the entire probe sets on a microarray assigned to a specific GO term or the list of differentially expressed genes (DEGs) in a particular experiment (http://www.plexdb.org//modules/glSuite/gl_main.php).

16.5. *Model Genome Interrogator (MGI):* By using MGI (http://www.plexdb.org//modules/MGI/), a user can input a list of genes from GeneChips to mine location and sequence data from model genomes (currently, Rice and Arabidopsis). MGI predicts homologies, shows gene structures and secondary information for annotated genes and full-length cDNAs. It also retrieves corresponding sequences, and offers direct links to various genome browsers such as GRAMENE (Ware et al. 2002), TAIR (Rhee et al. 2003), Rice Genome Browser (http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/?name=LOC_Os09g29170.4).

16.6. *Gene OscilloScope:* Allows users to search for experiments where expression of queried genes varies (oscillate) to a greater extent. If the expression of a probe set (gene) is affected by some of the treatments in an experiment, it shows a higher Coefficient of Variation (CV) i.e. more variation where as if the expression is less affected by the treatments, a lower CV is observed (negligible variation) (http://www.plexdb.org//modules/tools/genoscope/genoscope.php).

16.7. *Fluctuation Filter:* Allows users to search for the probe sets/genes whose fluctuations in expression (CV) matches their specifications (http://www.plexdb.org//modules/tools/datamine/fluctuation_filter_v2.php).

16.8. *PLEXdb Blast:* It allows users to map fasta sequences or probe sets to microarrays in PLEXdb using BLAST (Altschul et al. 1990) (http://www.plexdb.org//modules/tools/plexdb_blast.php).

16.9. *Search Experiments:* Search for experiments using keywords in title, treatments, experimental factors, array design, organism, etc. (http://www.plexdb.org//modules/PD_browse/queryExperiments.php).

References

Allen AM, Barker GL, Berry ST, Coghill JA, Gwilliam R et al (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). Plant Biotechnol J 9:1086–1099

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Amborella Genome Project (2013) The Amborella genome and the evolution of flowering plants. Science 342:1241089

Arnaoudova EG, Bowens PJ, Chui RG, Dinkins RD, Hesse U et al (2009) Visualizing and sharing results in bioinformatics projects: GBrowse and GenBank exports. BMC Bioinformatics 10:A4

Asamizu E, Ichihara H, Nakaya A, Nakamura Y, Hirakawa H et al (2014) Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases. Plant Cell Physiol 55, e8

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31:248–250
Bal S, Reddy KP (2014) Plant Genomic Databases for Oilseeds Crop Improvement. IJCB 3:43–47
Bandaranayake PCG, Yoder JI (2013) Haustorium initiation and early development, Chapter 4. In: Joel DM, Gressel J, Musselman LJ (eds) Parasitic orobanchaceae—parasitic mechanisms and control strategies. Springer, Heidelberg, pp 61–74
Barker MS, Vogel H, Schranz ME (2009) Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol Evol 1:391–399
Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
Braun P, Carvunis AR, Charloteaux B, Dreze M, Ecker JR et al (2011) Evidence for network evolution in an Arabidopsis interactome map. Science 333:601–607
Brown NP, Leroy C, Sander C (1998) MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14:380–381
Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S et al (1999) Genetic definition and sequence analysis of Arabidopsis centromeres. Science 286:2468–2474
Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res 40:D1194–D1201
Donlin MJ (2007) Using the generic genome browser (GBrowse). Curr Protoc Bioinformatics Chapter 9:Unit 9.9
Erwin TA, Jewell EG, Love CG, Lim GA, Li X et al (2007) BASC: an integrated bioinformatics system for Brassica research. Nucleic Acids Res 35:D870–D873
Evans HC (2007) Cacao diseases-the trilogy revisited. Phytopathology 97:1640–1643
Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341
Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230
Flicek P, Amode MR, Barrell D, Beal K, Brent S et al (2011) Ensembl 2011. Nucleic Acids Res 39:D800–D806
Fukami-Kobayashi K, Nakamura Y, Tamura T, Kobayashi M (2014) SABRE2: a database connecting plant EST/full-length cDNA clones with Arabidopsis information. Plant Cell Physiol 55, e5
Garcia-Hernandez M, Berardini TZ, Chen G, Crist D, Doyle A et al (2002) TAIR: a resource for integrated Arabidopsis data. Funct Integr Genomics 2:239–253
Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH et al (2007) A predicted interactome for Arabidopsis. Plant Physiol 145:317–329
Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186
Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University
Hebbar PK (2007) Cacao diseases: a global perspective from an industry point of view. Phytopathology 97:1658–1663
Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K et al (2013) Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA Res 20:221–233
Ilic K, Berleth T, Provart NJ (2004) BlastDigester--a web-based program for efficient CAPS marker design. Trends Genet 20:280–283, Erratum in: Trends Genet 2005 21:36
International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800
Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–664
Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, et al (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J 50:347–363
Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7
Lai K, Lorenc MT, Edwards D (2012) Genomic databases for crop improvement. Agronomy 2:62–73
Lee TH, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res 41:D1152–D1158
Li W, Liu B, Yu L, Feng D, Wang H et al (2009) Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants. BMC Evol Biol 9:90
Love CG, Robinson AJ, Lim GA, Hopkins CJ, Batley J et al (2005) Brassica ASTRA: an integrated database for Brassica genomic research. Nucleic Acids Res 33:D656–D659
McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367
McCouch SR, Paul E (1993) RiceGenes, an International Genome Database and Bulletin Board for Rice. DNA Link 3:40–41
McKnight TD, Fitzgerald MS, Shippen DE (1997) Plant telomeres and telomerases. A Rev Biochem (Mosc) 62:1224–1231
Mi H, Thomas P (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol 563:123–140
Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R et al (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol 138:1310–1317
Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q et al (2003) Molecular and cytological analyses of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163:759–770
NussbaumerT,KuglerKG,BaderKC,SharmaS,SeidelMetal(2014a)RNASeqExpressionBrowser--a web interface to browse and visualize high-throughput expression data. Bioinformatics 30:2519–2520
Nussbaumer T, Kugler KG, Schweiger W, Bader KC, Gundlach H et al (2014b) chromoWIZ: a web tool to query and visualize chromosome-anchored genes from cereal and model genomes. BMC Plant Biol 14:348
Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC et al (2013) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res 41:D1144–D1151
Ogata H, Goto S, Sato K, Fujibuchi W, Bono H et al (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27:29–34
Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363
Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556
Popescu SC, Popescu GV, Bachan S, Zhang Z, Seay M et al (2007) Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. Proc Natl Acad Sci U S A 104:4730–4735
Popescu SC, Popescu GV, Bachan S, Zhang Z, Gerstein M et al (2009) MAPK target networks in Arabidopsis thaliana revealed using functional protein microarrays. Genes Dev 23:80–92
Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T et al (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell 21:3718–3731
Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D et al (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43:D974–D981

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 31:224–228

Ronald PC (2014) Lab to farm: applying research on plant genetics and genomics to crop improvement. PLoS Biol 12, e1001878

Saski CA, Feltus FA, Staton ME, Blackmon BP, Ficklin SP et al (2011) A genetically anchored physical framework for Theobroma cacao cv. Matina 1-6. BMC Genomics 12:413

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37:501–506

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115, Erratum in: Science 2012 337:1040

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R et al (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–107, Erratum in: Genome Res 14:786

Shen L, Gong J, Caldo RA, Nettleton D, Cook D et al (2005) BarleyBase--an expression profiling database for plant genomics. Nucleic Acids Res 33:D614–D618

Shirasawa K, Asamizu E, Fukuoka H, Ohyama A, Sato S, et al (2010a) An interspecific linkage map of SSR and intronic polymorphism markers in tomato. Theor Appl Genet 121:731–739

Shirasawa K, Isobe S, Hirakawa H, Asamizu E, Fukuoka H, et al (2010b) SNP discovery and linkage map construction in cultivated tomato. DNA Res 17:381–391

Shirasawa K, Hirakawa H (2013) DNA marker applications to molecular genetics and genomics in tomato. Breed Sci 63:21–30

Shirasawa K, Fukuoka H, Matsunaga H, Kobayashi Y, Kobayashi I et al (2013) Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. DNA Res 20:593–603

Shirasawa K, Hirakawa H, Nunome T, Tabata S, Isobe S (2015) Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato. Plant Biotechnol J 14(1):51–60

Shishido R, Sano Y, Fukui K (2000) Ribosomal DNAs: an exception to the conservation of gene order in rice genomes. Mol Gen Genet 263:586–591

Smedley D, Haider S, Ballester B, Holland R, London D et al (2009) BioMart--biological queries made easy. BMC Genomics 10:22

Stein LD, Thierry-Mieg J (1998) Scriptable access to the Caenorhabditis elegans genome sequence and other ACEDB databases. Genome Res 8:1308–1315

Stein LD, Mungall C, Shu S, Caudy M, Mangone M et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J 43:153–163

Ware D, Jaiswal P, Ni J, Pan X, Chang K et al (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res 30:103–105

Watanabe M, Mochida K, Kato T, Tabata S, Yoshimoto N et al (2008) Comparative genomics and reverse genetics analysis reveal indispensable functions of the serine acetyltransferase gene family in Arabidopsis. Plant Cell 20:2484–2496

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wilkinson PA, Winfield MO, Barker GL, Allen AM, Burridge A et al (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. BMC Bioinformatics 13:219

Wise RP, Caldo RA, Hong L, Shen L, Cannon E et al (2007) BarleyBase/PLEXdb. Methods Mol Biol 406:347–363

Yi G, Sze SH, Thon MR (2007) Identifying clusters of functionally related genes in genomes. Bioinformatics 23:1053–1060

Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G (2011) Gramene database in 2010: updates and extensions. Nucleic Acids Res 39:D1085–D1094

QTL Analysis in Plants: Ancient and Modern Perspectives

Muhammad Jamil, Aamir Ali, Khalid Farooq Akbar, Abdul Aziz Napar, Alvina Gul, and A. Mujeeb-Kazi

Contents

M. Jamil (✉) • A. Ali
Department of Botany, University of Sargodha, Sargodha, Pakistan
e-mail: jshahid80@yahoo.com

K.F. Akbar
Department of Botany, Govt. Post-Graduate College, Sahiwal, Pakistan

A.A. Napar
Department of Plant Sciences, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan

A. Gul (✉)
Atta-ur-Rehman School of Applied Biosciences (ASAB), National University of Science and Technology (NUST), Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

A. Mujeeb-Kazi
Wheat Wide Crosses Program, National Agricultural Research Center (NARC), Islamabad, Pakistan

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_3

Abstract Quantitative traits exhibit continuous variation, indicating their control through multiple genes. Segregating populations are used to mine out associations between phenotypic and genotypic variations. Phenotyping performed for a specific trait and its variation in the population is justified with genotypic variation obtained through genetic markers application. A snapshot of genotypic variation is strictly dependent on the number and density of the markers applied. Parental and marker information is required to correlate genetic and phenotypic data for quantitative trait loci (QTL) analysis. For many years (now becoming obsolete), it has been of core importance to identify QTL with such methodology. Failure had to be faced by the researcher because the DNA region identified for phenotypic variation was much wider, and needed to be narrowed down by further dense marker application in that area to obtain required and accurate results. Nowadays the focus is on high-throughput technologies to obtain genome-wide resolution: high-throughput sequencing (HTS) is one of them. A comprehensive map of genomic variations can be produced with resequencing or reference genome sequences. Along with expression profiling, new molecular markers can be searched out with QTL analysis. Genomic-assisted breeding by studying the evolutionary variations in crops has many applied aspects as well. As compared to the conventional biparental population, presently the focus is on raising multiparent advanced generation inter-cross (MAGIC) populations to explore the genetic basis of quantitative traits. Probabilities of alleles of interest across the whole genome are calculated through the Hidden Markov Model (HMM). Different software packages (such as R-package, Qgene) are used for the estimates. Such whole-genome approaches in QTL analysis are a powerful and recently used technique. In this chapter, all these recent and modified modern techniques are reviewed with the most recent upcoming details. Traditional and modern QTL analyses have clearly been differentiated on applicable grounds.

Keywords QTL • Genotyping • Phenotyping • Mapping • Next-generation sequencing

1 Introduction

Plant breeding is the core area to develop genetic variations by which required traits are incorporated through selection. Traits of interest such as yield and biotic and abiotic stress resistances often are under the influence of more than one gene. Hence, to unravel the segregation pattern of such polygenic inheritance is of vital importance. The phenotype exhibited by such traits might be the aggregated action of many genes and the environment. Such assessable phenotypes have a continuous distribution pattern among individuals. The segregation pattern of such traits has previously been studied through simple statistical tools. At that time, by the involvement of the molecular markers (such as RFLPs, RAPDs, or SSRs) and visual measurement it became possible to have two types of expression (genotypic and phenotypic) of the examined individuals. Polymorphism shown by the molecular markers (genotypic variation) and through the recorded phenotypic variations compelled researchers to detect the association of genotypic variation with phenotypic patterns. To probe the segregation of required polygenic traits, biparental populations with a high number of individuals were developed. Before going into further details of such phenomena, we should know more about quantitative trait loci (QTL).

1.1 Quantitative Trait Loci

Asins (2002) reviewed that concepts of quantitative trait loci (QTL) detection had been developed from the work of Sax (1923). The acronym QTL was first coined by Geldermann (1975), reviewed by Slate (2005). The region of DNA responsible for influencing a trait that is recorded on a linear (continuous) scale is called a QTL. The expression of a quantitative trait is regulated by hundreds or even thousands of such QTLs (Mackay et al. 2009). On the basis of DNA markers positioned on a linkage map, QTLs are allotted on a chromosome in the vicinity where the statistical probability is significant. As DNA markers are not affected by the environment, after detecting their polymorphism, these can be used as a tool in mapping QTL. Quantitative traits can have varying phenotypic concentration depending upon the allelic diversity at a QTL region, and the functional markers found associated with these QTLs established the importance of QTL analysis. Thus, polygenic traits that could hardly be analyzed by the utilization of customary breeding methods could easily be labeled with DNA markers.

1.2 Essentiality of QTL Analysis

Because QTL is the region flanked by two markers (Erickson et al. 2004), it is obligatory to detect the linkage between the marker and QTL. Here mapping becomes useful, to arrange the markers, genes, or QTLs in a sequence on the

chromosome, highlighting the relative distance among them (Touré et al. 2000). When genes or QTLs linked with traits of interest are to be detected, it necessitates the construction of such maps. Without finding the association between the trait of interest and QTL, it is hard to avail the genetic diversity. By increasing the DNA marker density on the chromosome, a detailed genetic map can be produced. These maps created the importance of present-day QTL mapping (Doerge 2002). Narrowing down the distance (by increasing the number) between the markers and the QTL, a stronger linkage between marker and trait can thus be detected. The stronger the marker–trait linkage, the more authenticated the usage will be. With the help of DNA markers, it seems very important to detect the QTL linked with the trait of interest if we want to utilize that character in further breeding strategies.

1.3 Principle of QTL Analysis

QTL analysis is devised on the principle that genes and markers which segregate during meiosis, if tightly linked, must be transmitted together from parent to progeny (Collard et al. 2005). As a quantitative trait is the expression of many genes at the same time, there must be a region or locus (QTL) that if found linked with markers can thus be analyzed for further benefits. The development of a segregating population first and then the detection of marker–trait association with the help of genetic and phenotypic profiles are basic components of QTL analysis.

2 Methodology Involved

The science of quantitative genetics has been predominantly occupied by biometric mathematics. Sophisticated statistical tools are involved to extract and correlate the variation in genotypic and phenotypic diversity among individuals. QTL analysis can be performed if we have the following information:

1. A model segregating the population in which a QTL for the required trait is to be detected.
2. Genetic dissection of the population with markers.
3. A record of phenotypic variations for the trait of interest.
4. Software packages to depict marker–trait association.

2.1 Mapping Population

Breeding populations differ from natural populations because they are selected according to the breeders' interests. Based on required traits, the genetic properties of breeding population are highly confined and focused. All breeding disciplines

obey a general pattern of creating new genotypic variations. Crossing the lines with required traits has a high probability of detecting a QTL (Würschum 2012). To study the segregation of any polygenic trait of interest, parent selection is very crucial. Parents must be phenotypically evaluated and should have contrast in trait expression; for example, P1 (disease resistant) and P2 (disease susceptible). In self-pollinated species, the mapping population should be initiated from highly homozygous (inbred) parents (Collard et al. 2005). In cross-pollinated species, the F_1 generation can be developed by pair-crossing of heterozygous parent plants that are significantly different for required traits (Barrett et al. 2004). F_2 populations from F_1 hybrids, backcross-derived lines, are the usual types most often programmed for self-pollinated species and can easily be developed in a short time. Recombinant inbred lines (RILs) and doubled haploid (DH) lines are also developed. RILs and DH lines are used if homozygous lines are to be increased without any particular genetic alteration (Collard et al. 2005). In QTL mapping, construction of a mapping population must have a strategy of creating a correlation between the strength of linkage and the degree of linkage disequilibrium (Gardner and Latta 2007). Linkage disequilibrium (LD) arises when an allele at locus A is nonrandomly associated with the allele at locus B. It can befall when these two loci are unlinked (Flint-Garcia et al. 2003) (Fig.1).

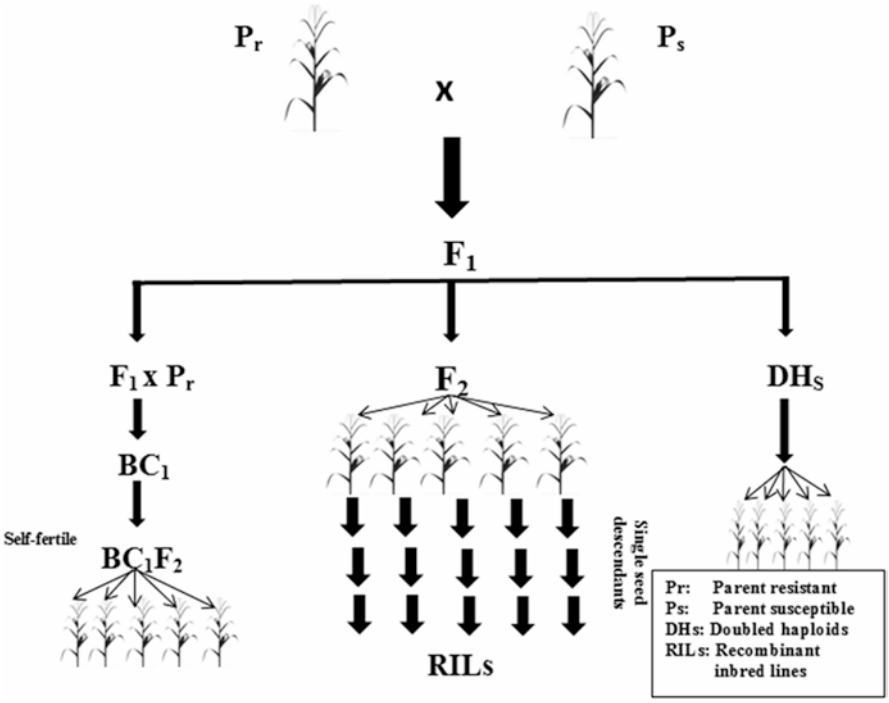


Fig. 1 Usual types of mapping population for self-pollinating species

Based on the biparental and diverse panel, there are two basic mapping approaches, that is, family mapping and population mapping. Family mapping detects only a limited number of alleles per locus at one time. Population mapping involves a diverse panel of genotypes with multiple families also, and each family with a small family size (Myles et al. 2009). The type of the population that should be used depends on the plant species, type of markers used, and the trait to be mapped (Touré et al. 2000).

2.2 Genotyping

Screening of a population along with the parents with the help of DNA markers (polymorphic) to obtain such a diversity pattern resulting from polymorphism of the markers is called genotyping (Collard et al. 2005). In the pregenomic era, populations used to be screened with a few markers of any type [restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs), etc.], depending upon the suitability and the nature of trait segregation. The genotypic profile obtained by the marker analysis contains the number of markers used and the polymorphism shown in the population including parents. All the previous and present-day genotyping techniques have the same purpose: how fast and how many of the markers can be processed quickly. The objective of genotyping has always been to have the polymorphism indicated with few base differences underlying allelic diversity. Current aspects of genotyping are discussed in this chapter under the heading of modern perspectives. Instead of extracting and analyzing DNA from every individual of a segregating population, bulk segregation analysis can also be performed. Four DNA bulks, two from individuals of extreme phenotypes (e.g., highly susceptible and highly resistant) along with two parents are prepared. For this purpose, we need to scan the genotypes with intensive application of markers (Cheng and Chen 2010) (Fig. 2).

2.3 Phenotyping

To dissect a trait, the genotypic variation pattern shown by markers as well as phenotypic diversity display are needed. A quantitative trait that is expressed in a continuous distribution pattern is scored, and the entire population and parents are screened. This method is foremost to detect a QTL when phenotypic data must be available. The data are usually obtained by combining multiple experiments but comes with unbalanced inferences (Würschum 2012), whereas balanced data sets are found beneficial in minimizing false-positive QTLs (Wang et al. 2012). Even then the phenotypic data generated without prior balanced experimental design can also be used for QTL detection.

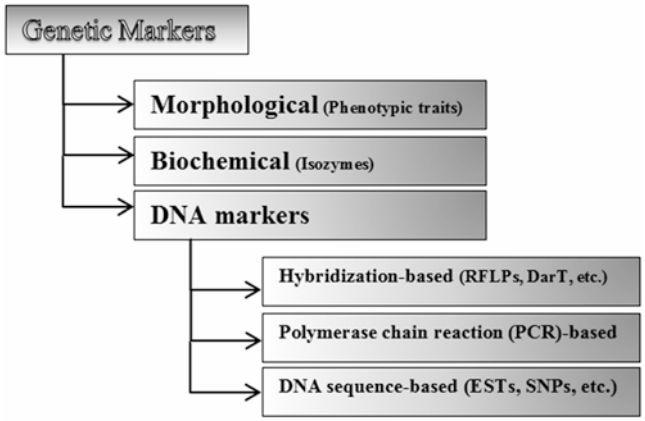


Fig. 2 Flowsheet presentation of genetic markers

Phenotyping intensity is also a notable factor for a precise QTL analysis. High heritability and low bias in the measurements are prerequisites for authenticated QTL detection (Bradbury et al. 2011; Liu et al. 2012). The obtained phenotypic profile will contain the number of individuals in a population and parents and the variation in trait expression among them. With the combination of modern approaches in phenotyping (discussed later in the chapter) and statistical tools, it has become convenient to have a more detailed and accurate pattern of phenotypic variation in the studied trait.

2.4 Software Used

After obtaining the genotypic and phenotypic picture of the mapping population, statistical tools come into use. Analysis of variance for the studied phenotypic trait in the population is mandatory. Sorting out linked loci and their strength of linkage with the phenotypic variation becomes vital. Without the aid of software packages, it seems impossible to handle the data produced by such extensive phenotypic and genotypic observations. These packages need input files that may be opened as any spreadsheet (Excel, SPSS, Statistica, SAS, Statistix, XLStats, etc.) software also. Results of marker applications and visual scoring are arranged in spreadsheet files and then formatted as the software requirements. The file format needed by the software can be found in template files of the software help manual. Most such software can be downloaded free from their respective websites along with their user manual and help files. Software that is often used in QTL mapping includes the following.

2.4.1 QTL Cartographer

QTL Cartographer is used for single-marker regression and interval mapping. It can analyze the data set obtained from F_2, inbred lines, and also from backcross-derived populations (Luciano et al. 2012). A useful software then results to be expressed in graphs. Various QTL models can be explored by generating simulation data and varying parameter settings. Rmap input and output files are used for creating linkage maps. Data files are Rcross input files and input files for QTL information are in Rqtl format in version 1.17 (Basten et al. 2004). An updated version 2.5, along with a user manual, is now also available (Wang et al. 2013).

2.4.2 MQTL

When the data set is from multiple environments, homozygous biparental progeny (recombinant inbred lines, doubled haploid lines), and the mapping is to be simple interval mapping or a simplified form of composite interval mapping, then MQTL software is the best choice. This software is specialized to handle large data sets primarily (Tinker and Mather 1995).

Nowadays there is *MetaQTL*, a Java package designed to analyze the combined data from different gene mapping experiments such as molecular markers, QTL, and candidate genes. This package is the assembly of various Java written programs executing different purposes (Veyrieras et al. 2007).

2.4.3 MapQTL

When we are concerned with experimental population of BC1, F_2, RIL, DH, and an outbred full-sib family in diploid species, then MapQTL V. 6 proves itself a user-friendly choice. For easier comparison of results from advanced backcross inbred lines, advanced intermated inbred lines, and doubled haploids derived from F_2, this software can be used. Combined analysis of multiple populations with simple experimental design along with covariance by selecting an automatic marker cofactor in a single project is the key feature of this software. Extended options for QTL chart presentation and adjustable data exportable tables are additional functions (Van Ooijen et al. 2000; Van Ooijen 2004; Van Ooijen and Kyazma 2009).

2.4.4 Joinmap

Estimation about linkage groups is the most technical task in QTL mapping. This software enables the user to study linkage group formation depending on independence test logarithm of odds (LOD) score, linkage LOD score, independence test *P* value, and recombination frequency. A linkage map can be constructed after fixing the linkage groups. With increased key components, high-quality charts to

express the maps up to the required preferences can be prepared. All the map charts can be exported to pdf format and can also be copied to MS-Word or Excel; moreover, the charts can be printed easily (Van Ooijen 2006).

2.4.5 Map Manager

Older versions of the software were Map Manager Classic, Map Manager QT, Map Manager XP, and the latest among them was Map Manager QTX (Manly et al. 2001). Map Data set from the dominant markers can also be manipulated by Manager QTX. It is modified and equipped with cross-platform libraries and designed for multiple computer platforms.

2.4.6 QGene

This software was reported by Nelson (1997) for the analysis of marker-based large amounts of genomic information in which raw genetic markers were reduced to numerical summary statistics along with prompt graphic display of both data and statistics. This software can now be downloaded from the website www.qgene.org along with its user manual. It can handle large amounts of the genotypic and phenotypic data obtained from F_2, F_3 families, BCF1, DHs, and RILs as well. A simple notepad file is prepared with marker data first; then trait data below; along with a Java Development Kit (JDK) extension. The detailed procedure is given in the user manual, and a sample data file is available with the software (Joehanes and Nelson 2008).

2.4.7 SAS

If mapping is to be performed only with single-marker analysis, then SAS is used. It can identify QTLs by detecting associations between marker genotype and phenotype of the quantitative trait. Analysis of variance, *t* test, general linear model, and regression analysis can also be performed with this package (Akbarpour et al. 2014; Rahman et al. 2014; Zambrano et al. 2014) (Table 1).

2.5 Interpreting Results

The birth of quantitative genetics was derived from the fusion of Mendelism and biometry. The combination of molecular genetic techniques and powerful statistical methods enables the researcher to dissect the complicated quantitative traits (Mauricio 2001). After having the detailed and authenticated genotypic and phenotypic profiles of all the individuals of a segregating population and the parents

Table 1 Computer software used in quantitative trait loci (QTL) mapping

Plant species	Software	QTL	Reference
Cotton	Join map	65	Tang et al. (2015)
Peanut	QTL Cartographer	mQTL	Pandey et al. (2014)
Brassica oleraceae	MapQTL v.4	13	Lv et al. (2014)
Potato	Illumina software	mQTL	Prashar et al. (2014)
Grape	MapQTL	3	Ban et al. (2014)
Vicia faba	Map manager v. 20	4	Kaur et al. (2014)
Yellow croaker	Join map	7	Ye et al. (2014)
Rice	WinQTLCart. v. 2.5	3	Yun et al. (2014)
Maize	QTL Network v. 2	55	Liu et al. (2014)
Eggplant	QGene	71	Frary et al. (2014)
Wheat	SAS v. 8.1	4	Daoura et al. (2014)

involved, the need to analyze the results is fulfilled by the computer software. Hence, comprehensive interpretation is necessitated. The following details are mandatory to understand and interpret the results.

2.5.1 Isolation of Linked Markers

Linkage analysis for a high number of markers cannot be done manually. With the help of computer software, as already mentioned, linkage can be determined using odds ratios. Understandable expression of this ratio as the logarithm is based on the hypothesis that among the total number of markers how many are linked and the rest are unlinked. So, the logarithm of odd ratios is "the ratio of linkage versus no linkage" (Collard et al. 2005). A logarithm of odds (LOD) value greater than 3 is usually applicable for mapping. If any two markers have a LOD value of 3, the chances of their linkage is more than 1000:1 (linkage:no linkage). LOD is basically a *Z*-distribution (Morton 1955).

$$\begin{aligned} \text{LOD} = Z &= \log_{10}\left[\begin{array}{l}\text{Probability : that two markers are linked /}\\ \text{Probability that two markers are unlinked}\end{array}\right] \\ &= \log 10\left(1-\theta\right)^{\text{NR}} \times \theta^{R} / 0.5^{(\text{NR}+R)} \end{aligned}$$

where LOD is the logarithm of odds, θ is the recombinant fraction $= R/(\text{NR}+R)$, NR is the number of nonrecombinants, and R is the number of recombinants.

For large data sets, LOD can easily be calculated by the software mentioned. As many as the number of individuals in the population, the authenticity for determination of genetic distance between the markers and their sequence will be increased (Collard et al. 2005).

2.5.2 Mapping Function

This function It is required to convert a recombination fraction to the centimorgan (cM). It has been observed that recombination frequency and crossing over are not related in a linear order (Hartl and Jones 2001). Mapping function is also calculated by recombination values. Mapping functions are mathematical adjustments used in the measurement of genetic distances between two loci (Vinod 2011). Vinod (2011) has also emphasized that there are three options, to choose any of the three of the mapping functions:

- Complete interference does not permits double crossover; thus, Morgan's mapping function is there to be applied to cover additives.
- Incomplete interference enables double crossover to a certain extent, so we have Kosambi's mapping function to be used.
- No interference compels us to use Hadan's mapping function.

The genetic distance between the markers or genes is not directly related to the physical distance on DNA between genetic markers but also corresponds to the genome size of the plant species (Han and Ming 2014). As we know, markers split the mapping population into different clusters. Then, we have two types of grouping, one made by the markers and the other made by the visual observation of a continuously varying trait. We also have information about the linked markers. The statistical significance between the groups made by the markers and phenotypic trait means is then of prime importance (Young 1996).

2.5.3 Single-Marker Analysis

The single-marker effect can be analyzed statistically by *t* test, analysis of variance, and linear regression. Coefficient of determination (R^2) from the marker that explains the variation shown by a quantitative trait describes to what extent the marker and the QTL are linked to each other (Collard et al. 2005).

2.5.4 Interval Mapping

Sometimes an issue faced after single-marker regression is the effect of QTL magnitude and position (Erickson et al. 2004) and is resolved by the interval mapping techniques. It confines the QTL between the interval of a pair of two genetic markers with the help of a LOD (maximum likelihood) score (Collard et al. 2005). With the help of the interval, the mapping effect between the QTL and marker distance is expressed as a magnitude. Interval mapping is mainly of four types: simple interval mapping (SIM) as performed by Nelson (1997), composite interval mapping (CIM) (Basten et al. 2004), and multiple interval mapping (MIM) by Zeng et al. (1999).

The results obtained can be presented in a tabular form including highly linked markers or by graphs made by software (Burton et al. 2014, 2015; Oakley et al. 2014).

3 Modern Perspectives in QTL Analysis

With every passing day, progress is made in each field of the sciences as in electronics equipment, software packages, and chemical sciences. Traditional QTL mapping with respect to phenotyping and genotyping has been modernized as well. By the advances in genotypic and phenotypic platforms, it has now become possible to perform multi-trait analyses to unravel pleiotropy and the gene control mechanisms of complicated traits (Alonso-Blanco and Méndez-Vigo 2014).

The search for genetic polymorphism by using intensive and a variety of markers for a particular species always opens the door for detailed and modified phenotyping, which is helping breeders, plant pathologists, and physiologists as well. Having detailed genotypic information, it has become feasible to check the association of phenotypic diversity with new genetic regions. Understanding of genetic and molecular bases of polygenic traits has always been the major objective of the geneticists. As much as the loci are involved in controlling a particular trait, to detect their interaction and interference becomes vital, demanding a high level of specialization.

Meta-analysis of QTL mapping is a promising tool in which multiple quantitative trait loci are analyzed. Results obtained from different studies demand more statistical potential for QTL identification. Thus, meta-analysis can produce stronger inferences than other univariate studies. Details of the meta-analytical aspects of QTL have been a focus by Wu and Hu (2012). Possibly the genetic bases of quantitative traits are related to phenotypic level, which also fluctuates between simple oligogenic and complex polygenic inheritance (Joseph et al. 2013).

3.1 Genotyping to Genomics

For QTL analysis, much detail about genotypic variations is available so as to better analyze the QTL responsible for the trait of interest. In the present day, screening of a mapping population with only a few hundred markers has now been shifted up to 9K, 90K iSelect SNP (Avni et al. 2014). DArtT markers (Grzebelus et al. 2014) kits are also available, and genotyping by sequencing (GBS) is another platform to assess genetic diversity among the segregating population individuals (Verde et al. 2012; De Donato et al. 2013; Buckler 2014; Larson et al. 2014; Liu et al. 2014).

To detect single-nucleotide polymorphisms (SNPs), there is the widely adapted technique known as microarray technology that identifies SNPs through hybridization of DNA to oligonucleotides fixed on a chip. This microarray-based genotyping

detects allelic diversity by locating thousands of SNPs quickly (Huang and Han 2014). Five high-throughput genotyping methods have been reviewed by Huang and Han (2014): microarray-based genotyping, sequencing-based genotyping, genotyping-based sequencing, RNA-seq-based genotyping, and exon-sequencing-based genotyping.

As a crop reference genome is published, it become easier to characterize genome-wide variation for genetic mapping (Lai et al. 2010; Jiao et al. 2012).

3.2 Phenotyping to Phenomics

Large numbers of quantitative traits have been traditionally dissected at different levels of biological organization, not only because of details provided by advanced genotyping platforms but also simple to modern phenotypic techniques. The shift from phenotyping to phenomics is characterized by measurement of physical and biochemical traits of the organism as they respond to genetic diversity and environmental fluctuation. High-throughput 2D and 3D image analyses are being used to produce a phenotypic profile for QTL analysis (Topp et al. 2013; Joosen et al. 2012). Fieldwork for phenotyping is still very difficult, particularly when experimental crops have been planted on multiple environments in a vast area. Presently, some sensor-based platforms have been made for measuring biomass traits. Near-infrared spectroscopy on agricultural harvesters and spectral reflectance of plant canopies reflecting that future development in phenotyping will enables QTL analysis to be more detailed and widely applied in the gene discovery of food crops (Huang and Han 2014).

In phenotype cover interface between the genome and the environment, the phenotypic architecture is often equipped with an explained set of biodiversities (Burleigh et al. 2013).

3.3 Multiparent Advanced Generation Inter-Cross (MAGIC) Populations

MAGIC populations were first reported by Mott et al. (2000), and further development of such populations was performed by Kover et al. (2009) when it was hypothesized that QTL can be analyzed with improved accuracy along with cloning. A first panel of MAGIC lines with a set of 527 RILs of *Arabidopsis thaliana* was produced. Many known QTL with high precision and some important QTL for germination data and bolting time were detected. It has been recommended that the usage of MAGIC lines for other organisms can analyze QTL with more authenticity. QTL analysis using MAGIC lines using the probability of inheriting founder alleles across the whole genome at a time, and the whole-genome approach was estimated during a simulation study and proved itself a powerful method of analysis (Verbyla et al. 2014).

3.4 Next-Generation Sequencing (NGS)

NGS is a high-throughput sequencing-based genotyping technique. NGS technology is being widely adopted in which millions of DNA fractions at a time are being synthesized and sequenced. A genomic DNA sample is sliced into a library of small fragments that are uniformly and exactly sequenced in millions of parallel reactions. Newly detected lengths of bases, which are then called reads, again reunite using a known reference genome, and the full set of arranged reads represents the entire sequence of each chromosome (Grada and Weinbrecht 2013). Quail et al. (2012) compared three major sequencing platforms (Torrent's PGM, Pacific Bioscience RS, and the IlluminaMiSeq) with IlluminaHiSeq and concluded that all three fast turnaround sequencers were able to generate usable sequences but that crucial differences were found among the quality of the data.

Burleigh et al. (2013) purposed a next-generation phenomics project to facilitate biologists working with phenotypic data. Three prominent areas have been focused: (a) computer vision techniques to detect and record trait, (b) to increase the speed of the scoring and producing data sets supported with labeled anatomical images, and (c) to extract character data, natural languages will be processed.

NGS technologies offering latest moves toward fine-mapping as well as gene identification are greatly beneficial for food crop research. Trick et al. (2012) employed bulk segregant analysis (BSA) to fine-map the genes in tetraploid wheat lines and discovered SNPs with the help of next-generation sequences data.

4 Practical Potential of QTL Analysis

Mineral nutrition along with micro and trace elements (Lowry et al. 2012), primary and secondary metabolites (Joosen et al. 2013), and some flavonoids influencing quantitative traits have been studied in recent years (Routaboul et al. 2012). QTL analysis has explored certain levels of transcriptomic field encompassing transcript variations. Cubillos et al. (2012) while studying the RILs of *Arabidopsis* stated that genetic makeup that is responsible for transcriptional variation can assist knowing the phenotypic variation. During the study of epigenetic variations, a new class of methyl QTL has also been reported by Schmitz et al. (2013). Differentially methylated regions (DMR) have also been mapped and depict that a major part of such epigenetic quantitative variations is the consequence of genetic variation in *cis*-methyl QTL and *trans*-methyl QTL. Alonso-Blanco and Méndez-Vigo (2014) reviewed that DMRs are found associated with gene expression variation; hence, it can be assumed that methyl QTL are about to display another molecular level controlling expression and ultimately a higher level of quantitative traits. Differential QTLs for micronutrients in seed structure have also been reported by Blair et al. (2013). Moscou et al. (2011) detected a *cis*-eQTL gene as candidate for a major fungal resistance locus as well as *trans*-eQTL colocating with an enhancer of the resistance (reviewed by Alonso-Blanco and Méndez-Vigo 2014).

4.1 Crops with Improved Breeding Strategies

In crops, to dissect complex QTLs such as grain yield and stress tolerance, a huge sample size up to thousands of individuals is required. Now it has become possible to genotype such large samples using advanced genotypic methods. Crop breeding based on Marker-assisted selection is beneficial for simple Mendelian traits, but it is trouble creating for complex quantitative traits such as stress tolerance. Sometime through marker-assisted selection an unexpected QTL appear and fails in trait expression with little phenotypic variation being observed. Such trouble shoots can be overcome through genomic selection. Genomic selection is a simple and powerful approach in which breeding values are assigned using their phenotypes and marker genotypes (Deshmukh et al. 2014). By applying molecular breeding techniques, food crops such as maize, rice, potato, and wheat have been greatly advanced and developed with respect to their yield and stress tolerance.

4.2 Revealing the Genetic Bases of Abiotic Stress Tolerance

Abiotic stresses such as drought, heat, and salinity have massive influence on food crop yield. Mechanism of abiotic stress tolerance and exact phenotyping for such aim has been poorly formulated so far. Irrespective to the constant and single environment, a number of QTLs under the influence of environmental interaction has been identified so far. Studies such as germination, growth, and flowering time are being performed under variable field conditions of temperature and moisture in different environments (Fournier-Level et al. 2011; Ågren et al. 2013; Leinonen et al. 2013). Crops having ability to adapt extreme environmental conditions can be a significant source for crop improvement to fulfill the food needs of the ever-increasing populace (Huang and Han 2014). An abiotic stress tolerance mechanism can be traced out with phenomics and genomic tactics. Molecular bases of environmental tolerance are being probed through high-throughput phenotyping and genotyping platforms (Roy et al. 2011).

For drought tolerance, a QTL hotspot has been reported during the study of three populations in maize (Almeida et al. 2014). Constitutive and adoptive regions for drought tolerance were earlier reported by Almeida et al. (2013). Doubled haploid (DH) lines of canola were examined to associate root and leaf traits with drought tolerance with the help of QTL analysis (Mekonnen 2013). A high-throughput phenotyping platform has been used to identify drought tolerance QTL in wild barley introgression lines (Honsdorf et al. 2014). Using SNPs and haplotypes, QTL for height and biomass as secondary traits of drought tolerance were detected in maize by Lu et al. (2012).

Using SNPs, QTLs for heat tolerance in rice have been mapped by Ye et al. (2012). Paliwal et al. (2012) mapped QTLs on 7DS in hexaploid wheat using the composite interval mapping approach. Talukder et al. (2014) mapped QTLs for the

traits responsible for heat tolerance in wheat. Family-based QTL mapping of heat tolerance in *Triticum turgidum* has been performed by Ali et al. (2013).

As far as salt tolerance is concerned, QTL mapping has been performed in a variety of valuable crops to make the optimal use of saline or salt-affected lands. Chankaew et al. (2014) mapped QTLs for salt resistance in *Vigna marina*. In *Zoysia japonica* QTL analysis was performed by (Guo et al. 2014). Validation of the dominant salt tolerance gene in cultivated soybeans was mapped by Guan et al. (2014). In wild soybean, a major salt tolerance QTL was mapped by Ha et al. (2013).

4.3 Exposing Genetic Dissection of Biotic Stress Resistance

Stress resistance mechanisms are governed by many genes in most plant species. Plant–pathogen interactions underlie the effect of many genes responsible for plant defense. QTL analysis successfully helps in genetic dissection of the resistance mechanism. After detecting the QTL region involved in biotic stress resistance, marker-assisted selection enables the breeder to produce more resistant crops. QTLs for disease resistance found and utilized in breeding create durable resistance in genotypes, which proves an active method to achieve such broad-spectrum resistance, and thus these modified crops can be a good genetic resource (Kou and Wang 2010).

By using the marker-assisted selection (MAS) approach, gene pyramiding is performed to create broad-spectrum resistance in plant species (Tester and Langridge 2010). MAS for Lr 34, Yr 18, and powdery mildew 38 resistance in wheat and barley have been performed by Miedaner and Korzun (2012). Joshi and Nayak (2010) reviewed that durable stress resistance in crops can be achieved through gene pyramiding. Gene pyramiding for rice blast management through host-plant resistance has been reported by Sharma et al. (2012). To avail the gene pyramiding technique, Grimmer et al. (2014) analyzed four different wheat mapping populations being segregated for partial resistance to four contrasting foliar pathogens. It was stated using simple multiplicative survival (SMS) that with an increased number of loci, an enhanced level of disease resistance was achieved in wheat lines. In rice, quantitative resistance genes *pi21*, *Pi34,* and *Pi35* have been pyramided by Yasuda et al. (2015). Rice breeding lines with three pyramided resistance genes have been developed for broad-spectrum resistance against bacterial blight (Suh et al. 2013).

One major QTL for leaf spot and rust resistance in groundnut has been reported by Khedikar et al. (2010). In wheat, QTL mapping for multiple foliar disease and root lesion nematode resistance has been focused by Zwart et al. (2010). In potato, working on late blight resistance, a consensus map and QTL meta-analysis were performed by Danan et al. (2011). Genome-wide association mapping revealed disease resistance QTLs in barley (Gutiérrez et al. 2013). QTL mapping for fruit rot resistance in a *Capsicum annuum* population was done by Naegele et al. (2013).

A massive literature has become available in recent years to highlight the significance of QTL analysis. Here we present one view of a survey in table form (Table 2).

Table 2 Trends in QTL analysis in plants

Plant species	Type of population	Trait studied	QTL/method	Markers	Reference
Peanut	BC derived	Flower color, growth habit	Chi-square	115 SSRs	Fonceka et al. (2012a, b)
Cotton	RILs $F_{6:8}$	Fiber quality	50 QTLs CIM	5742 SSRs	Sun et al. (2012)
Maize	RILs	Kernel quality	26 QTLs	GWAS	Cook et al. (2012)
Peanut	BC derived	Yield components	95 QTLs SIM	SSRs	Fonceka et al. (2012a, b)
Wheat	RILs	Seedling traits	380 QTLs	SSRs	Guo et al. (2012)
Eggplant	156-F_2 plants	Anthocyanin contents	6 QTLs Interval Map.	SNPs, RFLPs, COSII	Barchi et al. (2012)
Wheat	RILs	APR Ug99	QTL on 7AS 2,3BS,5BL,	DArT	Singh et al. (2013)
Peach	126-F_1 Plant	VOCs, pest resistance	72 QTLs	SNPs, SSRs	Eduardo et al. (2013)
Rice	NILs BC_5F_4	Grain wt. Panical spike	QTLs	SSRs INDELs	Luo et al. (2013)
Grapevine	Pseudo F_1 Progeny	Skin color	eQTL	Transcript	Huang et al. (2013)
Sesamum indicum L.	P1, P2, F1, F2 BC1, BC2	Seed coat color	4QTLs CIM	653 SSRs AFLPs, RSAMP	Zhang et al. (2013)
Zoysia grass	120-F1	Salt tolerance	3QTLs Interval Map.	217-SRAP 25-RAPDs	Guo et al. (2014)
Capsicum annum	RILs F_6	Fruit rot resistance	QTLs	High-density Map.	Naegele et al. (2014)
Maize	RILs 3-populations	Root anatomical traits	6QTLs	Map used	Burton et al. (2014, 2015)
Eggplant	F2	Agronomic traits	7QTLs Interval Map.	Map used	Portis et al. (2014)
Wheat	RILs 3-populations	Yield traits	165 QTLs	DArT	Cui et al. (2014)
Rice	RILs	Seed vigor	8QTLs	Map used	Xie et al. (2014)
Soybean	304-Short season lines	Agronomic traits	GWAS	GBS	Sonah et al. (2015)

5 Conclusions and Future Perspective

From the Mendelian era to Morgan's linkage analyses, extension of knowledge from qualitative traits to quantitative, it has become clear that in days to come nucleotides and their expression will be comprehensively understood. Fine mapping, chromosome walking, and dissection of the quantitative traits with the help of highly dense maps has annexed genetics with all other biological sciences. To see genetic change and diversity has now become very clear with the aid of next-generation sequencing platforms. With the passage of time, the science of genetics is becoming laboratory based, but to the common man the threat of hunger and starvation still prevail with the increasing populace. What has been done in the field of molecular breeding has not yet advanced the contribution of the green revolution: even at that time such sophisticated tools were not available. Now with the advances in laboratory science there comes a dire responsibility to the agricultural researcher to feed the world populace of more than 9 billion in coming years. Can such fancy laboratory techniques fill the empty stomachs of nutritionally deprived peasants? There is a need to integrate laboratory science with fieldwork and to target it as per the demands of the market and common people. Such research should have an impact rather than being an impact factor.

References

Ågren J, Oakley CG, McKay JK, Lovell JT, Schemske DW (2013) Genetic mapping of adaptation reveals fitness tradeoffs in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 110:21077–21082

Akbarpour O, Dehghani H, Sorkhi B, GauchJr H (2014) Evaluation of genotype × environment interaction in barley (*Hordeum vulgare* L.) based on AMMI model using developed SAS program. J Agric Sci Technol 16:909–920

Ali MB, Ibrahim AM, Malla S, Rudd J, Hays DB (2013) Family-based QTL mapping of heat stress tolerance in primitive tetraploid wheat (*Triticum turgidum* L.). Euphytica 192:189–203

Almeida GD, Makumbi D, Magorokosho C, Nair S, Borém A, Ribaut J-M, Bänziger M, Prasanna BM, Crossa J, Babu R (2013) 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

Almeida GD, Nair S, Borém A, Cairns J, Trachsel S, Ribaut J-M, Bänziger M, Prasanna BM, Crossa J, Babu R (2014) Molecular mapping across three populations reveals a QTL hotspot region on chromosome 3 for secondary traits associated with drought tolerance in tropical maize. Mol Breed 34:701–715

Alonso-Blanco C, Méndez-Vigo B (2014) Genetic architecture of naturally occurring quantitative traits in plants: an updated synthesis. Curr Opin Plant Biol 18:37–43

Asins M (2002) Present and future of quantitative trait locus analysis in plant breeding. Plant Breed 121:281–291

Avni R, Nave M, Eilam T, Sela H, Alekperov C, Peleg Z, Dvorak J, Korol A, Distelfeld A (2014) Ultra-dense genetic map of durum wheat × wild emmer wheat developed using the 90K iSelect SNP genotyping assay. Mol Breed 34:1549–1562

Ban Y, Mitani N, Hayashi T, Sato A, Azuma A, Kono A, Kobayashi S (2014) Exploring quantitative trait loci for anthocyanin content in interspecific hybrid grape (*Vitis labruscana* × *Vitis vinifera*). Euphytica 198:101–114

Barchi L, Lanteri S, Portis E, Valè G, Volante A, Pulcini L, Ciriaci T, Acciarri N, Barbierato V, Toppino L (2012) A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. PLoS One 7:e43740

Barrett B, Griffiths A, Schreiber M, Ellison N, Mercer C, Bouton J, Ong B, Forster J, Sawbridge T, Spangenberg G (2004) A microsatellite map of white clover. Theor Appl Genet 109: 596–608

Basten CJ, Weir BS, Zeng Z-B (2004) QTL Cartographer, version 1.17. Department of Statistics, North Carolina State University, Raleigh, NC

Blair MW, Izquierdo P, Astudillo C, Grusak MA (2013) A legume biofortification quandary: variability and genetic control of seed coat micronutrient accumulation in common beans. Front Plant Sci 4:275

Bradbury P, Parker T, Hamblin MT, Jannink J-L (2011) Assessment of power and false discovery rate in genome-wide association studies using the BarleyCAP germplasm. Crop Sci 51:52–59

Buckler ES (2014) Applying genotyping-by-sequencing to characterize and map in diverse maize. In: Plant and Animal Genome XXII Conference: Plant and Animal Genome

Burleigh JG, Alphonse K, Alverson AJ, Bik HM, Blank C, Cirranello AL, Cui H, Daly M, Dietterich TG, Gasparich G (2013) Next-generation phenomics for the Tree of Life. PLoS Curr 5

Burton AL, Johnson JM, Foerster JM, Hirsch CN, Buell C, Hanlon MT, Kaeppler SM, Brown KM, Lynch JP (2014) QTL mapping and phenotypic variation for root architectural traits in maize (*Zea mays* L.). Theor Appl Genet 127:2293–2311

Burton AL, Johnson J, Foerster J, Hanlon MT, Kaeppler SM, Lynch JP, Brown KM (2015) QTL mapping and phenotypic variation of root anatomical traits in maize (*Zea mays* L.). Theor Appl Genet 128(1):93–106

Chankaew S, Isemura T, Naito K, Ogiso-Tanaka E, Tomooka N, Somta P, Kaga A, Vaughan DA, Srinives P (2014) QTL mapping for salt tolerance and domestication-related traits in *Vigna marina* subsp. *oblonga*, a halophytic species. Theor Appl Genet 127:691–702

Cheng P, Chen X (2010) Molecular mapping of a gene for stripe rust resistance in spring wheat cultivar IDO377s. Theor Appl Genet 121:195–204

Collard B, Jahufer M, Brouwer J, Pang E (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica 142:169–196

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834

Cubillos FA, Yansouni J, Khalili H, Balzergue S, Elftieh S, Martin-Magniette M-L, Serrand Y, Lepiniec L, Baud S, Dubreucq B (2012) Expression variation in connected recombinant populations of *Arabidopsis thaliana* highlights distinct transcriptome architectures. BMC Genomics 13:117

Cui F, Zhao C, Ding A, Li J, Wang L, Li X, Bao Y, Li J, Wang H (2014) Construction of an integrative linkage map and QTL mapping of grain yield-related traits using three related wheat RIL populations. Theor Appl Genet 127:659–675

Danan S, Veyrieras J-B, Lefebvre V (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

Daoura BG, Chen L, Du Y, Hu Y-G (2014) Genetic effects of dwarfing gene *Rht-5* on agronomic traits in common wheat (*Triticum aestivum* L.) and QTL analysis on its linked traits. Field Crop Res 156:22–29

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013) Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One 8:e62137

Deshmukh RK, Sonah H, Patil G, Chen W, Prince S, Mutava R, Vuong T, Valliyodan B, Nguyen HT (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Plant Genet Genomics 5:244

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43–52

Eduardo I, Chietera G, Pirona R, Pacheco I, Troggio M, Banchi E, Bassi D, Rossini L, Vecchietti A, Pozzi C (2013) Genetic dissection of aroma volatile compounds from the essential oil of peach fruit: QTL analysis and identification of candidate genes using dense SNP maps. Tree Genet Genomes 9:189–204

Erickson DL, Fenster CB, Stenøien HK, Price D (2004) Quantitative trait locus analyses and the study of evolutionary process. Mol Ecol 13:2505–2522

Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Fonceka D, Tossim H-A, Rivallan R, Vignes H, Faye I, Ndoye O, Moretzsohn MC, Bertioli DJ, Glaszmann J-C, Courtois B (2012a) Fostered and left behind alleles in peanut: interspecific QTL mapping reveals footprints of domestication and useful natural variation for breeding. BMC Plant Biol 12:26

Fonceka D, Tossim H-A, Rivallan R, Vignes H, Lacut E, de Bellis F, Faye I, Ndoye O, Leal-Bertioli SC, Valls JF (2012b) Construction of chromosome segment substitution lines in peanut (*Arachis hypogaea* L.) using a wild synthetic and QTL mapping for plant morphology. PLoS One 7:e48642

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011) A map of local adaptation in *Arabidopsis thaliana*. Science 334:86–89

Frary A, Frary A, Daunay M-C, Huvenaars K, Mank R, Doğanlar S (2014) QTL hotspots in eggplant (*Solanum melongena*) detected with a high resolution map and CIM analysis. Euphytica 197:211–228

Gardner KM, Latta RG (2007) Shared quantitative trait loci underlying the genetic correlation between continuous traits. Mol Ecol 16:4195–4209

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. Theor Appl Genet 46:319–330

Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133:e11

Grimmer M, Boyd L, Clarke S, Paveley N (2014) Pyramiding of partial disease resistance genes has a predictable, but diminishing, benefit to efficacy. Plant Pathol 64(3):748–753

Grzebelus D, Iorizzo M, Senalik D, Ellison S, Cavagnaro P, Macko-Podgorni A, Heller-Uszynska K, Kilian A, Nothnagel T, Allender C (2014) Diversity, genetic mapping, and signatures of domestication in the carrot (*Daucus carota* L.) genome, as revealed by Diversity Arrays Technology (DArT) markers. Mol Breed 33:625–637

Guan R, Chen J, Jiang J, Liu G, Liu Y, Tian L, Yu L, Chang R, Qiu L-J (2014) Mapping and validation of a dominant salt tolerance gene in the cultivated soybean (*Glycine max*) variety Tiefeng 8. Crop J 2(6):358–365

Guo Y, Kong F-M, Xu Y-F, Zhao Y, Liang X, Wang Y-Y, An D-G, Li S-S (2012) QTL mapping for seedling traits in wheat grown under varying concentrations of N, P and K nutrients. Theor Appl Genet 124:851–865

Guo H, Ding W, Chen J, Chen X, Zheng Y, Wang Z, Liu J (2014) Genetic linkage map construction and QTL mapping of salt tolerance traits in *Zoysia* grass (*Zoysia japonica*). PLoS One 9:e107249

Gutiérrez L, Berberian N, Capettini F, Falcioni E, Fros D, Germán S, Hayes PM, Huerta-Espino J, Herrera S, Pereyra S (2013) Genome-wide association mapping identifies disease-resistance QTLs in barley germplasm from Latin America, Advances in barley sciences. Springer, Dordrecht, pp 209–215

Ha B-K, Vuong TD, Velusamy V, Nguyen HT, Shannon JG, Lee J-D (2013) Genetic mapping of quantitative trait loci conditioning salt tolerance in wild soybean (*Glycine soja*) PI 483463. Euphytica 193:79–88

Han J, Ming R (2014) Molecular genetic mapping of papaya, Genetics and genomics of papaya. Springer, New York, pp 143–155

Hartl D, Jones E (2001) Introduction to molecular genetics and genomics. In: Hartl DL, Jones EW (eds) Genetics: analysis of genes and genomes, 5th edn. Jones & Bartlett, Mississauga, ON, Canada, pp 1–35

Honsdorf N, March TJ, Berger B, Tester M, Pillen K (2014) High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. PLoS One 9:e97047

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531–551

Huang Y-F, Bertrand Y, Guiraud J-L, Vialet S, Launay A, Cheynier V, Terrier N, This P (2013) Expression QTL mapping in grapevine: revisiting the genetic determinism of grape skin colour. Plant Sci 207:18–24

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44:812–815

Joehanes R, Nelson JC (2008) QGene 4.0, an extensible Java QTL-analysis platform. Bioinformatics 24:2788–2789

Joosen RVL, Arends D, Willems LAJ, Ligterink W, Jansen RC, Hilhorst HW (2012) Visualizing the genetic landscape of *Arabidopsis* seed performance. Plant Physiol 158:570–589

Joosen RVL, 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

Joseph B, Corwin JA, Züst T, Li B, Iravani M, Schaepman-Strub G, Turnbull LA, Kliebenstein DJ (2013) Hierarchical nuclear and cytoplasmic genetic architectures for plant growth and defense within *Arabidopsis*. Plant Cell 25:1929–1945

Joshi RK, Nayak S (2010) Gene pyramiding: a broad spectrum technique for developing durable stress resistance in crops. Biotechnol Mol Biol Rev 5:51–60

Kaur S, Kimber RB, Cogan NO, Materne M, Forster JW, Paull JG (2014) SNP discovery and high-density genetic mapping in faba bean (*Vicia faba* L.) permits identification of QTLs for *Ascochyta* blight resistance. Plant Sci 217:47–55

Khedikar Y, Gowda M, Sarvamangala C, Patgar K, Upadhyaya H, Varshney R (2010) A QTL study on late leaf spot and rust revealed one major QTL for molecular breeding for rust resistance in groundnut (*Arachis hypogaea* L.). Theor Appl Genet 121:971–984

Kou Y, Wang S (2010) Broad-spectrum and durability: understanding of quantitative disease resistance. Curr Opin Plant Biol 13:181–185

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:e1000551

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030

Larson WA, Seeb LW, Everett MV, Waples RK, Templin WD, Seeb JE (2014) Data from geotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). Evol Appl 7(3):355–369

Leinonen PH, Remington DL, Leppälä J, Savolainen O (2013) Genetic basis of local adaptation and flowering time variation in *Arabidopsis lyrata*. Mol Ecol 22:709–723

Liu W, Maurer HP, Reif JC, Melchinger A, Utz H, Tucker MR, Ranc N, Della Porta G, Würschum T (2012) Optimum design of family structure and allocation of resources in association mapping with lines from multiple crosses. Heredity 110:71–79

Liu Y, Wang L, Sun C, Zhang Z, Zheng Y, Qiu F (2014) Genetic analysis and major QTL detection for maize kernel size and weight in multi-environments. Theor Appl Genet 127:1019–1037

Lowry DB, Sheng CC, Zhu Z, Juenger TE, Lahner B, Salt DE, Willis JH (2012) Mapping of ionomic traits in *Mimulus guttatus* reveals Mo and Cd QTLs that colocalize with MOT1 homologues. PLoS One 7:e30730

Lu Y, Xu J, Yuan Z, Hao Z, Xie C, Li X, Shah T, Lan H, Zhang S, Rong T (2012) Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. Mol Breed 30:407–418

Luciano Da Costa ES, Wang S, Zeng Z-B (2012) Composite interval mapping and multiple interval mapping: procedures and guidelines for using Windows QTL cartographer. Methods Mol Biol 871:75–119

Luo X, Ji S-D, Yuan P-R, Lee H-S, Kim D-M, Balkunde S, Kang J-W, Ahn S-N (2013) QTL mapping reveals a tight linkage between QTLs for grain weight and panicle spikelet number in rice. Rice 6:33

Lv H, Wang Q, Zhang Y, Yang L, Fang Z, Wang X, Liu Y, Zhuang M, Lin Y, Yu H (2014) Linkage map construction using InDel and SSR markers and QTL analysis of heading traits in *Brassica oleracea* var. *capitata* L. Mol Breed 34:87–98

Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565–577

Manly KF, Cudmore RH Jr, Meer JM (2001) Map Manager QTX, cross-platform software for genetic mapping. Mamm Genome 12:930–932

Mauricio R (2001) Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat Rev Genet 2:370–381

Mekonnen MD (2013) QTL mapping of root and leaf traits associated with drought tolerance in a canola (*Brassica napus* L.) doubled haploid population. Colorado State University, Fort Collins

Miedaner T, Korzun V (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. Phytopathology 102:560–566

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277

Moscou MJ, Lauter N, Steffenson B, Wise RP (2011) Quantitative and qualitative stem rust resistance factors in barley are associated with transcriptional suppression of defense regulons. PLoS Genet 7:e1002208

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 USA 97:12649–12654

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 Online 21:2194–2202

Naegele R, Ashrafi H, Hill T, Chin-Wo SR, Van Deynze A, Hausbeck M (2014) QTL mapping of fruit rot resistance to the plant pathogen *Phytophthora capsici* L. in a recombinant inbred line *Capsicum annuum* L. population. Phytopathology 104(5):479–483

Nelson JC (1997) QGENE: software for marker-based genomic analysis and breeding. Mol Breed 3:239–245

Oakley CG, Ågren J, Atchison RA, Schemske DW (2014) QTL mapping of freezing tolerance: links to fitness and adaptive trade-offs. Mol Ecol 23:4304–4315

Paliwal R, Röder MS, Kumar U, Srivastava J, Joshi AK (2012) QTL mapping of terminal heat tolerance in hexaploid wheat (*T. aestivum* L.). Theor Appl Genet 125:561–575

Pandey MK, Wang ML, Qiao L, Feng S, Khera P, Wang H, Tonnis B, Barkley NA, Wang J, Holbrook CC (2014) Identification of QTLs associated with oil content and mapping FAD2 genes and their relative contribution to oil quality in peanut (*Arachis hypogaea* L.). BMC Genet 15:133

Portis E, Barchi L, Toppino L, Lanteri S, Acciarri N, Felicioni N, Fusari F, Barbierato V, Cericola F, Valè G (2014) QTL mapping in eggplant reveals clusters of yield-related loci and orthology with the tomato genome. PLoS One 9:e89499

Prashar A, Hornyik C, Young V, McLean K, Sharma SK, Dale MFB, Bryan GJ (2014) Construction of a dense SNP map of a highly heterozygous diploid potato population and QTL analysis of tuber shape and eye depth. Theor Appl Genet 127:2159–2171

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and IlluminaMiSeq sequencers. BMC Genomics 13:341

Rahman H, Kebede B, Zimmerli C, Yang R-C (2014) Genetic study and QTL mapping of seed glucosinolate content in *Brassica rapa* L. Crop Sci 54:537–543

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

Roy SJ, Tucker EJ, Tester M (2011) Genetic analysis of abiotic stress tolerance in crops. Curr Opin Plant Biol 14:232–239

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552

Schmitz RJ, He Y, Valdés-López O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G (2013) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res 23:1663–1674

Sharma T, Rai A, Gupta S, Vijayan J, Devanna B, Ray S (2012) Rice blast management through host-plant resistance: retrospect and prospects. Agric Res 1:37–52

Singh S, Singh RP, Bhavani S, Huerta-Espino J, Eugenio L-VE (2013) QTL mapping of slow-rusting, adult plant resistance to race Ug99 of stem rust fungus in PBW343/Muu RIL population. Theor Appl Genet 126:1367–1375

Slate J (2005) Invited review: Quantitative trait locus mapping in natural populations: progress, caveats and future directions. Mol Ecol 14:363–379

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2015) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J 13(2):211–221

Suh J-P, Jeung J-U, Noh T-H, Cho Y-C, Park S-H, Park H-S, Shin M-S, Kim C-K, Jena KK (2013) Development of breeding lines with three pyramided resistance genes that confer broad-spectrum bacterial blight resistance and their molecular analysis in rice. Rice 6:1–11

Sun F-D, Zhang J-H, Wang S-F, Gong W-K, Shi Y-Z, Liu A-Y, Li J-W, Gong J-W, Shang H-H, Yuan Y-L (2012) QTL mapping for fiber quality traits across multiple generations and environments in upland cotton. Mol Breed 30:569–582

Talukder SK, Babar MA, Vijayalakshmi K, Poland J, Prasad PV, Bowden R, Fritz A (2014) Mapping QTL for the traits associated with heat tolerance in wheat (*Triticum aestivum* L.). BMC Genet 15:97

Tang S, Teng Z, Zhai T, Fang X, Liu F, Liu D, Zhang J, Liu D, Wang S, Zhang K (2015) Construction of genetic map and QTL analysis of fiber quality traits for Upland cotton (*Gossypium hirsutum* L.). Euphytica 201:195–213

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

Tinker N, Mather D (1995) MQTL: software for simplified composite interval mapping of QTL in multiple environments. J Agric Genomics 1(2)

Topp CN, Iyer-Pascuzzi AS, Anderson JT, Lee C-R, Zurek PR, Symonova O, Zheng Y, Bucksch A, Mileyko Y, Galkovskyi T (2013) 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. Proc Natl Acad Sci USA 110:E1695–E1704

Touré A, Haussmann B, Jones N, Thomas H, Ougham H (2000) Construction of a genetic map, mapping of major genes, and QTL analysis In: Haussmann BIG, Geiger HH, Hess DE, Hash CT, Bramel-Cox P (eds) Application of molecular markers in plant breeding. Training manual for a seminar held at IITA, Ibadan, Nigeria, from 16–17 August 1999. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India, p 17–34

Trick M, Adamski NM, Mugford SG, Jiang C-C, Febrer M, Uauy C (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. BMC Plant Biol 12:14

Van Ooijen J (2004) MapQTL® 5, software for the mapping of quantitative trait loci in experimental populations. Kyazma, Wageningen, Netherlands, p 63

Van Ooijen J (2006) JoinMap 4. Software for the calculation of genetic linkage maps in experimental populations. Kyazma, Wageningen, Netherlands

Van Ooijen J, Kyazma B (2009) MapQTL 6. Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma, Wageningen, Netherlands

Van Ooijen J, Boer M, Jansen R, Maliepaard C (2000) MapQTL 4.0: software for the calculation of QTL positions on genetic maps (user manual)

Verbyla AP, George AW, Cavanagh CR, Verbyla KL (2014) Whole-genome QTL analysis for MAGIC. Theor Appl Genet 127:1753–1770

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E (2012) Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. PLoS One 7:e35668

Veyrieras J-B, Goffinet B, Charcosset A (2007) MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. BMC Bioinformatics 8:49

Vinod K (2011) Kosambi and the genetic mapping function. Resonance 16:540–550

Wang H, Smith KP, Combs E, Blake T, Horsley RD, Muehlbauer GJ (2012) Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor Appl Genet 124:111–124

Wang S, Basten C, Zeng Z (2013) Windows QTL Cartographer 2.5. User manual. Department of Statistics, North Carolina State University, Raleigh

Wu X-L, Hu Z-L (2012) Meta-analysis of QTL mapping experiments. In: Rifkin SA (ed) Quantitative trait loci (QTL). Springer, New York, pp 145–171

Würschum T (2012) Mapping QTL for agronomic traits in breeding populations. Theor Appl Genet 125:201–210

Xie L, Tan Z, Zhou Y, Xu R, Feng L, Xing Y, Qi X (2014) Identification and fine mapping of quantitative trait loci for seed vigor in germination and seedling establishment in rice. J Integr Plant Biol 56(8):749–759

Yasuda N, Mitsunaga T, Hayashi K, Koizumi S, Fujita Y (2015) Effects of pyramiding quantitative resistance genes pi21, Pi34, and Pi35 on rice leaf blast disease. Plant Dis 99(7):904–909

Ye C, Argayoso MA, Redoña ED, Sierra SN, Laza MA, Dilla CJ, Mo Y, Thomson MJ, Chin J, Delaviña CB (2012) Mapping QTL for heat tolerance at flowering stage in rice using SNP markers. Plant Breed 131:33–41

Ye H, Liu Y, Liu X, Wang X, Wang Z (2014) Genetic mapping and QTL analysis of growth traits in the large yellow croaker *Larimichthys crocea*. Mar Biotechnol 16:729–738

Young N (1996) QTL mapping and quantitative disease resistance in plants. Annu Rev Phytopathol 34:479–501

Yun B-W, Kim M-G, Handoyo T, Kim K-M (2014) Analysis of rice grain quality-associated quantitative trait loci by using genetic mapping. Am J Plant Sci. doi:10.4236/ajps.2014.59125

Zambrano JL, Jones MW, Francis DM, Tomas A, Redinbaugh MG (2014) Quantitative trait loci for resistance to maize rayado fino virus. Mol Breed 34:989–996

Zeng Z-B, Kao C-H, Basten CJ (1999) Estimating the genetic architecture of quantitative traits. Genet Res 74:279–289

Zhang H, Miao H, Wei L, Li C, Zhao R, Wang C (2013) Genetic analysis and QTL mapping of seed coat color in sesame (*Sesamum indicum* L.). PLoS One 8:e63898

Zwart R, Thompson J, Milgate A, Bansal U, Williamson P, Raman H, Bariana H (2010) QTL mapping of multiple foliar disease and root-lesion nematode resistances in wheat. Mol Breed 26:107–124

Transposon Activity in Plant Genomes

Nermin Gozukirmizi, Aslihan Temel, Sevgi Marakli, and Sibel Yilmaz

Contents

N. Gozukirmizi (✉) • A. Temel • S. Marakli • S. Yilmaz
Department of Molecular Biology and Genetics, Faculty of Science, Istanbul University, 34134, Vezneciler, Istanbul, Turkey
e-mail: nermin@istanbul.edu.tr

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_4

Abstract Transposable elements (TEs) were first discovered in maize plants. However, they exist in all plant species investigated so far. Although plants with small genomes have smaller transposon percentages, plants with large genomes have high transposon percentages. For example, *Arabidopsis thaliana* has a genome size of 125 Mb, which comprises 14 % transposons, and the *Hordeum vulgare* genome (5300 Mb) has 80 %. TEs are classified into two major groups based on their transposition mechanism. Class I elements are characterized by DNA sequences with homology to reverse transcriptase, and they are often referred to as retroelements, retrotransposons, or retrovirus-like elements. Retrotransposons function by a copy-and-paste transposition mechanism. Class II TEs (DNA transposons) move by a cut-and-paste mechanism. TEs affect the genome dynamics of plants by regulation of gene expression and chromosomal mutations (such as duplications, insertions/deletions, and structural variations). Transposition rates among generations are about 10^{-3} to 10^{-4}, which is a higher rate than spontaneous mutations. All TEs in a cell are named as transposomes, and transposomics is a new area to work with transposomes. Although some bioinformatics software has recently been developed for the annotation of TEs in sequenced genomes, there are very few computational tools strictly dedicated to the identification of active TEs using genome-wide approaches. In this review article, after a brief introduction and review of the transposable elements, we discuss the effects of TEs in plant gene expression and evolution, and also present our recent research data on barley retrotransposons.

Keywords Mobile elements • Plant genome dynamics • Transposomic • Plant evolution

1 Introduction

Transposon, a segment of DNA that moves to a new location in a chromosome, or to another chromosome or cell, and alters the existing genetic structure, sometimes causes significant changes. Transposons were first described by Barbara McClintock, a maize cytogeneticist who was rewarded with the Nobel prize (1984) 30 years later than her exploration of the relationship between chromosome breaks and maize grain color alterations. Today, we are aware that gene and genome dynamics caused by transposons exist in somatic tissues not only of plants but also of almost all living organisms. Different terms, such as jumping genes, mobile genetics elements, controlling elements, and transposable elements, are used in synonymous ways. Today, we have gained incredible knowledge about the structure, types, and life cycles of these genomic sequences; however, their origin, functions, roles in gene expression, and evolutionary processes still need to be investigated. Developments in DNA sequencing, combined with advances in functional genomics and bioinformatics, have affected studies with transposons because they concern the genomes of living organisms, and the majority of these transposable elements (TEs) are either defective,

fossilized copies, or potentially active copies that are restrained by host silencing systems. However, active transposition evidenced by instances of mutagenic (yet potentially evolutionarily significant) insertions has been demonstrated. For example, TEs have been shown to silence or alter expression of genes adjacent to insertion sites, contribute to chromosomal rearrangements via recombination, epigenetically alter regional methylation patterns, and provide template sequences for RNA interference (Feschotte et al. 2002; Bennetzen 2005; Morgante et al. 2007; Weil and Martienssen 2008; Slotkin et al. 2012; Zhao et al. 2016). These diverse functional impacts of TEs, and their intrinsic contribution to genomic plasticity, suggest that these elements have a major function in molecular diversification and, ultimately, species divergence. In this chapter, we provide the reader with the fundamentals of TE biology, with an emphasis on plant elements. We begin with an overview of TE classification and transposition mechanisms, followed by an examination of the extensive variability in both inter- and intraspecific TE content across diverse plant taxa, and transposon markers, transposomics, and their effects on gene expression and evolutionary processes.

2 Structure of Transposons

Transposons are classified into two classes: class I and class II (retrotransposons and DNA transposons, respectively). Before we discuss the types and classification of transposons, a general explanation of their structures will be helpful to understand the following parts of this review. Transposons use many different enzymes for their transposition. Although some transposons can encode these enzymes (autonomous), others cannot (nonautonomous) and use enzymes of autonomous transposons.

Retrotransposons have a more complex structure than DNA transposons. Group-specific antigen (GAG) is the first protein-coding region of a retrotransposon (Fig. 1). It encodes proteins that pack retrotransposon mRNA in the cytoplasm, and this structure is named virus-like particle (VLP). Reverse transcription of retrotransposon mRNA occurs in VLPs (Jaaskelainen et al. 1999). Although it usually has the same open reading frame (ORF) as other domains, it is transcribed from a different ORF in some retrotransposons (Ohtsubo et al. 1999).

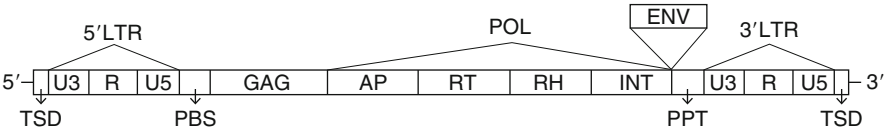


Fig. 1 Schematic demonstration of a retrotransposon having LTR regions. *LTR* long terminal repeat, *R* repeated region, *U3* unique for 3′-end of RNA, *U5* unique for 5′-end of RNA, *PBS* primer binding site, *GAG* group-specific antigen, *POL* polyprotein, *AP* aspartic peptidase, *RT* reverse transcriptase, *RH* ribonuclease H, *INT* integrase, *ENV* envelope, *PPT* polypurine tract, *TSD*, target site duplication

The other coding region of retrotransposons is the polyprotein (POL), which region encodes four enzymes: aspartic peptidase (AP), reverse transcriptase (RT), RNase H (RH), and integrase (INT). These enzymes can act either together as part of a protein or individually. AP is a proteolytic enzyme that is responsible for maturation of proteins required for the retrotransposon life cycle. RT is an RNA-dependent DNA polymerase that is necessary for reverse transcription of retrotransposon mRNA. RH is a hydrolytic enzyme that is responsible for the hydrolysis of the original RNA template that is part of the RNA/DNA hybrid generated after reverse transcription. INT catalyzes the insertion of new retrotransposon copies into the genome (Flavell 1995; Malik and Eickbush 2001). Some retrotransposons have a tyrosine recombinase (YR) domain instead of INT in the POL region. YR is typically involved in site-specific recombinations between similar or identical DNA sequences (Coates et al. 2005). Similar to YR, apurinic endonuclease (*APE*) is also encoded instead of INT in certain types of retrotransposons and provides site-specific insertion (Yadav et al. 2009).

In addition to these domains, some retrotransposons can have extra domains that give retrotransposons several properties. The envelope (ENV) domain is one of these domains and was first defined in viruses. This domain encodes a protein that is responsible for cell-to-cell transfer of virus. Most of the retrotransposons lost their ENV domains during the evolutionary process. However, the domain is still present in some retrotransposons (Wright and Voytas 1998).

Retrotransposons also have some sequences that do not encode a protein product but are essential for the life cycle of the retrotransposon. Long terminal repeats (LTRs) sequences are direct repeats that are present at two borders (the 5′- and 3′-ends) of some retrotransposons. They vary within the retrotransposon families. However, the LTR sequence of a retrotransposon is well conserved between plant species during the evolutionary process. LTR sequences do not encode a protein product but have important functions for transposition. An LTR sequence has three regions: Unique for the 3′-end of RNA, Repeated region, and Unique for the 5′-end of RNA (U3, R, and U5, respectively). In the U3 region, the promoter and some *cis*-acting elements are present (Pouteau et al. 1994; Takeda et al. 1998). The R region functions during reverse transcription of retrotransposon mRNA, and the U5 region constitutes the first portion of the retrotranscribed genome (SanMiguel et al. 1998; Jiang et al. 2002).

In addition to LTR sequences, retrotransposons have three noncoding regions: primer-binding site (PBS), polypurine tract (PPT), and target site duplication (TSD). PBS is a region about 18 nucleotides in length located between the 5′-LTR and GAG domains. During reverse transcription of retrotransposon mRNA, the 3′-end of a cellular tRNA binds to this region and acts as a primer (Wilhelm and Wilhelm 2001). PBS sequences are used for construction of a sequencing library to identify LTR sequences and their insertion sites (Monden et al. 2014). PPT is located between 3′-LTR and the internal domains: it contains about ten purine (A-G) bases and is involved in the synthesis of double-stranded cDNA from single-stranded (Mak and Kleiman 1997; Gabus et al. 1998). TSD is a short direct repeat typically four to eight nucleotides in length and that is present at the two borders. TSD is characteristic of both retrotransposons and DNA transposons.

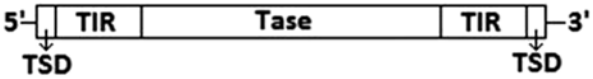


Fig. 2 Schematic demonstration of a DNA transposon. *TIR* terminal inverted repeat, *Tase* transposase, *TSD* target site duplication

Despite the complex structure of retrotransposons, DNA transposons have a more simple structure (Fig. 2). DNA transposons encode a transposase (Tase) enzyme. This enzyme cuts the DNA transposon and integrates it to a new location.

DNA transposons also contain noncoding repeat regions, such as LTR, at both borders, but these repeats are present in inverted orientation (Poulter and Goodwin 2005). Thus, these repeats are called terminal inverted repeats (TIRs). TIRs are essential for the transposition of most transposons. Generally, this is a characteristic feature of DNA transposons. However, some retrotransposon groups also have these repeat sequences. Generally, two functional regions are present in TIR sequences. Although the first region is involved in DNA cleavage and strand transfer reactions, the second one is required for specific recognition and binding (Szabo et al. 2010).

3 Types and Classification of Transposons

After the great discovery of transposons by Barbara McClintock, numerous studies in many areas have been performed regarding transposons. It was found that transposons are present as billions of copies in almost every organism investigated so far. In these studies, transposons with different features were also defined. Therefore, a classification system was needed, and the first classification was proposed by Finnegan (1989). Finnegan divided transposons into two classes, class I and class II, based on their intermediate of transposition. In the following years, this simple classification system has become inadequate. Some hierarchical classifications in detail were proposed by the International Committee on the Taxonomy of Viruses (ICTV) and Capy et al. (1998) for certain types of transposons. In this review, we describe a different type of transposons based on a unified classification system proposed by Wicker et al. (2007). In this hierarchical system, transposons are classified in five levels as class, subclass, order, superfamily, and family. In the first level, class, transposons are divided based on the presence or absence of an RNA transposition intermediate, as previously proposed by Finnegan (1989). The second level is the subclass, and transposons are separated according to their transposition mechanisms, either copy-paste or cut-paste. Because all class I transposons (retrotransposons) use copy-paste transposition, only class II transposons (DNA transposons) have different subclasses. Order, the third level, bases major differences in the insertion mechanisms and enzymology between transposons that belong to the same subclass. In the superfamily, the fourth level, transposons are grouped based on structural differences of their proteins or noncoding regions. The fifth level, family, is based on DNA sequence similarities.

3.1 Class I Transposons (Retrotransposons)

Retrotransposons are mobile genetic elements that transpose an RNA intermediate and have an important function in genome evolution. Because of their copy-paste transposition mechanism, they cause genome enlargement. They are present at high copy numbers in most plants, constituting more than 50 % of the nuclear genome in many cases. Retrotransposons resemble retroviruses because of both their structure and replication mechanism. Because all retrotransposons have copy-paste mechanisms, they do not have different subclasses. All belong to the same subclass, which is divided into five orders.

3.1.1 Order 1: LTR Retrotransposons

LTR retrotransposons are the largest group in all transposons. They are flanked by variable-size long terminal repeats (LTRs) in both borders. The LTR retrotransposon order is divided into five superfamilies: *Copia, Gypsy, Bel/Pao, Retrovirus*, and *Endogenous Retrovirus (ERV)*.

Copia is the first superfamily and shows identical function and similar genome organization with other superfamilies of LTR retrotransposons. However, it differs from others in the position of the INT domain that located between the AP and RT domains. It is widespread in almost all organisms: plants, animals, fungi, etc. In addition to the classification proposed by Wicker et al. (2007), the International Committee on the Taxonomy of Viruses (ICTV) divided it into three genera: *Pseudovirus, Hemivirus*, and *Sirevirus*. In these genera, only *Sirevirus* have ENV domain in this superfamily.

The second superfamily is *Gypsy* and, similar to *Copia*, it is one of the most abundant groups in living organisms. Its INT domain is located between the RH domain and the 3′-LTR. Some retrotransposons that belong in this superfamily also have an ENV domain and resemble retroviruses. Detailed classification of the *Gypsy* superfamily is complicated. Previously, it was divided into two genera, as *Metaviridae* and *Errantiviridae*, but lately is divided as *Chromavirus* and *non-Chromavirus*.

The last three superfamilies of LTR retrotransposons—*Bel/Pao, Retrovirus*, and *Endogenous Retrovirus (ERV)*—have the same organization with internal domains as in *Gypsy*. They exist only in Metazoan genomes (for more details of these groups, see Eickbush and Jamburuthugoda 2008; Llorens et al. 2009).

3.1.2 Order 2: DIRS

This order is characterized by a tyrosine recombinase (YR) that is typically involved in site-specific recombinations between similar or identical DNA sequences, instead of INT in the POL region. Therefore, this order is also called tyrosine recombinase

retroelements. The DIRS order is divided into three superfamilies: DIRS, Ngaro, and VIPER (Goodwin et al. 2004; Lorenzi et al. 2006). Analyses of the reverse transcriptase sequences of the first superfamily of DIRS suggested that it was distantly related to the *Gypsy* superfamily. However, DIRS elements differ from *Gypsy* by their terminal repeats in addition to the YR domain: DIRS elements are bordered by TIRs whereas *Gypsy* has LTRs (Poulter and Goodwin 2005). The DIRS superfamily contains internal ORFs and an internal complementary region (ICR) derived from the duplication of TIR sequences (Piednoel et al. 2011). Only the DIRS superfamily of the DIRS order is present in plants. The other two superfamilies of the DIRS order, Ngaro and VIPER, have not been found in plants to date. Although orientation of their internal domains is the same with DIRS, their terminal repeats are present as a direct position as in LTR retrotransposons.

3.1.3 Order 3: PENELOPE (PLE)

This order contains only one superfamily, named Penelope. Up to now this order has been described in many animals, plants, protists, and fungi. Retrotransposons that belong this group also contain LTRs in both borders. However, these LTR sequences might be either direct or inverted orientations. This order has a coding region that contains RT and endonuclease (EN) domains. The RT of PLE is related to the RT of telomerase, but EN is similar to an intron-encoded endonuclease (Pyatkov et al. 2004).

3.1.4 Order 4: LINE

This order is also known as non-LTR retrotransposons because elements of this order have neither LTR nor TIR sequences at borders. They are widely distributed in eukaryotic genomes. They usually have a poly-A tail at the 3′-end and some deletions at the 5′-end. This order is divided into five superfamilies: R2, RTE, Jockey, L1, and I (Wicker et al. 2007). The first three superfamilies, R2, RTE, and Jockey, are unique for metazoans whereas the last two, L1 and I, are found in various organisms including plants. Because of their capability to encode the enzymes required for transposition, they are also accepted as autonomous non-LTR retrotransposons (Eickbush and Jamburuthugoda 2008). Although R2 encode RT and EN enzymes, others encode APE and RT. Only the last superfamily, I, has an RH domain together with APE and RT.

3.1.5 Order 5: SINE

The last order of class I transposons, is also known as non-LTR retrotransposons, similar to LINE. This order lacks both repeat regions, TIR and LTR, at the borders and has a poly-A tail. In contrast to the LINE order, SINE does not have internal domains. Because SINE does not encode their own reverse transcriptase, it has been

proposed that SINE uses the enzymatic machinery of LINE for transposition (Wallace et al. 2008; Kroutter et al. 2009). Therefore, elements belonging to the SINE order are accepted as nonautonomous non-LTR retrotransposons. This order shows similarity with reverse-transcribed RNAs transcribed by RNA polymerase III into tRNA, rRNA, and other small nuclear RNAs (Malik and Eickbush 1998). The SINE order is divided into three superfamilies, tRNA, 7SL, and 5S. The first two superfamilies are found in plantsbut the last one is not.

3.2 Class II Transposons (DNA Transposons)

DNA transposons have a different transposition mechanism than retrotransposons. They move through a DNA intermediate, in either the cut-paste or copy-paste mechanism. For this reason, DNA transposons are divided into two subclasses. The first subclass uses the cut-paste mechanism and the second uses copy-paste. Further, both these subclasses have two orders.

3.2.1 Subclass I Order I: TIR

This order shows characteristic structures of DNA transposons, a Tase enzyme that is flanked by two TIR sequences. During transposition, both strands of these elements are cut and inserted into new locations so that their copy numbers remain constant. However, if they transpose during DNA replication (S-phase of the cell cycle), the copy number can increase. The TIR order is divided into nine superfamilies: Tc1/mariner, hAT, Mutator, P, PIF/Harbinger, CACTA, Merlin, Transib, and PiggyBac. The first six superfamilies are present in plants but the last three are not (Wicker et al. 2007).

3.2.2 Subclass I Order II: Crypton

The order Crypton also uses the cut-paste transposition mechanisms as does the TIR order, but in contrast the Crypton order has neither TIR sequences nor Tase domain. Instead of Tase, they have an YR domain. It only has one superfamily, Crypton, that has been found only in fungi so far.

3.2.3 Subclass II Order I: Helitron

Although helitrons are DNA transposons, they use the copy-paste transposition mechanism as do retrotransposons. However, their transpositions do not include an RNA intermediate. This DNA-intermediated transposition mechanism is achieved by cleavage of just one strand of a DNA transposon. Then, a complementary strand

of released single-strand DNA is synthesized and inserted to a new location. The Helitron order has just one superfamily, Helitron. This is the only replicative DNA transposon type defined in plants so far.

3.2.4 Subclass II Order II: Maverick

The last order of transposons is Maverick, which also uses the DNA-mediated replicative form of transposition. This order has only one superfamily, which is not present in plants. Differing from Helitron, DNA transposons belonging to the Maverick superfamily have TIRs at both borders. These transposons can have 11 domains that encode proteins needed for transposition, but the number and orientation of these domains vary from one transposon to another.

3.3 Transposons in Different Plant Species

After their discovery and characterization in maize, transposons have been found in all organisms investigated so far, with just one exception: in some species of *Plasmodium*, transposons were not found (Vitte and Bennetzen 2006; Huang et al. 2012). Percentages and types of transposons in the genomes can vary among species (Feschotte and Pritham 2007). Although transposons compose 1–3 % of prokaryotic genomes, in eukaryotic organisms, such as plants, their percentage may reach 85 % or more. In animals, transposons constitute 3–45 % of the genome. The human genome contains about 45 % transposons, and the most abundant type is non-LTR retrotransposons (7–33 %). However, LTR retrotransposons are the most abundant transposon type in the genome of many organisms but they constitute just 8.5 % of the human genome (Cordaux and Batzer 2009). Generally plants, especially cereals, have the highest percentage of transposons compared with others, such as prokaryotes, animals, and mammals.

4 Transposon Markers

A molecular marker is defined as a particular segment of DNA that is representative of the differences at the genome level. An ideal molecular marker should be polymorphic and evenly distributed throughout the genome, generate multiple, independent, and reliable markers, be simple, quick, and inexpensive, need only small amounts of DNA samples, etc. (Agarwal et al. 2008). Many features of LTR retrotransposons in plant genomes have made them excellent sources of molecular markers (Kalendar and Schulman 2006; Poczai et al. 2013; Yuzbasioglu et al. 2016a). A schematic demonstration of some LTR retrotransposon-based marker techniques is given in Fig. 3. Retrotransposon sequences alone or combined with

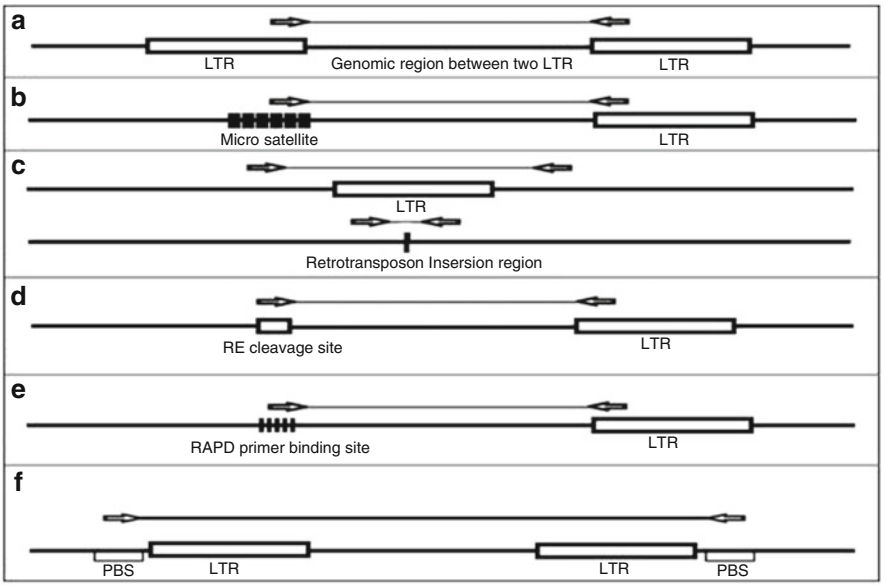


Fig. 3 LTR retrotransposon-based marker techniques: (**a**) IRAP, (**b**) REMAP, (**c**) RBIP, (**d**) SSAP, (**e**) RRAP, (**f**) IPBS (Gozukirmizi et al. 2015)

various sequences in the genome provide primer-binding sites. LTRs of a retrotransposon have conserved sequences between different organisms. Thus, a primer pair designed for a specific organism can be used for other organisms. However, different retrotransposons of an organism have different LTR sequences so a primer pair designed for a specific retrotransposon cannot bind to another LTR. In the following section, some of the transposon-based marker techniques are explained and compared, and their advantages and disadvantages are discussed.

4.1 Inter-Retrotransposon Amplified Polymorphism (IRAP)

IRAP is a multiplex and dominant technique. This technique was first developed using the *BARE-1* ('*BArley RetroElement 1*') sequence (Kalendar et al. 1999). The *BARE-1* family has approximately 1.5×10^4 full-length copies and 1.7×10^5 solo LTRs (Vicient et al. 1999). IRAP uses PCR primers designed in an outward direction from the conserved sequences of LTR. Internal regions between two LTRs or solo LTRs (without retrotransposon) are amplified. Band sizes of amplification products support the local clustering yet large-scale dispersion of *BARE-1* elements (Kalendar et al. 1999). Retrotransposons tend to be nested into another element in cereal genomes (Vicient et al. 2001). Targets (5′- or 3′-end of LTR) of primers produce different amplicons and alter banding patterns. The appearance of a band

reflects a new insertion that is in an amplifiable distance of another element. In this technique, DNA samples are amplified with a single or two primers. The major advantages of this technique are that restriction digestion or ligation steps do not exist and the polymerase chain reaction (PCR) products are separated on agarose gels. Because of the high band numbers, DNA amounts should be optimized. High template DNA concentrations may cause overamplification (Kalendar et al. 1999; Kalendar and Schulman 2006). Slow electrophoresis with high-quality agaroses improves the resolution of the bands. Impurities in DNA samples may interfere with PCR (Schulman et al. 2004).

4.2 Retrotransposon-Microsatellite-Amplified Polymorphism (REMAP)

This method is also a dominant and multiplex marker system. REMAP resembles IRAP but uses one LTR primer and a primer specific to a nearby microsatellite. Therefore, not only regions between two LTRs but also regions between an LTR and a microsatellite are amplified (Kalendar et al. 1999). Microsatellites, also known as simple sequence repeats (SSRs), are a type of repetitive sequences in plants and animals. These motifs can consist of a single base pair or a small number of bases (usually ranging from 1 to 6) that are repeated several times (Saha et al. 2003). SSRs are abundant, highly polymorphic, and dispersed throughout the genome (Li et al. 2002). SSRs also provide an independent molecular marker technique (Zietkiewicz et al. 1989). Moreover, SSRs are associated with retrotransposons in barley (Davila et al. 1999). Microsatellite primers in the REMAP method are different from original SSR primers that anneal to flanking regions. In other words, a LTR primer is combined with a primer consisting of SSR itself. PCR conditions and analysis of PCR products are similar to IRAP conditions (Kalendar et al. 1999; Kalendar and Schulman 2006). However, some bands may be derived solely from SSR primers (Leigh et al. 2003).

4.3 Retrotransposon-Based Insertional Polymorphism (RBIP)

RBIP was considered as an alternative to the multiplex S-SAP method. Individual transposon insertions are detected by RBIP using host-specific PCR primers and a transposon-specific primer. Two PCR primers are designed from the target (host) where the transposon is inserted and an additional transposon-specific primer was designed. DNA samples are amplified with either host primers or a host primer combined with a transposon primer. Insertion of a transposon alters the size of the locus. Polymorphism can be determined by agarose gel electrophoresis or dot-blot analysis. Hence, allelic states of a particular locus are determined. RBIP is locus

specific and codominant as are SSRs. The most important limitation to both RBIP and SSR markers is the need for sequence information from the chromosomal region surrounding each marker. Although SSR markers are obtained using by screening genomic libraries with simple sequence repeats, RBIP markers can be identified using the S-SAP approach (Flavell et al. 1998). DNA quality is not an issue (Schulman et al. 2004).

A major advantage of RBIP is that it can easily be automated, using gel-free procedures such as TaqMan or DNA chip technology to increase sample throughput.

4.4 Sequence-Specific Amplified Polymorphism (S-SAP)

S-SAP is a dominant and multiplex marker system. It is the first retrotransposon-based molecular marker and was modified from amplified fragment length polymorphism (AFLP) (Waugh et al. 1997). The AFLP technique is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA (Vos et al. 1993). In S-SAP, genomic DNA is first digested with two restriction endonucleases and ligated with adaptors specific to restriction endonucleases. The template is subjected to preamplification with primers from adaptors and then selectively amplified with two primers, one designed from a conserved region of a LTR and the other a selective adaptor primer. Selective adaptor primers have additional nucleotides (one to three) at the 3′-end. PCR products are denatured with gel loading buffer containing formamide and then resolved on polyacrylamide gel electrophoresis (PAGE). Although S-SAP generally yielded fewer PCR bands than AFLP, the polymorphism percentage was higher in S-SAP (Waugh et al. 1997). S-SAP is very similar to the transposon display (TD) method (van den Broeck et al. 1998).

4.5 RAPD Retrotransposon-Amplified Polymorphism (R-RAP)

R-RAP was proposed as a combination of RAPD (random amplified polymorphic DNA) and IRAP techniques, which may have some limitations. In the case of RAPD, low reproducibility may be a problem and in the case of IRAP, primer-binding sites may be too far apart to produce an amplification product. DNA samples are amplified using one random 10-mer primer and a LTR primer. Melting temperatures of RAPD and IRAP differ greatly; therefore, identification of optimal annealing temperature by gradient PCR is necessary. Amplification products are resolved on agarose gels. The number of R-RAP loci varies from 10 to 17 with high resolution and reproducibility (Aalami et al. 2012).

5 Sequencing and Transposomics

Transposon-insertion sequencing is a powerful technique for the rapid connection of genotype to phenotype, particularly in bacteria. A number of studies have improved our understanding of basic gene functions, establishing requirements for colonization and infection, mapping complex metabolic pathways, and exploring noncoding genomic regions (Barquist et al. 2013). Several methods were developed concurrently for high-throughput sequencing of transposon-insertion sites: transposon-directed insertion site sequencing (TraDIS) (Langridge et al. 2009), insertion sequencing (INSeq) (Goodman et al. 2009), high-throughput insertion tracking by deep sequencing (HITS) (Gawronski et al. 2009), and ransposon sequencing (Tn-seq) (van Opijnen et al. 2009), followed by Tn-seq circle (Gallagher et al. 2011). All these protocols have the same basic workflow with minor variations (Fig. 4) (Barquist et al. 2013).

The Tn-seq and INSeq methods are highly similar but INSeq includes a PAGE purification step following adaptor ligation and PCR whereas Tn-Seq includes an agarose gel purification at this point. Goodman et al. (2011) introduced additional steps to the original INSeq protocol: a linear PCR step using a biotinylated primer and subsequent purification of the product with magnetic streptavidin beads were added following adapter ligation. These steps reduce both the amount of sample and

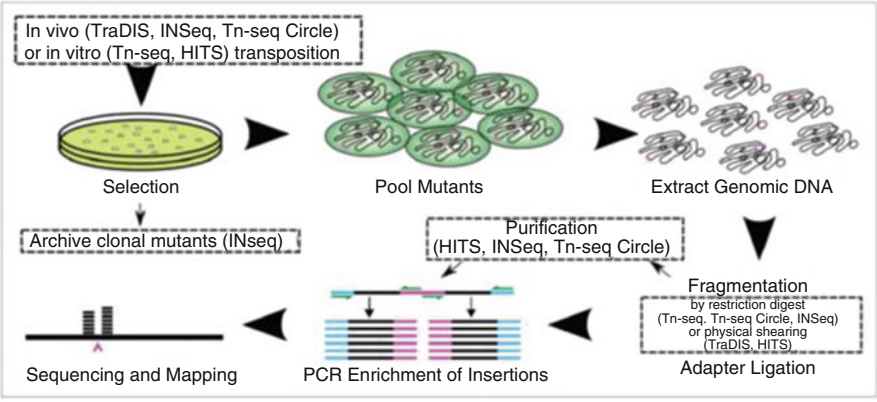


Fig. 4 An illustration of the typical workflow of transposon-insertion sequencing protocols. Transposons are represented by *pink lines*, sequencing adaptors by *blue*, genomic DNA by *black*, and PCR primers by *green*. Mutants are generated through either in vivo or in vitro transposition and subsequent selection for antibiotic resistance. These mutants are pooled, and genomic DNA is extracted and fragmented by restriction digest or physical shearing. Sequencing adaptors are ligated; some protocols then perform a step to purify fragments containing transposon insertions. Adaptor-specific primers are used to specifically enrich for transposon-containing fragments. The fragments are then sequenced and mapped back to a reference genome to uniquely identify insertion sites. *Dashed boxes* indicate steps that differ between protocols. (From Barquist et al. 2013)

the amount of enzymes needed. Although they make the protocol more laborious, the results suggest that these modifications increase the sensitivity of the technique. Moreover, HITS and TraDIS methods are more similar to each other than to Tn-seq and INSeq: after shearing of the DNA, the DNA ends are repaired, and a poly-A tail is added. However, the methods diverge after the PCR step; in HITS, the PCR products undergo size selection (on a gel) and affinity purification before sequencing, whereas in TraDIS, the PCR products are sequenced directly (van Opijnen and Camilli 2013).

Tn-seq was applied to the pathogen *Streptococcus pneumoniae* to identify genes that are essential for growth in rich medium, to determine the fitness effect (advantageous or disadvantageous) of each nonessential gene and to identify 97 genetic interactions between five query genes and the rest of the genome (van Opijnen et al. 2009). Moreover, Tn-seq has been used in *Pseudomonas aeruginosa* to identify antibiotic resistance genes (Gallagher et al. 2011) and *Mycobacterium tuberculosis* (Griffin et al. 2011) to identify essential genes and pathways that are involved in the utilization of cholesterol. A slight variation on the Tn-seq method, the so-called Tnseq circle, was used to track large numbers of transposon mutants of *Pseudomonas aeruginosa* to study its inherent resistance to the aminoglycoside antibiotics (Gallagher et al. 2011). The INSeq method was used to determine whether the human symbiont *Bacteroides thetaiotaomicron* harbors specific genes that are necessary for survival in the colon (Goodman et al. 2009). The HITS method was applied to a library of approximately 75,000 *Haemophilus influenzae* mutants (Gawronski et al. 2009). TraDIS was used to screen for essential genes in *Salmonella typhimurium*, the causative agent of typhoid fever (Langridge et al. 2009).

Transposon sequencing is not restricted to screening for the fitness effects of gene knockouts. In one study (Christen et al. 2011), this technology was used to construct a dense library of *Caulobacter crescentus* insertion mutants (with 8-bp resolution). In another study, by combining RNA-seq with Tn-seq, 56 new putative sRNAs in the noncoding regions of *S. pneumoniae* (Mann et al. 2012) were identified. Transposon sequencing has the potential to be used for other purposes. For example, probing for differences in genotype–phenotype interactions between strains or closely related species has the potential to reveal divergent evolutionary trajectories for shared genes or pathways, such as functional pathways that have undergone reconfiguration either in terms of their regulation or to accomplish new tasks (van Opijnen and Camilli 2013). Although it has been applied almost exclusively to bacterial species, transposon sequencing could be expanded to study other microorganisms. For example, it could be used to probe gene function and gene networks in viruses. All that is needed is a means of introducing transposon insertions into the viral genome; this could be accomplished by building an inducible transposition system within the host, such that insertions are introduced into the viral genome during viral replication (van Opijnen and Camilli 2013).

In vitro and in vivo transposition have been linked by the use of Transposomes™ (Goryshin et al. 2000). A transposome is the complex of transposase and transposon that is created in the early steps of the IVT (in vitro transposition) when the transposase attaches to the transposon for excision. At that point there is no recipient

DNA or cations to activate transposition. The transposome is introduced into a host by transformation and, because of the internal cation concentration, transposition into the host genome follows. This technology is dependent on a host transformation system and overcomes the host barrier for in vivo transposition and the need for homologous recombination (Hamer et al. 2001). In conclusion, the application of transposon sequencing give highly valuable information for discovering gene function, and further technological advances are likely to see it continue to be applied as a useful, high-throughput screening method for some time to come (van Opijnen and Camilli 2013).

6 Transposons and Gene Expression

Heterochromatin is commonly regarded as silent DNA. It consists of large regions of repetitive nucleotide sequences and transposons. Transposons, however, must be suppressed because they constitute two dangers for the genome: (1) their repeated units can cause spurious homologous recombination; and (2) their ability to transpose can lead to disruption or misregulation of important genes. Both these dangers are suppressed by heterochromatinization (Madlung and Comai 2004). Insertion of a transposable element into a gene causes mutation. A proper instance comes from human genetics: for example, retrotransposition of L1 element (LINE) into factor VIII gene causes hemophilia (Kazazian and Moran 1998). In the case of plants, transposon-like insertion into the starch-branching enzyme interrupts gene expression and causes wrinkled seeds in maize (Bhattacharyya et al. 1990). In addition, tases can fulfill important functions for the host as in the case of *Arabidopsis thaliana*. A tase, named DAYSLEEPER, binds to a specific motif in the upstream region of a repair gene. Plants lacking DAYSLEEPER exhibit developmental abnormalities (Bundock and Hooykaas 2005). Transposable elements produce small RNAs that regulate expression of specific genes in *Drosophila* and *Arabidopsis* (reviewed by McCue and Slotkin 2012). Insertion of a transposable element, *Mu*, generates a recessive mutation referred as *hcf106* that causes pale green seedlings and is lethal in maize. However, when the *Mu* element is inactivated by hypermethylation, plants exhibit a normal phenotype despite the continued presence of the transposon within the gene (Martienssen et al. 1990). Shortly after this publication, the same group reported that the *Mu* element lies within the 5′-untranslated leader of the *Hcf106* mRNA and, when active, this insertion interferes with the accumulation of *hcf106* mRNA (Barkan and Martienssen 1991). *Flowering locus C* (*FLC*) encodes a MADS domain protein and is a repressor of flowering in *Arabidopsis* (Michaels and Amasino 1999). The Landsberg *erecta* (Ler) ecotype of *Arabidopsis* has a Mutator-like (MULE) transposable element insertion in the first intron of FLC gene (Gazzani et al. 2003; Michaels et al. 2003), and this insertion causes FLC-Ler to be expressed at low levels (Michaels et al. 2003). It was proposed that microRNA (miRNA) genes have been evolved from miniature inverted-repeat transposable elements (MITEs). *hsa-mir-548*, a family of human miRNA genes, were found to be derived from

Made1 elements, which are members of MITEs (Piriyapongsa and Jordan 2007). miR441 and miR446, miRNA sequences of rice, were reported to have originated from *Stowaway1* (Li et al. 2011). TE-derived siRNA inhibits the formation of a host protein that represses TE activity (McCue et al. 2013).

7 Transposons and Plant Evolution

7.1 Genome Size

TEs have found to be important in genome structure and genetic diversity. For example, in some plant species such as maize, *Gossypium* spp., and *Oryza australiensis* (a wild relative of rice), TE amplifications have at least doubled the genome size within the last 5 million years (SanMiguel et al. 1998; Piegu et al. 2006; Hawkins et al. 2006). TE content strongly correlates with genome size variation. TE-derived DNA mostly makes up the 20–30 % of the genome even for plant species with small genomes such as *Brachypodium distachyon* and *Arabidopsis* spp. (The *Arabidopsis* Genome Initiative 2000; The International *Brachypodium* Initiative 2010). Species with larger genomes, such as maize and barley, have larger TE-derived DNA content, up to 85 % (Wicker et al. 2005; Schnable et al. 2009). Evidence suggests that repetitive DNA may be removed more slowly from species with larger genome sizes than smaller genomes. A study on *Copia* retrotransposons showed that they remained intact and active for much longer time periods in the larger barley and wheat genomes than the smaller *A. thaliana* and rice genomes (Wicker and Keller 2007). The TE proportion of the genomes not only differs between species but also within species. A comparison of a 1.2-Mb region of the *indica* and *japonica* subspecies of rice revealed that 13 % of the sequence was not shared and that the difference was caused by differential TE insertions (Ma and Bennetzen 2004). Again, the important differences in genome size between *A. thaliana* and *A. lyrata* have also shown to be caused by a reduction in TE activity and potentially more efficient TE elimination in *A. thaliana* (Hu et al. 2011). Moreover, two sequenced maize genomes revealed different genome sizes by 22 %, with 90 % of this difference caused by repetitive elements (Wang and Dooner 2006).

7.2 Genome Organization and Maintenance

Most eukaryotic organisms accommodate high numbers of retrotransposons in centromeres and telomeres. Plant centromeres contain infused TEs within short centromeric repeats (Ma et al. 2007). Pericentromeric regions, similarly, are composed mostly of silenced TEs and pseudogenes (Hall et al. 2006). A centromere-specific LTR-retrotransposon has been identified in rice (Cheng et al. 2002) and some

other grasses (Miller et al. 1998). Additionally, in maize, the centromere-specific retroelements have been demonstrated to have an interaction with the kitetochore protein CENH3 (Zhong et al. 2002). On the other hand, the exact role of these sequences in centromere or telomere function is still not clear (Joly-Lopez and Bureau 2014).

Allopolyploidization has been shown to be connected with rapid structural and functional alterations in genomes, especially in the repetitive content (Leitch and Leitch 2008). Broad epigenetic rearrangements are induced by TEs after polyploidy (Parisod et al. 2010). About one third of the identified imprinted genes in *A. thaliana* are located near transposable elements or repeated sequences that encounter demethylation. It could be implemented that transposon insertions near gene regulatory regions will recruit DNA methylation machinery that targets the invading foreign DNA and eventually cause silencing of the affected gene (Jians 2012). In plants, methylated TEs are found further away from genes than are unmethylated TEs (Hollister and Gaut 2009).

7.3 Generation of Variation and Evolutionary Innovation

Transposons have a significant role in constructing eukaryotic genomes by creating interspecies/intraspecies diversity and potential for adaptation to changing environments. Some TEs stimulate genome rearrangements, including inversion, duplication, or deletion of adjacent DNA, by chromosome breaking, aborted transposition, or ectopic recombination between homologous transposable elements at different chromosomal locations (Feschotte and Pritham 2007). The transposons not only contribute to the generation of allelic diversity in natural populations but also to shaping the genomic and epigenetic properties of their hosts and to the divergence of new genes. MULEs have been observed to capture genic sequences in both maize and *Arabidopsis* (Talbert and Chandler 1988; Yu et al. 2000). MULEs have also been shown to be a major feature of rice genome construction (Jiang et al. 2004).

TE insertions disrupt the coding sequence of a gene and inhibit the production of the gene product. However, insertions within promoters, introns, and untranslated regions can directly change the phenotype by genetic or epigenetic regulations (Feschotte and Pritham 2007). There are also transposase-induced rearrangements that have a strong potential for chromosome restructuring. Chromosomal inversions, duplications, and deletions of more than 100 kb can occur depending on the orientation of the TEs at the chromosomal location (Zhang and Peterson 2004). These events can lead to exon shuffling and create new functional genes. Studies on flowering plants demonstrate a high degree of gene creation by transposon capture and exon shuffling. Late-flowering *Arabidopsis* accessions possess an epigenetically silenced TE insertion at the first intron of the *FLC* (flowering locus C) gene (Liu et al. 2004). Wheat, barley, and maize flowering studies also highlighted the TEs on quantitative variation via modulation of gene expression (Yan et al. 2006;

Salvi et al. 2007). Additionally, as class II *Gyno-hAT* transposon causes methylation spreading to the promoter of the *CmWIP1* transcription factor gene, transition from male to female flowers occurs in melon (Martin et al. 2009).

The duplication of the host genes and exon shuffling by TEs may be important in the genome evolution. TEs use three modes: alteration of gene functions through insertion, induction of chromosomal rearrangements, and divergence of new genes and regulatory sequences (Feschotte and Pritham 2007). Transposons assist in the arrangement of the chromosomal domains through distinct epigenetic marks and transcriptional activity. These marks can be changed by genetic stress, such as DNA damage, interspecific hybridization or polyploidization, and environmental cues such as pathogen infection, and abiotic stresses help increase homologous recombination occurrence and chromosomal rearrangements. Such changes can release the transposons from epigenetic control, allowing stress-inducible TEs to propagate stress-inducible promoters to other genes through transposition (Ito et al. 2011) and can influence the genome organization both somatically and heritably (Boyko et al. 2010), offering an opportunity for natural selection to establish new chromosomal domains and regulatory circuits, eventually leading to speciation (Feschotte and Pritham 2007).

Plant resistance genes have been shown to be rapidly evolving. Race-specific resistance genes are often found in clusters forming large tandem repeats of polymorphic genes (Casacuberta and Santiago 2003). Likely, the rice *Xa21* gene family is demonstrated to incorporate a high number of TEs located within different genes (Richter and Ronald 2000). In plants, genes with similar or identical functions can be positioned in very different chromosomal locations in different species. For instance, the *adh1* gene is found in an entirely new chromosome in maize and sorghum when compared to more distant relatives such as rice or wheat, disrupting chromosomal synteny (Ilic et al. 2003).

The term TE exaptation can be referred to molecular domestication where donation of the protein-coding sequences contributes to the evolution of new genes with new functions and to the host phenotype. For instance, different genes captured by the same transposon are transcribed as a single compound mRNA, which can then evolve into new genes encoding novel proteins (Morgante et al. 2005). The host defends its genome against transposons via DNA methylation, small RNAs, cytosine deaminases, and DNA-repair factors (Levin and Moran 2011). In case TEs produce beneficial phenotypes, domesticated genes will be less silenced by the host silencing machineries. TE insertions can generate new promoters, enhancers, insulators, and regulatory regions. Analysis of TE-derived genes proposes that transposase-encoding DNA transposon sequences are the most frequently domesticated parts of TEs (Feschotte and Pritham 2007).

8 Our Work with Transposons

Our group used retrotransposon-based molecular markers mainly for analysis of somaclonal variation. The stability of aging barley calli and callus-regenerated shoots was investigated by IRAP using a primer derived from *BARE-1* (Evrensel et al. 2011;

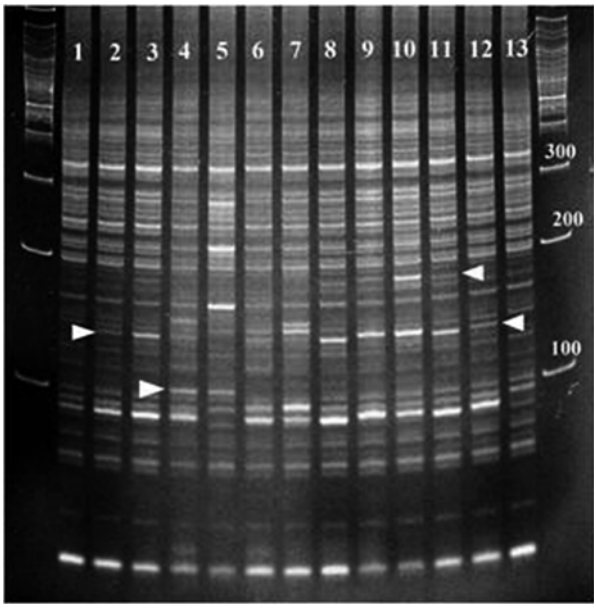


Fig. 5 Inter-retrotransposon amplified polymorphism (IRAP) profiles of calli and shoots for *BAGY2*. *1*, noncultured embryo; *2–13*, tissue culture materials (*2*, *6*, *10* 45-day-old calli; *3*, *7*, *11* shoots regenerated from 45-day-old calli; *4*, *8*, *12* 90-day-old calli; *5*, *9*, *13* shoots regenerated from 90-day-old calli. (From Yilmaz et al. 2014)

Yilmaz and Gozukirmizi 2013) and *Nikita* (Bayram et al. 2012) sequences. Callus culture conditions activate *BARE-1* and *Nikita* element. In the paper by Marakli et al. (2012), similarity of mature embryo, leaf, and root tissues grown from the same barley plant were investigated in terms of *BARE-1* and *BAGY2* movements. *BAGY2* was found to be more stable than *BARE-1*. Not all callus induction conditions increase retrotransposon activity (Temel and Gozukirmizi 2013). However, in addition to a *BAGY2* retrotransposon-specific IRAP polymorphism (Fig. 5), they also caused increase in copy numbers of internal domains of *BAGY2* (Yilmaz et al. 2014). Transformation of tobacco plant with the *dehE* gene, which degrades the herbicide Dalapon, was shown to cause activation of *Tto1* retrotransposon, one of the few active retrotransposons in tobacco (Kaya et al. 2013). Moreover, we investigated the nonautonomous retrotransposon *Sukkula* in barley (Kartal et al. 2014). Our group utilized the IRAP technique to assess the genotoxicity of some drugs, for example, epirubicine (Hamat-Mecbur et al. 2014) and amiprophos-methyl (Temel and Gozukirmizi 2014). Recently, we used IRAP markers for identification of variation in single seed derived leaves and roots in rice (Yuzbasioglu et al. 2016b).

9 Conclusions and Future Perspective

Transposable elements, particularly retroelements, are an important part of the genomes of complex organisms. Retroelements affect various aspects of human health (St. Laurent et al. 2010), but discussion of this impact is beyond the scope of

this chapter. Yet, review of recent developments in transposon research is essential. In this chapter, we tried to summarize the structure, types, and classification of transposons and present information about the most widely used transposon markers. We also mentioned the relationship of transposons with sequencing technologies, regulation of gene expression, and genome evolution. Data of our studies with transposons were briefly mentioned. We tried to cite as many papers as possible, but we apologize to those authors whose works went unmentioned in this chapter.

Acknowledgments We are grateful to the Research Fund of Istanbul University for financial support (Projects 20212 and 20316).

References

Aalami A, Safiyar S, Mandoulakani BA (2012) R-RAP: a retrotransposon-based DNA fingerprinting technique in plants. Plant Omics 5:359–364

Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 27:617–631

Barkan A, Martienssen RA (1991) Inactivation of maize transposon *Mu* suppresses a mutant phenotype by activating an outward-reading promoter near the end of *Mu1*. Proc Natl Acad Sci USA 88:3502–3506

Barquist L, Boinett CJ, Cain AK (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. RNA Biol 10:1161–1169

Bayram E, Yilmaz S, Hamat-Mecbur H, Kartal-Alacam G (2012) *Nikita* retrotransposon movements in callus cultures of barley (*Hordeum vulgare* L.). Plant Omics 5:211–215

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–627

Bhattacharyya MK, Smith AM, Noel Ellis TH, Hedley C, Martin C (1990) The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell 60:115–122

Boyko A, Blevins T, Yao Y, Golubov A, Bilichak A, Ilnytskyy Y, Hollander J, Meins F Jr (2010) Transgenerational adaptation of *Arabidopsis* to stress requires DNA methylation and the function of dicer-like proteins. PLoS One 5:e9514

Bundock P, Hooykaas P (2005) An *Arabidopsis* hAT-like transposase is essential for plant development. Nature (Lond) 436:282–284

Capy P, Bazin C, Higuet D, Langin T (1998) Dynamics and evolution of transposable elements. Library of Congress, Austin, TX

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene (Amst) 311:1–11

Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14:1691–1704

Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, Fero MJ, McAdams HH, Shapiroa L (2011) The essential genome of a bacterium. Mol Syst Biol 7:1–7

Coates CJ, Kaminski JM, Summers JB, Segal DJ, Miller AD, Kolb AF (2005) Site-directed genome modification: derivatives of DNA-modifying enzymes as targeting tools. Trends Biotechnol 23:407–419

Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10:691–703

Davila JA, Loarce Y, Ramsay L, Waugh R, Ferrer E (1999) Comparison of RAMP and SSR markers for the study of wild barley genetic diversity. Hereditas 131:5–13

Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134:221–234

Evrensel C, Yilmaz S, Temel A, Gozukirmizi N (2011) Variations in *BARE-1* insertion patterns in barley callus cultures. Genet Mol Res 10:980–987

Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103–107

Flavell AJ (1995) Retroelements reverse transcriptase and evolution. Comp Biochem Physiol B Biochem Mol Biol 110:3–15

Flavell AJ, Knox MR, Pearce SR, Ellis TH (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J 16:643–650

Gabus C, Ficheux D, Rau M, Keith G, Sandmeyer S, Darlix JL (1998) The yeast Ty3 retrotransposon contains a 5′–3′ bipartite primer-binding site and encodes nucleocapsid protein NCp9 functionally homologous to HIV-1 NCp7. EMBO J 17:4873–4880

Gallagher LA, Shendure J, Manoil C (2011) Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. MBio 2(1):e00315–e00310

Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. Proc Natl Acad Sci USA 106:16422–16427

Gazzani S, Gendall AR, Lister C, Dean C (2003) Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Physiol 132:1107–1114

Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD, Lozupone CA, Knight R, Gordon JI (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe 6:279–289

Goodman AL, Wu M, Gordon JI (2011) Identifying microbial fitness determinants by insertion sequencing using genome-wide transposon mutant libraries. Nat Protoc 6:1969–1980

Goodwin TJD, Poulter RTM, Lorenzen MD, Beeman RW (2004) DIRS retroelements in arthropods: identification of the recently active TcDirs1 element in the red flour beetle *Tribolium castaneum*. Mol Genet Genomics 272:47–56

Goryshin I, Jendrisak J, Hoffman L, Meis R, Reznikoff W (2000) Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. Nat Biotechnol 18:97–100

Gozukirmizi N, Yilmaz S, Marakli S, Temel A (2015) Retrotransposon-based molecular markers: tools for variation analysis in plants. In: Tashki-Ajdukovic K (ed) Applications of molecular markers in plant genome analysis and breeding. Research Signpost/Transworld Research Network, Ontario, pp 19–45

Griffin J, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, Sassetti CM (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog 7:e1002251

Hall AE, Kettler GC, Preuss D (2006) Dynamic evolution at pericentromeres. Genome Res 16:355–364

Hamat-Mecbur H, Yilmaz S, Temel A, Sahin K, Gozukirmizi N (2014) Effects of epirubicin on barley seedlings. Toxicol Ind Health 30:52–59

Hamer L, DeZwaan TM, Montenegro-Chamorro MV, Frank SA, Hamer JE (2001) Recent advances in large-scale transposon mutagenesis. Curr Opin Chem Biol 5:67–73

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Hollister JD, Gaut BS (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

Huang CRL, Burns KH, Boeke JD (2012) Active transposition in genomes. Annu Rev Genet 46:651–675

Ilic K, SanMiguel PJ, Bennetzen JL (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum and rice genomes. Proc Natl Acad Sci USA 100:12265–12270

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J (2011) An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature (Lond) 472:115–119

Jaaskelainen M, Mykkanen AH, Arna T, Vicient CM, Suoniemi A, Kalendar R, Savilahti H, Schulman AH (1999) Retrotransposon BARE-1: expression of encoded proteins and formation of virus-like particles in barley cells. Plant J 20:413–422

Jiang N, Bao Z, Temnykh S, Cheng Z, Jiang J, Wing RA, McCouch SR, Wessler SR (2002) *Dasheng*: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. Genetics 161:1293–1305

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature (Lond) 431:569–573

Jians H (2012) Evolution, function, and regulation of genomic imprinting in plant seed development. J Exp Bot 63:4713–4722

Joly-Lopez Z, Bureau TE (2014) Diversity and evolution of transposable elements in *Arabidopsis*. Chromosome Res 22:203–216

Kalendar R, Schulman AH (2006) IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat Protoc 1:2478–2484

Kartal G, Yilmaz S, Marakli S, Gozukirmizi N (2014) *Sukkula* retrotransposon insertion polymorphism in barley. Russ J Plant Physiol 61:828–833

Kaya Y, Yilmaz S, Gozukirmizi N, Huyop F (2013) Evaluation of transgenic *Nicotiana tabacum* with *dehE* gene using transposon-based IRAP markers. Am J Plant Sci 4:41–44

Kazazian HH, Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19:19–24

Kroutter EN, Belancio VP, Wagstaff BJ, Roy-Engel AM (2009) The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. PLoS Genet 5:e1000458

Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G et al (2009) Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. Genome Res 19:2308–2316

Leigh F, Lea V, Law J, Wolters P, Powell W, Donini P (2003) Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. Euphytica 133:359–366

Leitch AR, Leitch IJ (2008) Genomic plasticity and the diversity of polyploid plants. Science 320:481–483

Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and their hosts. Nat Rev Genet 12:615

Li Y, Li C, Xia J, Jin Y (2011) Domestication of transposable elements into microRNA genes in plants. PLoS One 6:e19212

Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in *Arabidopsis*. Genes Dev 18:2873–2878

Llorens C, Munoz-Pomer A, Bernad L, Botella H (2009) Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol Direct 4:41

Lorenzi HA, Robledo G, Levin MJ (2006) The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. Mol Biochem Parasitol 145:184–194

Ma JX, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Ma J, Wing RA, Bennetzen JL, Jackson SA (2007) Plant centromere organization: a dynamic structure with conserved functions. Trends Genet 23:134–139
Madlung A, Comai L (2004) The effect of stress on genome regulation and structure. Ann Bot 94:481–495
Mak J, Kleiman L (1997) Primer tRNAs for reverse transcription. J Virol 71:8087–8095
Malik HS, Eickbush TH (1998) The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol Biol Evol 15:1123–1134
Malik HS, Eickbush TH (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR-Retrotransposable elements and retroviruses. Genome Res 11: 1187–1197
Mann B, van Opijnen T, Wang J, Obert C, Wang YD, Carter R, McGoldrick DJ, Ridout G, Camilli A, Tuomanen EI, Rosch JW (2012) Control of virulence by small RNAs in *Streptococcus pneumoniae*. PLoS Pathog 8:e1002788
Marakli S, Yilmaz S, Gozukirmizi N (2012) *BARE1* and *BAGY2* retrotransposon movements and expression analyses in developing barley seedlings. Biotechnol Biotechnol Equip 26:3451–3456
Martienssen R, Barkan A, Taylor WC, Freeling M (1990) Somatically heritable switches in the DNA modification of Mu transposable elements monitored with a suppressible mutant in maize. Gene Dev 4:331–343
Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A (2009) A transposon induced epigenetic change leads to sex determination in melon. Nature (Lond) 461:1135–1138
McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792–801
McCue AD, Slotkin RK (2012) Transposable element small RNAs as regulators of gene expression. Trends Genet 28:616–623
McCue AD, Nuthikattu S, Slotkin RK (2013) Genome-wide identification of genes regulated in trans by transposable element small interfering RNAs. RNA Biol 10:1379–1395
Michaels SD, Amasino RM (1999) *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11:949–956
Michaels S, He Y, Scortecci KC, Amasino R (2003) Attenuation of *FLOWERING LOCUS C* activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. Proc Natl Acad Sci USA 100:10102–10107
Miller JT, Dong F, Jackson SA, Song J, Jiang J (1998) Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics 150:1615–1623
Monden Y, Yamaguchi K, Tahara M (2014) Application of iPBS in high-throughput sequencing for the development of retrotransposon-based molecular markers. Curr Plant Biol doi:10.1016/j.cpb.2014.09.001
Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A (2005) Gene duplication and exon shuffling by Helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997–1002
Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10:149–155
Ohtsubo H, Kumekawa N, Ohtsubo E (1999) *RIRE2* a novel *gypsy*-type retrotransposon from rice. Genes Genet Syst 74:83–91
Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol 186:37–45
Piednoel M, Goncalves IR, Higuet D, Bonnivard E (2011) Eukaryote DIRS1-like retrotransposons: an overview. BMC Genomics 12:621
Piegu B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Piriyapongsa J, Jordan IK (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One 2:e203
Poczai P, Varga I, Laos M, Cseh A, Bell N, Valkonen JPT, Hyvonen J (2013) Advances in plant gene-targeted and functional markers: a review. Plant Methods 9:6
Poulter RT, Goodwin TJ (2005) DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet Genome Res 110:575–588
Pouteau S, Grandbastien MA, Boccara M (1994) Microbial elicitors of plant defence responses activate transcription of a retrotransposon. Plant J 5:535–542
Pyatkov KI, Arkhipova IR, Malkova NV, Finnegan DJ, Evgen'ev MB (2004) Reverse transcriptase and endonuclease activities encoded by *Penelope*-like retroelements. Proc Natl Acad Sci USA 101:14719–14724
Richter TE, Ronald PC (2000) The evolution of disease resistance genes. Plant Mol Biol 42: 195–204
Saha S, Karaca M, Jenkins JN, Zipf AE, Reddy UK, Kantety RV (2003) Simple sequence repeats as useful resources to study transcribed genes of cotton. Euphytica 130:355–364
Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci USA 104:11376–11381
Sanmiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45
Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115
Schulman AH, Flavell AJ, Ellis THN (2004) The application of LTR retrotransposons as molecular markers in plants. Methods Mol Biol 260:145–173
Slotkin RK, Nuthikattu S, Jiang N (2012) The impact of transposable elements on gene and genome evolution. Plant Genome Divers 1:35–58
St. Laurent G III, Hammell N, McCaffrey TA (2010) A LINE-1 component to human aging: do LINE elements exact a longevity cost for evolutionary advantage? Mech Ageing Dev 131:299–305
Szabo M, Kiss J, Olasz F (2010) Functional organization of the inverted repeats of IS30. J Bacteriol 192:3414–3423
Takeda S, Sugimoto K, Otsuki H, Hirochika H (1998) Transcriptional activation of the tobacco retrotransposon *Tto1* by wounding and methyl jasmonate. Plant Mol Biol 36:365–376
Talbert LE, Chandler VL (1988) Characterization of a highly conserved sequence related to Mutator transposable elements in maize. Mol Biol Evol 5:519–529
Temel A, Gozukirmizi N (2013) Analysis of retrotransposition and DNA methylation in barley callus culture. Acta Biol Hung 64:86–95
Temel A, Gozukirmizi N (2014) Genotoxicity of metaphase-arresting methods in barley. Turk J Biol. doi:10.3906/biy-1405-58
The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature (Lond) 408:796–815
The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature (Lond) 463:763–768
van den Broeck D, Maes T, Sauer M, Zethof J, de Keukeleire P, D'hauw M, van Montagu M, Gerats T (1998) Transposon display identifies individual transposable elements in high copy number lines. Plant J 13:121–129
van Opijnen T, Camilli A (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11:435–442
van Opijnen T, Bodi KL, Camilli A (2009) Tn-seq: highthroughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6:767–772
Vicient CM, Suoniemi A, Anamthawat-Johnsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769–1784

Vicient CM, Jaaskelainen M, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125:1283–1292

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1993) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Wallace N, Wagstaff BJ, Deininger PL, Roy-Engel AM (2008) LINE-1 ORF1 protein enhances Alu SINE retrotransposition. Gene (Amst) 419:1–6

Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. Proc Natl Acad Sci USA 103:17644–17649

Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BB, Powell W (1997) Genetic distribution of *Bare-1*-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253:687–694

Weil C, Martienssen R (2008) Epigenetic interactions between transposons and genes: lessons from plants. Curr Opin Genet Dev 18:188–192

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in *Triticeae*, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. Genome Res 17:1072–1081

Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, Graner A, Stein N (2005) A detailed look at 7 million years of genome evolution in a 439-kb contiguous sequence at the barley *Hv-eIF4E* locus: recombination, rearrangements and repeats. Plant J 41:184–194

Wicker T, Sabot F, Hua-Van A, Benetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al (2007) A unified classified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wilhelm M, Wilhelm FX (2001) Reverse transcription of retroviruses and LTR-retrotransposons. Cell Mol Life Sci 58:1246–1262

Wright DA, Voytas DF (1998) Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/*gypsy* retrotransposons that encode envelope-like proteins. Genetics 149:703–715

Yadav VP, Mandal PK, Rao DN, Bhattacharya S (2009) Characterization of the restriction enzyme-like endonuclease encoded by the *Entamoeba histolytica* non-long terminal repeat retrotransposon EhLINE1. FEBS J 276:7070–7082

Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene *VRN3* is an orthologue of FT. Proc Natl Acad Sci USA 103:19581–19586

Yilmaz S, Gozukirmizi N (2013) Variation of retrotransposon movement in callus culture and regenerated shoots of barley. Biotechnol Biotechnol Equip 27:4227–4230

Yilmaz S, Marakli S, Gozukirmizi N (2014) *BAGY2* retrotransposon analyses in barley calli cultures and regenerated plantlets. Biochem Genet 52:233–244

Yu Z, Wright SI, Bureau TE (2000) Mutator-like elements in *Arabidopsis thaliana*: structure, diversity and evolution. Genetics 156:2019–2031

Yuzbasioglu G, Yilmaz S, Gozukirmizi N (2016a) Houba retrotransposon-based molecular markers: a tool for variation analysis in rice. Turk J Agric For doi:10.3906/tar-1509-2

Yuzbasioglu G, Yilmaz S, Marakli S, Gozukirmizi N (2016b) Analysis of *Hopi/Osr27* and *Houba/Tos5/Osr13* retrotransposons in rice. Biotechnol Biotech Eq 30:213–218

Zhang J, Peterson T (2004) Transposition of reversed Ac element ends generates chromosome rearrangements in maize. Genetics 167:1929–1937

Zhao D, Ferguson AA, Jiang N (2016) What makes up plant genomes: The vanishing line between transposable elements and genes. Biochimica et Biophysica Acta 1859:366–380

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell 14:2825–2836

Zietkiewicz E, Rafalski A, Labuda D (1989) Genome singerprinting by simple sequence sepeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183

Next-Generation Sequencing: Advantages, Disadvantages, and Future

Şule Ari and Muzaffer Arikan

Contents

Abstract It has been more than 35 years since the development of the groundbreaking method for DNA sequencing by Frederick Sanger and colleagues. This revolutionary study triggered the improvement of new methods that have provided great opportunities for low-cost and fast DNA sequencing. Strikingly after the Human Genome Project, the time interval between each sequencing technology started decreasing while amount of scientific knowledge has continued growing exponentially. Considering Sanger sequencing as the first generation, new generations of DNA sequencing have been introduced consequently. The development of the next-generation sequencing (NGS) technologies has contributed to this trend substantially by reducing costs and producing massive sequencing data. Hitherto, four sequencing generations have been defined. Second-generation sequencing that is currently the most commonly used NGS technology consists of library preparation, amplification, and sequencing steps while in third-generation sequencing,

Ş. Ari (✉) • M. Arikan
Molecular Biology and Genetics Department, Faculty of Science, İstanbul University, Vezneciler, İstanbul 34134, Turkey
e-mail: sari@istanbul.edu.tr

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_5

individual nucleic acids are sequenced directly in order to avoid biases and have higher throughput. Recently described fourth-generation sequencing aims conducting genomic analysis directly in the cell. Classified to different generations, NGS has led to overcome the limitations of conventional DNA sequencing methods and has found usage in a wide range of molecular biology applications. On the other hand, plenty of technical challenges, which need to be deeply analyzed and solved, emerged with these technologies. Every sequencing generation and platform, by reason of its methodological approach, carries characteristic advantages and disadvantages which determine the fitness for certain applications. Thus, assessment of these features, limitations, and potential applications help shaping the studies that will determine the route of omic technologies.

Keywords Next-generation sequencing • Genomics • High-throughput sequencing • Massively parallel sequencing • DNA-sequencing technologies

1 Introduction

Today's life sciences focus broadly on the interpretation of the relationship between the genome and phenotype from the cellular level to complex biological process like development or diseases. After the enunciation of "central dogma" for action mechanism of genes by Crick (1958), we can now assess the impact of omic technologies on description of the complexity of a functioning living cell. In conjunction with the Human Genome Project (HGP), omic technologies provided us large data sets that are paving the way toward a comprehensive and holistic understanding of genotype–phenotype interaction in 2000s. Since genes and genome structure together with their functions are essential for biological activities, elucidation of entire genome sequence was the main purpose of the HGP. Before the genomic revolution, use of genome-mapping approaches required probability statistics to identify the gene positions, followed by positional cloning to identify the underlying genes. Therefore, omic technologies were utilized mainly in the first instance for genome sequencing, and increasing efforts involved improvements in molecular genetic technology such as gene cloning, contig construction, and DNA sequencing.

Even though technologies for sequencing DNA, RNA, and proteins were available since 1970s (Sanger et al. 1977), revolutionary improvements in DNA-sequencing techniques have given rise to a large amount of DNA sequences. Developments in algorithms to analyze the vast amount of data, computer science, nanobiology, robotics, and information technology, as well as bioinformatics have helped high-throughput (HT) DNA sequencing. Genome sequencing has become synonymous with HT sequencing hereafter. The completion of a human genome reference sequence allowed next-generation sequencing (NGS). Since 2005, next generation of sequencing instruments that substantially reduced DNA sequencing

time and cost, and remarkably increased data-production capacity, have been introduced by commercial manufacturers. As their details will be discussed in this chapter, NGS technologies rely on a set of methods for DNA template preparation, massively parallel reading of sequenced millions to billions short DNAs, real-time image capturing, alignment of sequences, sequence assembly, and variant detection. Current NGS technologies differ mainly on methods for clonal amplification and sequencing of DNA fragments. Each has specific advantages for criteria: read length, accuracy, run time, and throughput. In addition to analysis of DNA sequences, progression of sequencing technologies has resulted in analysis of other biological components such as RNA and protein, as well as how they interact in complex cellular networks. NGS technologies allowed new biological research areas—finer analysis of transcriptome dynamics, genome structure, genomic variation—(Soon et al. 2013) and provided a chance to researchers to study systematically the complex interactions of structure and function of biological systems and to solve important biological questions which was not possible before. Expected benefits from these understanding are—in a broad range—applications such as personalized medicine, molecular cloning (from biotechnological point of view), breeding, and comparative and evolution studies. In spite of widely and routine application of HT sequencing technologies in biology and healthcare, challenges still remain for individualistic sequencing.

In this chapter, NGS technologies are reviewed and the advantages and disadvantages specifically associated with each system are discussed. The chapter is organized as follows: we outline the historical development of DNA sequencing elaborating Sanger sequencing in the next section. In the consequent three sections, we introduce various types of second-, third-, and fourth-generation sequencing technologies, respectively. In the sixth section, we present new technologies that are under development and discuss the possible new directions of research. Finally, in conclusion section, we summarize the current state of sequencing technologies and point to the horizon where new approaches, analyses, and tools may be needed.

2 Sanger Sequencing: The First Generation

From discovery of double helix structure of DNA (Watson and Crick 1953) till introduction of the dideoxy method based on chain termination commonly known as “Sanger method” (Sanger et al. 1977), researchers had to cope with multifarious difficulties in DNA sequencing. Need for short sequences, which can be noted as an important limitation also for today’s sequencing technologies, was one of the main handicaps. Although there were some approaches developed to obtain short sequences, it was not yet possible to locate them in chains (Sanger 1988). Ergo, most of the studies had been conducted targeting bacteriophage genomes which are comparatively small thereby providing suitable experimental systems to temporarily overcome sequence length limitation (Metzker 2005). Withal, absence of

restriction enzymes was another impediment preventing substantial improvements in DNA-sequencing field. Despite the studies on RNA sequencing in which ribonuclease T1 targeting G motifs could be used, there was lack of an enzyme sharing similar concept for DNA (França et al. 2002). Also, the structure of DNA that consists of only four different nucleotides can be considered as a particular difficulty, since it complicated the separation of sequences from each other. Consequently, while the first protein sequencing was successfully performed for insulin already in early 1950s (Sanger and Tuppy 1951), the first DNA sequences barely could be obtained—after almost two decades—in an exceptional study by Wu and Kaiser who sequenced cohesive 3′ ends of bacteriophage lambda employing DNA polymerase for copying DNA (Wu and Kaiser 1968). Then, Sanger and Coulson introduced a new rapid approach to sequence DNA that is called "plus and minus method" to overcome explained obstacles (Sanger and Coulson 1975). This method was based on activity of DNA polymerase I being able to produce a complementary strand starting from a primer sequencing in presence of deoxynucleotide triphosphates (dNTPs) using ^{32}P labeling for one of them. After producing a mixture of different-sized DNA molecules having the same 5′ end, the products were purified and new reaction cycles were applied. To perform these reactions, the produced chains were separated into eight samples with equal volume and nucleotide incorporation by DNA polymerase was blocked by addition of three of dNTPs called as "minus" or one of the nucleotides called as "plus." Then, gel electrophoresis was used for differentiation of the DNA chains and determination of the sequence contents (Hutchison 2007) Although it brought important improvements to DNA sequencing, plus and minus method is time consuming and has certain disadvantages like obtaining no bands for internal positions of sequences but for the beginning and end of runs.

In 1977, Sanger and colleagues published first DNA sequences of bacteriophage ϕX174 using revolutionary "the dideoxy method based on chain termination" technique that caused substantial changes in genomics field (Sanger et al. 1977). The *Sanger method* requires a mixture of four dNTPs and a dideoxynucleotide triphosphates (ddNTP), with single-strand DNA and DNA polymerase. The synthesis of complementary DNA is performed by DNA polymerase using dNTPs and the enzyme terminates the reaction in the presence of ddNTP. The optimization of sequence length of nested fragments obtained through the reaction is achieved by testing different ddNTP/dNTP ratios. According to the information of ddNTP that is employed to terminate the reaction, sequencing products are discriminated on a gel and results are analyzed to determine the complete DNA sequence (Metzker 2005). In the original method, ^{32}P molecule is utilized to label sequences and four lanes were needed for each sample due to usage of one type of dye for all ddNTPs. Although it was more efficient compared to other methods, original Sanger method was also time consuming and carrying a risk by cause of the radioisotopes used for labeling (França et al. 2002). Thus, throughout the years many modifications have been introduced to this first protocol to enhance efficiency and ease the procedures. Among notable advances, deoxyadenosine 5′-(α-[S]thio)triphosphate incorporating

into DNA sequences was proposed for labeling to increase resolution and sharpness of bands on autoradiography a few years later (Biggin et al. 1983). A decade after development of Sanger method, chemiluminescent labeling was proposed as a new method instead of radioactive isotopes bringing the advantage of sequencing PCR products (Beck et al. 1989). In this method, an oligonucleotide sequence bound with a biotin molecule in 5′end is used as primer. Alkaline phosphatase is bound to oligonucleotide from 5′end with a streptavidin conjugate and starts a chemical reaction that causes emission of photons which are detected on a photographic film. Other advantages of this approach include no requirement of cloning before sequencing, capability of conducting multiple reactions in one lane and performing sequence reading by using specific primers for each reaction. Smith et al. (1986) developed a system based on four different florescent dyes bound to 5′ end of a primer and correlated each dye with one ddNTP—fluorescein isothiocyanate (FITCH) for A reaction, NBD aminohexanoic acid for T reaction, tetramethylrhodamine isothiocyanate (TM) for G reaction, and Texas Red for C reaction. Hence, all reactions could be loaded to one lane as a postsequencing process and checked through gel. An alternative approach including attachment of four fluorescent dyes to ddNTP molecules instead of sequencing primers (so all reactions could be performed in one lane) eliminated the need for application of reactions in four separate lanes and pooling them later for gel imaging process (Prober et al. 1987). Development of polymerase chain reaction can be mentioned as a significant advance to simplify sample preparation (Mullis et al. 1986). Employing T7 DNA polymerase for amplification and internal labeling approach is another improvement that increased the resolution for obtained DNA sequences (Wiemann et al. 1996). Alternative fluorescent dyes showing differences in several features such as mobility in gel, fluorescence intensity, and lifetime were proposed and used for better resolution and efficiency (Müller et al. 1997; Flanagan et al. 1998). Moreover, since the development of Sanger method, usage of slab gels was time-consuming step and needed to be improved and automatized; therefore, new systems utilizing automated reloading of the capillaries with polymer matrix instead of slab gels were also introduced these years.

Automated DNA sequencing using approaches explained above was introduced in the middle of 1980s to make sequencing steps, such as sample preparation, electrophoresis, data analysis, practical, and boost data-production speed. The optimization of these devices and emergence of new techniques caused a considerable rise both in quality and data generation capacity for DNA sequencing. The first commercially automated DNA sequencer ABI Prism 310 was announced by PE Biosystems in 1996. This single *capillary instrument* allows automatization of all steps of sequencing from introduction of florescent labeled dyes till separation of sequenced fragments through capillary electrophoresis using performance optimized polymer 6 (POP-6) and determination of DNA sequences by data analysis. 400–450 bases of sequencing data can be produced in less than 1 h by rapid protocol of this instrument while it can reach over 600 bases with standard protocols (Watts and MacBeath 2001). Two years later, 96 capillary sequencers of GE

Healthcare (MegaBACE 1000) and PE Biosystems (ABI Prism 3700) that improved the outputs strikingly entered the sequencing market. Thanks to modifications introduced to injection processes, microfabricated sequencing devices allowed to overcome certain disadvantages of first automated instruments like background and resolution problems. In addition, a variety of materials used for the production of microchannels of these devices let accomplishing better sequencing quality and reduced the time (Paegel et al. 2003; Kan et al. 2004). Development of these new devices not only solved efficiency and resolution drawbacks but also increased sequencing data amount tens of times higher than first sequencing systems and decreased the costs per reaction (Metzker 2005). The original Sanger method is widely accepted and commonly used due to its open character to all these great modifications and improvements thereby allowing completion of big genome projects in a relatively short time of the history of science while started from little genomes at beginning.

As new methods and modifications have been developed, total throughput and costs have decreased gradually for Sanger sequencing. In 1994, average 100 kb data-production per person-year would result $1.5 per base net cost (Chen 1994). These automated DNA sequencers were successfully employed in the HGP and reduced both costs and time need to complete the project. Currently, it is possible to produce about 250 kb data (600–1000 bp DNA sequences) in a single run performing 384 reactions in parallel with Sanger sequencing and the average cost is about $0.0004 per base (Wetterstrand 2014). Total sequencing costs have decreased continuously in correlation with new advances and applications to the original method.

Sanger sequencing provided remarkable opportunities to life sciences and shaped and improved the knowledge about nucleic acids accordingly helped better understanding of cellular mechanisms and diseases. However, this method has certain limitations including data-production amount and speed, sequencing quality, obstacles in answering different applications of genomics, and labor-intensive character of the protocol. Discovery of new mechanisms and concepts in molecular biology and new horizons in almost every branch of life sciences has formed their own myriad of needs for sequencing technology. At this point, new game changer for DNA sequencing has emerged, NGS.

3 Second-Generation Sequencing

In 2000s, the concept of DNA sequencing underwent drastic changes. Particularly, it needs to be mentioned that shotgun sequencing approach introduced during HGP which includes random fragmentation and sequencing of DNA then utilizing computer programs for assembly of different overlapping reads caused an expansion in the perceptions by forming the idea of massively parallel sequencing. All NGS systems, without any exception, have rise on revolutionary idea of shotgun sequencing which perfectly fit with the goals of HGP, development of cheaper and faster sequencing technologies. Consequently, just a few years after completion of HGP,

NGS platforms based on different approaches have been released to the market. So far proposed second-generation sequencing systems are based on either "sequencing by synthesis" or "sequencing by ligation."

3.1 Pyrosequencing

Pyrosequencing that comprises "sequencing by synthesis" reaction and—unlike chain termination in Sanger sequencing—employs pyrophosphate (PPi) release was developed by Ronaghi et al. (1996). After almost 10 years of introduction of different modifications to this novel method, Rothberg and colleagues presented the first NGS sequencer based on pyrosequencing approach. The three main problems of the original pyrosequencing method consisted of dATP usage by luciferase, high signal background after every cycle and difficulties in sequencing of GC-rich templates. First of this issues has been solved by use of dATPαS instead of dATP as this molecule is a better substrate for luciferase (Ronaghi et al. 1998). In order to solve signal background problem that was mostly because of the inefficient removal of dNTPs after each sequencing cycle, apyrase (an enzyme degrades all four nucleotides) was employed (Nyrén 2007). Sequencing of GC-rich regions presented a difficult problem due to secondary structure formations. Hence, addition of single-strand binding protein was employed to address this difficulty (Ronaghi 2000). Due to its increased popularity, a lot of other improvements have been proposed to efficiency, throughput, and practicality. Easy use of double strand DNA as sequencing template, glycine betaine for increasing the temperature of sequencing reaction to 37 °C thus obtaining better enzyme performances, development of sepharose bead/vacuum protocol to increase robustness of the system and make it possible to prepare 96 samples in a short time can be mentioned among these important improvements (Novais and Thorstenson 2011).

In 2005, Rothberg and colleagues developed the first commercial NGS instrument utilizing massively parallel pyrosequencing. As most of the NGS technologies does, pyrosequencing involves sequencing of a DNA strand during the sythesis of complementary strand by DNA polymerase. By usage of a surface containing nanometer size wells (called PicoTiter Plate, PTP), each DNA fragment is sequenced in one well. The library preparation step of this method starts with fragmentation of DNA sample or production of the amplicons with suitable sizes. Addition of adapter and barcode sequences is the next basic step of library construction. Then, emulsion PCR (emPCR) approach that includes generation of a water-in-oil emulsion and then clonal amplification of each DNA molecule using a water droplet as a microreactor environment is employed for amplification of tagged DNA sequences. After that, the barcoded and clonally amplified sequences are used for sequencing procedure which comprises addition of four nucleotides, one for each cycle, to growing template DNA and recording signal profile obtained from sequencing surface using a charge coupled device (CCD) camera. After addition of matching nucleotide by DNA polymerase, PPi that is substrate of ATP sulfurylase when adenosine

5′ phosphosulfate exists in the environment. Thus, PPi is converted to ATP by ATP sulfurylase and luciferase uses this ATP to produce oxyluciferin from luciferin which causes production of light. Between each cycle, apyrase degrades nucleotides and ATP remained in the environment. In every cycle, CCD camera detects signals from spots on PTP which shows position of unique DNA fragments and recorded images are used for base calling process which includes quantitative correlation of light signals with specific order of added nucleotides and determine nucleotide content of every DNA fragment.

454, the first NGS platform, was developed and put on the market by Roche in 2005 and after 3 years, the company presented GS FLX Titanium series. 454 sequencing platform can produce 400 Mb per run with an average 400 bp read length. In 2011, the GS FLX was upgraded to FLX+ which is able to reach up to 1000 bp read length and 700 Mb throughput per run which takes 23 h in this system. In 2010, to provide a benchtop sequencer, Roche introduced GS Junior platform which yields approximately 40 Mb per run. GS FLX and GS Junior platforms are suitable for a wide range of applications including whole genome (for small genomes, bacteria, virus, etc.), exome, transcriptome, amplicon sequencing, sequence capture, or metagenomics. While sequencing accuracy for FLX+ platform is reported as 99.997 % at 15× coverage, one of the well-known problems of this platform is high sequencing error rate in homopolymeric regions especially when it is consisted of equal or more than five base repeats. This phenomenon is caused by basic mechanism of the method in which incorporation of more than one nucleotide in one sequencing cycle is possible. Since addition of one type of nucleotide only allows prevention of incorporation by other nucleotides, it is quite possible that when there is repeat region in the template fragment, more than one nucleotide can be incorporated by DNA polymerase. In perfect conditions, the intensity of light observed would be expected to rise due to number of incorporated nucleotides. However, total signal density and its correlation with added nucleotides may be misinterpreted by the analysis system due to noise background. Another disadvantage of this method is time-consuming sample preparation protocols which include emPCR. On the other hand, the long read length advantage of this platform makes it a good candidate for applications like amplicon sequencing or de novo sequencing of small genomes while whole genome sequencing of complex genomes such as human genome requires longer time and higher costs. Roche Diagnostics Corporation announced shutting down 454 updates and gradually finish the production of reagents of 454 platforms.

3.2 Sequencing by Reversible Terminator Chemistry

Coupling capillary electrophoresis with ddNTPs labeled by four different fluorescent dyes vastly facilitated DNA-sequencing procedures and hastened later improvements (Luckey et al. 1990). However, as a result of its nature, the limitations of sequencing technology based on dye terminators have become more noticeable in

large scale sequencing studies since to increase output of sequencing runs, one should increase the number of capillary tubes which prevents establishment of high-throughput sequencing systems (Chen 2014). For that reason, a variety of modifications for dye terminators has been exploited to allow design of HT platforms. The idea of employing reversible dye terminators in DNA sequencing to widen boundaries presented by Sanger sequencing initially was raised by researchers of Columbia University (Li et al. 2003). The development and usage of first nucleotide analogue dUTP analogue (dUTP-PC-BODIPY 5) in sequencing reaction was yielded a highly efficient incorporation and satisfying fluorescence signal. Moreover, cleavage of fluorescent dye was performed successfully and signal background noise was reduced notably. After obtaining successful results, it was concluded that four nucleotide analogues tagged with different fluorescent dyes can be used to conduct efficient massive sequencing by synthesis. Consequently, several reversible dye terminators have been introduced each of which carries specific advantages and limitations owing to its interaction with DNA polymerase and efficiency of incorporation and cleavage reactions (Wu et al. 2007; Bentley et al. 2008; Pushkarev et al. 2009; Litosh et al. 2011; Gardner et al. 2012). Reversible dye terminators developed so far, can be classified into two main classes—blocked and unblocked terminators—according to binding region of blocking chemical group. Reversible terminators can be formed by either attaching a chemical group to five carbon (pentose) sugar and linker region between base and fluorescent dye (blocked) or to the base together with fluorescent dye (unblocked) and cleaved after incorporation of nucleotide during chain elongation. Comparing different types, blocked reversible terminators show higher performance in termination process while unblocked ones are more efficient for sequence elongation being a better substrate for DNA polymerase with absence of any scar after cleavage of blocking group (Chen et al. 2013). Blocked reversible terminators are employed in second-generation sequencing systems whereas unblocked reversible terminators are developed for third-generation sequencing systems. For example, Illumina Solexa, a second-generation sequencing platform using reversible terminator chemistry, developed that type terminator 3′-O-azidomethy to utilize for DNA sequencing (Bentley et al. 2008). On the other hand, Helicos Biosciences Corporation developed an unblocked terminator named "virtual terminator" for its third-generation sequencing platform (Bowers et al. 2009). The cleavage of the unblocked reversible terminators can be accomplished using chemical treatments or UV (Pushkarev et al. 2009; Litosh et al. 2011).

General mechanism of reversible terminator chemistry-based sequencing technology consists of three main steps: library preparation, clonal amplification, and sequencing by synthesis. The procedures start with construction of library including DNA fragments with suitable sizes and tagging each fragment by adapter and index sequences. Clonal amplification is performed on a solid surface where primer sequences that are complementary to adapter sequences are immobilized to create clusters representing each unique DNA fragment and provide sufficient signal for during imaging process. Sequencing step involves nucleotide addition by DNA polymerase, washing away unincorporated nucleotides, signal detection, removal of fluorescent and terminator groups, and washing away all remnants. DNA polymerase

adds one of four nucleotides labeled with different fluorescent dyes and containing a 3′ blocking group to growing DNA chain. Next, the unincorporated nucleotides are washed away. After signal detection by fluorescent imaging, 3′ blocking group and fluorescent dye is cleaved from nucleotide structure thus DNA polymerase can add new nucleotides in the next cycle. Another washing step is conducted to remove all chemical remnants that may interfere with sequencing reaction later cycles (Bentley et al. 2008).

Sequencing by reversible terminator chemistry is currently the most commonly used NGS technology worldwide. NGS platforms of Illumina Inc. rely on sequencing technology consisting of bridge amplification on solid surfaces (Adessi et al. 2000) developed by Manteia Predictive Medicine and reverse termination chemistry and engineered polymerases (Bennett 2004) developed by Solexa. Combining bridge amplification and reversible termination technologies, Solexa team sequenced bacteriophage ϕX174 genome producing over 3 million bases in a single run and 1 year later, launched Genome Analyzer, first NGS platform based on reverse termination chemistry and produces 1 gigabase (Gb) data per run (Bennett et al. 2005). After acquiring Solexa, Illumina released new sequencing platforms and expanded data throughput. Currently, the company offers MiSeq, NextSeq 500, HiSeq 2500 platforms producing 15, 120 and 1000 Gb sequencing data per run and having maximum 2×300 bp, 2×150 bp, and 2×125 bp read length, respectively. Furthermore, each of these platforms provide solutions for different spectrum of applications thereby covering solutions for a variety of scientific questions.

The flow of procedures for Illumina platforms start with conversion of DNA sample to fragments with acceptable sizes. Library preparation step continues with addition of specific adapter and index sequences of Illumina systems to each DNA fragment. Then, DNA fragments are loaded to a flow cell containing (immobilized to the surface) two types of primer sequences that are complementary to adapters attached to the fragments in library preparation in order to amplify each fragment with a reaction named “bridge amplification.” After binding to the primers on the surface, the complementary sequence is produced and template strand is removed. After that, surface attached DNA strand bends over and anneals to the closest complementary primer, a new strand is synthesized and the replication is repeated. Consequently, millions of clusters consisting of clonally amplified fragments are formed on the flow cell. After removing one type of fragments, DNA polymerase, nucleotides containing 3′ blocking group and a flourophore, and first sequencing primers are added to perform sequencing reaction. DNA polymerase adds suitable nucleotide to the growing chain, unincorporated nucleotides are washed away and using a laser fluorescent attached to the incorporated nucleotide is activated, signal is detected by a CCD camera. After cleavage of blocking group and removal of fluorescent washing step is repeated and continue next cycle. Index sequences are read between two sequencing period. A barcode specific primer is released to the reaction and index sequence of each fragment is determined. To start second read, the synthesized complementary strands are removed with denaturation and bridge amplification is conducted. After amplification, opposite strands of fragments are

removed with chemical cleavage and sequencing reaction starts again binding reverse primer (second sequencing primer) and explained steps are followed.

Although it undoubtedly alleviated many problems of Sanger sequencing, reversible termination chemistry has brought its specific advantages and disadvantages. First of all, this technique increased throughput in DNA sequencing while reducing time spent per run drastically (Buermans and den Dunnen 2014). Today, it is possible to obtain 1 terabase (Tb) data per day with very small amount of input sample using a commercial platform based on this chemistry. The read lengths reached to 300 bp paired-end in some Illumina platforms (MiSeq) from 25 bp single-end reads of first platform produced by Solexa which can be noted as a significant improvement succeeded in one of the main disadvantages compared with FLX 454 and Sanger sequencing. The accuracy of Illumina platforms is reported as 99.9 % and standard reagents let barcoding up to 96 samples per run (Morey et al. 2013). Another strong side is that this technology shows a better performance in sequencing of homopolymeric regions compared with FLX 454 and Ion Torrent platforms since the method allows incorporation of one nucleotide per reaction thanks to blocking groups (Mardis 2013). Instead, substitution errors are more commonly observed in Illumina systems due to noise background growing each sequencing cycle (Hutchison 2007). Also, after cleavage of blocking group, scars remained on nucleotide structure which eventually caused interaction with proteins and decreased efficiency of sequencing reactions (Chen et al. 2013). Another problem about Illumina systems was GC bias introduced in bridge amplification step (Mardis 2013). These limitations originated from the nature of the method have been reduced with enhancements in its chemistry. Although engineering of DNA polymerase and rearrangement of flow cell channels has provided better accuracy and cluster densities, read length limitation still stays as the main issue for reversible terminator chemistry-based sequencing which presents noticeable obstacles especially in de novo sequencing (Chen et al. 2013).

Illumina platforms enable a variety of applications in both DNA and RNA sequencing. Whole genome, exome, targeted, and de novo sequencing can be listed as main DNA-sequencing applications while total RNA, mRNA, and small RNA sequencing are the basic applications in RNA sequencing. Methylation and ChIP sequencing is also among the most commonly used applications. Among the sequencing platforms of the company considering its output per run, HiSeq is suitable for whole genome sequencing of organisms having relatively bigger genome such as human whereas NextSeq is especially promoted for exome sequencing projects. MiSeq is the benchtop sequencer of the company presenting practical applications of different fields in which relatively smaller data output is required. It is also preferable platform for de novo sequencing of small genomes such as bacterial or viral genomes since it provides longest read lengths. Putting all together, while Illumina technology has its own advantages and disadvantages explained above, platforms cover almost every field in genomics and present different alternatives to address questions of scientific studies.

3.3 Sequencing by Ligation

Sequencing by ligation utilizes the enzyme DNA ligase to determine nucleotide sequence of target DNA whereas sequencing technologies explained so far employ DNA polymerase. The concept of this method firstly demonstrated by resequencing of a strain of *Escherichia coli* wherein the execution of base calls with 10^{-6} error rate was succeeded thus brought out a new, inexpensive, and accurate sequencing technology (Shendure et al. 2005). Also known as polony sequencing, this technology employs mate paired library construction by tagging DNA fragments, amplification by emPCR on bead surface to construct "polonies," immobilization in a polyacrylamide gel on a microscope glass, sequencing of immobilized short tagged DNA fragments by ligation, and florescent imaging due to specific labels of nucleotide sequences on the sequencing surface (Porreca et al. 2006).

SOLiD (Sequencing by Oligonucleotide Ligation and Detection), a commercial second-generation NGS platform, based on a modified and enhanced version of polony sequencing was put the market in 2007 by ABI and SOLiD 5500xl was introduced after 3 years. At the beginning, SOLiD platform was able to produce more sequencing data then platforms provided by Illumina. However, recent enhancements in Illumina's technology and release of HiSeq 2500 and HiSeq X Ten which allows to sequence much more DNA fragments in a single run terminated the superiority of SOLiD in data-production per run. Furthermore, at the beginning, the SOLiD had 35 bp read length thus generating 3 Gb sequencing data per run. A complete run can be finished 7–14 days depending on the chemistry. By different modifications added since 2005, the read length of the sequencing by ligation-based platforms reached 2×50 paired read lengths and 320 Gb output. A sequencing run can be finished 7–14 days depending on the chemistry. The current commercial SOLiD platforms (5500 and 5500xl W) has up to 99.99 % accurate reads after filtering due to its unique approach.

Library construction for sequencing by ligation bears a resemblance to the one for pyrosequencing. emPCR is employed for clonal amplification and enhanced beads are immobilized on a glass surface which is possible due to a modification formed at 3′ end of target DNA molecules. The sequencing reaction in SOLiD platforms depends on octamer nucleotide sequences which are used as detection probes and DNA ligase activity. Carrying definite two bases at 3′ end, the rest of each octamer sequence is consisted of degenerate bases which can bind any DNA sequence. According to specific combinations of two bases at 3′ end, octamer sequences compete to be ligated to template by DNA ligase activity. The sequence reaction starts with binding of universal sequencing primers to adapter sequences. After ligation of specific octamer to the template, 3 nucleotides at the 5′ end which are linked with a specific fluorescent dye for detection of two bases annealed to target molecule are removed. Then, a new octamer sequence is added to growing DNA chain in the next cycle. The sequences of target DNA molecules are determined including 3 nucleotide gaps between two base reads determined per cycle.

These gaps are filled in the next sequencing part with a shifting approach. First universal primers and synthesized DNA strand is removed by denaturation to start second sequencing part. After that, second universal primers are released to the surface. Second sequencing primer is designed to be composed of a nucleotide sequence causing one base shift in the sequencing reaction since the first base at the 3′ end of the second universal primer binds to the second base at the 5′ end of the adapter sequence. Accordingly, every octamer sequence added to the growing chain is one base shifted. The sequencing reaction is completed in 5 parts with 5 universal primers, one base shifted after first universal primer so all bases of DNA templates are read two times with this sequencing approach which increases accuracy of SOLiD sequencing systems.

Substitutions are the most common errors in sequencing by ligation-based applications (Shendure and Ji 2008). The higher background error rate due to biases and incorporation mistakes during the amplification step can be mentioned as another problem for this method. Moreover, beads that have a mixed DNA fragments instead of one unique fragment decreases the quality of the reaction and causes reduction in filtered data. Also, short distance between beads can cause misreading thus create false reads and low quality bases. In addition, decrease of signal intensity which is a common phasing effect and problems in removal of fluorescent dye cause an accrual in error rate as the ligation cycles continue (Kircher and Kelso 2010). A new technique named wildfire has started to be used instead of technically challenging emPCR procedures and carries potential to decrease time for library preparation and increase output due to efficient use of sequencing cells (Stranneheim and Lundeberg 2012). Although it provides whole genome, exome, transcriptome, methylation, ChIP, small RNA sequencing, the SOLiD system is best suited for resequencing projects demanding low error rates and transcriptome sequencing (Metzker 2010). On the other hand, the data analysis problem of SOLiD platforms points to strong need for new methodologies to be developed (Bao et al. 2011). Moreover, while the accuracy of the system is quite high, short read lengths, and obstacles in sequencing of palindromic sequences is still an important problem for this technology (Huang et al. 2012).

4 Third-Generation Sequencing

Second-generation sequencing technologies are currently the most commonly used NGS platforms. However, to reduce well-known limitations of these platforms such as high costs, biases resulting from library amplification step and time-consuming protocols, new methods forming “third-generation sequencing” have appeared in the past years. The main difference between second and third-generation sequencing technologies is that third-generation sequencing systems are mostly based on direct detection of nucleotide composition of target DNA molecules without any amplification step. Library preparation step is skipped, single DNA fragments

tagged with adapters are read and nucleotide incorporation is detected through different approaches. Single molecule fluorescent sequencing, single molecule real-time sequencing, semiconductor sequencing, and nanopore sequencing can be listed as third-generation sequencing technologies.

4.1 Single Molecule Fluorescent Sequencing

As reversible terminators employed for second-generation sequencing platforms carry certain drawbacks explained above, in 2009, virtual terminators for third-generation sequencing technologies have been introduced (Bowers et al. 2009). The basic working mechanism is similar for all reversible terminators. However, blocking groups and fluorescent dyes to modify nucleotides show different characteristics according to their structure and nucleotide binding region. The general features of virtual terminators are consisted of 3′ free hydroxyl group which makes interaction with DNA polymerase possible and a fluorescent dye bounded to the removable linker group. Virtual terminators are employed in HeliScope sequencing platform developed by Helicos Bioscience Company in single molecule fluorescent sequencing (Kumar et al. 2005; Korlach et al. 2010).

The fragmentation of DNA sequences to suitable lengths for sequencing platform is the first process in the protocol of HeliScope technology. The DNA libraries containing 100–200 bp fragments should be prepared for sequencing step of HeliScope platform. After generation of a poly A tail at the 3′ end of each fragment by an enzyme called terminal transferase, prepared libraries are bound to the primer sequences containing only dTTP nucleotides and attached to a solid surface. To prevent reading poly A tails, dTTP molecules are added to the reaction with virtual terminators of other nucleotides in the first step of sequencing reaction. As a result, any non-dTTP molecule that is encountered by DNA polymerase and added to the growing DNA chain terminates the reaction. Sequencing reaction takes place in consistent with the basic principles of sequencing by reversible terminator chemistry. Every nucleotide analogue added to template DNA by polymerase enzyme blocks the sequencing reaction and imaging process is conducted using four CCD camera and a confocal microscope after removal of fluorescent dye and blocking group (Thompson and Steinmann 2010).

HeliScope platform removes need for complex and time-consuming protocols for library preparation which is considered as an important handicap for second-generation sequencing systems. Thus, it is an advantageous alternative for especially RNA Seq applications which tend to be affected by PCR biases. On the other hand, the main disadvantage of this platform is relative shortness of sequencing reads. A HeliScope platform can produce 35 Gb data per run including 35 bp long sequences. To reduce high error rates caused by this technology, repetitive sequencing runs should be performed which increases costs per application. Helicos Biosciences filed for bankruptcy in 2012 but still DNA and RNA sequencing services are provided by a company called SeqLL. Despite its advantages for some applications and technological improvements, HeliScope systems could not have a broad use for sequencing.

4.2 Single Molecule Real-Time Sequencing

Single molecule real-time sequencing (SMRT) relies on sequencing by synthesis approach and real-time detection of incorporated fluorescently labeled nucleotides. Pacific Biosciences (PacBio) introduced this sequencing method in 2009 as first third-generation sequencing technique (Eid et al. 2009). Uninterrupted DNA polymerase activity and addition nucleotides labeled by different fluorescent dyes is monitored by an imaging system. Single molecule real-time sequencing technology does not require any library preparation step since single DNA molecules are read in this system (Schadt et al. 2010).

Although it shares sequencing by synthesis concept with some second-generation sequencing systems, SMRT technology differs in many aspect due to its distinct features. At first, DNA samples are fragmented and tagged with adapter sequences skipping clonal amplification of each DNA molecule before sequencing. Tagged DNA molecules are directly loaded to the sequencing surface and then immobilized DNA polymerase adds appropriate nucleotide to template DNA. Fluorescent dye linked to phosphate group is removed naturally and imaging process of single molecules can be performed by use of a technology called zero mode waveguide (ZMW), a light focusing dense array reducing signal noise background (Hui 2014).

Amplification originated biases introduced to sequencing data is not an obstacle for SMRT systems. In addition, an important increase in read lengths caused notable enhancements in de novo sequencing studies since short reads bring mistake in the assembly of DNA regions including repeats and GC-rich regions (Bahassi and Stambrook 2014). On the other hand, high error rate is an undeniable limitation for SMRT technology. SMRT technology has 5 % error rate which especially includes insertion and deletion mistakes thus causing errors in resequencing and de novo assembly processes (Roberts et al. 2013). Read length distribution of SMRT technology is different then second-generation sequencing technologies since the method depends on uninterrupted activity of DNA polymerase so when DNA polymerase and template DNA separates from each other, the sequencing process ends (Quail et al. 2012). PacBio platforms are able to detect methylation status or RNA splicing pattern without any chemical treatment (Song et al. 2012). Because of this reason, it enables researchers to conduct experiments in a variety of fields without introducing any intermediate stage. Comparing with second-generation sequencing platforms, SMRT technology produces much less data per run since it depends on sequencing of single molecules. This difference brings need of new approaches especially statistical methods for analysis of third-generation sequencing platform data (Schadt et al. 2010). PacBio currently offers, targeted sequencing, de novo sequencing, base modification detection, isoform sequencing for any organisms including microorganisms, plants, and animals.

The latest sequencing platform of Pacific Biosciences is PacBio RS II may allow over 14,000 bp read length for more than 50 % of the reads and up to 40,000 bp while throughput per run is approximately 400 Mb. The system detects minor variant having lower frequency than 0.1 % and error profile shows differences from second-generation sequencing systems by not having a specific type of error in the

DNA sequences. Hence, increasing coverage for a region sequenced with low quality can work effectively to reduce error rates. One of the main advantages of SMRT technology is that base modification status and RNA-based researches can be performed by using this method to get unbiased, higher quality results. Also, SMRT technology solves drawbacks of assembly process of de novo sequenced genomes by providing much longer reads which make it possible to create scaffolds in repeat regions. As a result, single molecule real-time sequencing, owing to its characteristic advantages and disadvantages, provide a practical and accurate alternative to second-generation sequencing platforms.

4.3 Semiconductor Sequencing

Searches for new DNA-sequencing techniques have brought quite different approaches and commercial platforms to the sequencing market. While introduced to the market by Ion Torrent (acquired by Life Technologies) in 2010 with Ion Torrent Personal Genome Machine (PGM) sequencing platform, semiconductor sequencing concept was developed at 2006 by Toumazou and colleagues (Rothberg et al. 2011). Semiconductor sequencing is based on measurement of proton release during nucleotide incorporation by sequencing by synthesis. Direct measurement of pH changes in the microenvironment eliminates the time-consuming imaging step by a special camera. Considering its different properties, semiconductor sequencing technology is classified either as second or third-generation sequencing technology. Although it includes amplification step before sequencing, due to its unique and new sequencing methodology, this technology is classified as third-generation sequencing technology (Srinivasan and Batra 2014).

Semiconductor sequencing shares similar protocol flow with pyrosequencing-based FLX 454 system. The sample preparation starts with fragmentation of DNA molecules, blunt ended, and tagged with adapter sequences. Then, emPCR is employed to clonally amplify each DNA molecule of the NGS library. The beads carrying amplified DNA sequences are immobilized on a chip surface. However, the sequencing process presents differences than 454 platform. Ion Torrent PGM platform employs pH measurement-based detection of nucleotide incorporation in each cycle. Instead of imaging of fluorescent labels of modified nucleotides, Ion Torrent technology uses unmodified nucleotides. Thus, detection of specific nucleotide is performed by adding nucleotides in a certain order. Owing to this approach, the sequencing and "base calling" processes are finished in shorter time but specific errors in homopolymeric regions are similar with FLX platform as a result of saturation of pH detector which causes misinterpretation of signals produced by >4 base homopolymers. The average error rate for semiconductor sequencing platforms is approximately 1 % while average read length is 400 bp (Mardis 2013). Currently, Ion Torrent platforms can produce up to 10 Gb sequencing data per run which is completed in approximately 2.5 h and it is planned to increase output per run to

64 Gb with new updates. Considering run time advantages current Ion Torrent systems can be compared Illumina's HiSeq platforms for total throughput capacity.

Ion Torrent platforms can be used mainly for targeted, exome, transcriptome, de novo, small RNA sequencing, viral, and bacterial typing studies. The increased output of these systems make them convenient for applications like exome or whole genome sequencing. However, high error rate for specific regions is still an important obstacle for Ion Torrent technology (Morey et al. 2013).

4.4 Nanopore Sequencing

Nanopore sequencing is a new technology that basically depends on DNA sequence translocation through nanometer size pores by applying an electric field and measuring physical changes (Rusk 2014). The basic concept of this technology was presented in the article published by Kasianowicz et al. (1996) in which translocation of DNA molecules through α-hemolysin nanopore was accomplished and the possibility of sequencing DNA or RNA molecules due to characteristic changes during this process was shown. Moreover, the stability and geometry of the pore, the speed of procedure, and the features of the signal detection system determine the efficiency of nanopore sequencing-based platforms (Wang et al. 2014). To optimize these main parameters, many nanopore sequencing approaches has been introduced so far. Although biological membrane systems like α-hemolysin and *Mycobacterium smegmatis* porin A (MspA) were used for nanopore sequencing in earlier studies successfully (Kasianowicz et al. 1996; Derrington et al. 2010), specific obstacles such as instability and dimension tuning still remain to be overcome. Thus, along with studies to improve biological membrane systems, studies focusing solid-state nanopores have been accelerated in recent years. To date, researchers have tried to create nanopores in acceptable size and structure by using a variety of materials including molybdenum disulfide (Liu et al. 2014), boron nitride (Liu et al. 2013a), single walled carbon nanotubes (Liu et al. 2013b), hafnium oxide (Larkin et al. 2013), glass (Li et al. 2013), silicon nitride (Heng et al. 2004), organic polymer (Siwy and Fuliński 2002), and graphene (Fischbein and Drndić 2008) to investigate their potential for DNA sequencing. Furthermore, the speed of translocation of molecules through nanopores and detection of signals during this movement have been optimized by various modifications. Exonuclease assisted nanopore sequencing, hybridization-based sequencing, sequencing by expansion, and Nano-Tag SBS sequencing can be mentioned as approaches that were improved mostly to address drawbacks of signal detection and introduce vital enhancements till now. Also, original and innovative methodologies such as sequencing by electronic tunneling, direct tunneling, hydrogen bond-mediated tunneling, measurement of transverse conductance of DNA bases, concurrent detection of ionic current blockage and other signals, and field effect transistor were proposed to achieve signal detection in nanopore sequencing systems (Wang et al. 2014). Inasmuch as nanopore

sequencing technology is highly expected to provide new opportunities for cheap, fast, and accurate DNA sequencing, it continues to attract attention of many companies and research groups for research and innovation studies.

Nanopore sequencing-based platforms have not yet been put on market. However, companies are racing to deliver first commercial nanopore sequencing platform. For instance, Genia Technologies acquired by Roche declared that they will provide sequencing chips with $100 cost in 2015 while Oxford Nanopore Technologies still continue MınION Access Program which was started in 2014 to enable researchers exploring features of MinION sequencing device and contribute improvement of its features. Thus, results of the studies focusing on characteristics and applications of Oxford Nanopore Technology have become available in recent months. Although it is still early to reach definite conclusion for this technology, some of published studies can be mentioned to give a hint about its characteristics. In a study of de novo sequencing and assembly of *Escherichia coli* genome, it was reported that MinION device produces 35 million reads per run and average read length is 6 kb (Loman et al. 2015). In another study to sequence *Saccharomyces cerevisae* genome, it was stated that 490 Mb sequencing data per run while average read length was 5743 bp (Goodwin et al. 2015). Furthermore, the provider company states, it is possible to produce reads with length up to 50 kb.

The important factors affecting the output of Oxford Nanopore systems are the average length of the libraries prepared for sequencing, run time, and performance. The specific modifications in the analyses of data produced by these platforms could increase performance as indicated in recent studies (Jain et al. 2015). Advantageous characteristics like longer read lengths show that nanopore sequencing can be considered an important alternative in specific applications. Be that as it may, it is still early to make proper evaluation of error rate, average output-read length, and other features of nanopore sequencing platform. Considering its potential and distinct advantages, nanopore sequencing technology which can be used for DNA, RNA, or protein analyses is expected to be more widely used and have broad application spectrum.

5 Fourth-Generation Sequencing

Fourth-generation sequencing systems have made in situ sequencing possible in fixed tissue and cells by use of second-generation sequencing technologies (Mignardi and Nilsson 2014). The study conducted by Ke et al. (2013) for multiplex gene expression profiling and analyses of point mutations in breast cancer tissue sections using in situ sequencing has provide principal concepts of this sequencing generation. The methodology of this study includes use of padlock probes to encircle short targets and filling gaps between arms of padlock probes by polymerase and ligase activities. After that, rolling circle amplification is employed for clonal amplification and ligation dependent sequencing is performed in final step. In another study, random hexamers tagged with sequencing adapters were used to produce cDNAs and self circularization before rolling circle amplification (Lee et al.

2014). In this method that can be used in many different cell types for production of the amplicons which bind cellular proteins through covalent bonds, target mRNA molecules were sequenced accurately by SOLiD technology-based sequencing method. In fourth-generation sequencing, RNA amount in the cell can cause intensity problems for sequencing process as it depends on differentiation of two different spots on a layer. This important point determines the physical limitation for this application. A solution was proposed by Ke et al. (2013) to this problem in which introduction of random mismatching primers decreases intensity of reading and allows sequencing only some of fragments in the library pool in each cycle. This approach provided up to 400 reads per reading cell which makes it possible to determine expression of thousands of genes in the cell simultaneously for different types of RNA molecules including mRNA, noncoding RNA, rRNA, and antisense RNA (Mignardi and Nilsson 2014).

Instead of representing an alternative option to be used in a broad application field when compared with other sequencing technologies and generations, fourth-generation sequencing is expected to be useful in certain applications. Particularly, the methodology carries advantages for the applications for which analysis of cell populations with single-cell resolution can provide important benefits. When compared with single cell sequencing, in situ sequencing makes screening whole cell population with single cell resolution possible and provides more detailed results owing to this profiling approach. On the other hand, the problems about standardization, cost effectiveness, practicality, and full integration to current sequencing systems need to be solved to increase efficiency thus usage of these methods. Considering the speed of improvements in NGS systems, these problems do not seem to take long time to be overcome.

6 Future Perspectives

While available sequencing platforms are updated continuously, research and innovation studies to create cheaper, faster, and more accurate sequencing technologies are being performed unabated. Nowadays several technologies based on nanopore sequencing are under development in different countries of the world. Base4, Genia, INanoBio, Nabsys, Noblegen Biosciences, Oxford Nanopore Technologies, Quantapore, and Quantum Biosystems can be listed among companies working in nanopore sequencing field. To briefly explain methodologies and concepts, for example, Nabsys is based on a method in which semiconductor-based nanodetectors are employed for direct electrical detection of tagged DNA while Genia is preparing to present a platform depending on biological membranes and optical detection systems. INanoBio has a nanopore technology that prevents decrease in signal by Fully Depleted Exponentially Coupled (FDEC) Field Effect Transistor (FET) nanowire sensors to increase sensitivity and selectivity. Moreover, Quantum Biosystems uses an approach that combines nanopore sequencing with tunneling electron detector. On the other hand, Base4 employs a chemical cascade reaction for detection of single nucleotides cleaved and separated by water–oil emulsion.

Noblegen Bioscience is currently developing a system called optipore. Also, Quantapore is developing a nanopore-based optical read out method. It should be noted that aforementioned companies and methodologies form a limited list of technologies that are currently under development as it is not possible to deeply analyze and discuss all of them in this study.

In addition to R&D activities of companies, government agencies, and foundations support many groups with considerable grants for development of new sequencing technologies as well. For example, only in 2014, NHGRI has provided US $14.5 million to research groups for development of new sequencing techniques expecting a substantial decrease in genome sequencing costs (less than US $1000) and extending application fields in order to allow introduction of new advances in clinic. The official and commercial investments in this field is expected to increase coming years thereby letting research studies bring better platforms to the market.

Since 2004, NHGRI has supported tens of groups as part of $1000 genome project. Hence, the cost of genome sequencing that was $3 billion for HGP announced in 2001 has been reduced substantially and testing $1000 border nowadays. On the other hand, a deceleration in the reduction of DNA sequencing costs can be observed since 2012 compared with the period between 2007 and 2012. Moreover, although new technologies that will be introduced to market seems to cause decrease of costs continuously in coming years and increase data-production levels, an obvious bottleneck of sequencing accuracy for all NGS technologies remains to be solved during this period. In addition, the considerable increase in throughput will bring new obstacles that will take to time find practical solutions for but a number of studies addressing these issues are being performed to provide new alternatives.

7 Conclusion

DNA-sequencing field has been witnessed many revolutionary advances in past 40 years. The dideoxy method developed by Sanger and colleagues can be seen as a beacon for incredible change in genomics field. Still the gold standard in the routine use, this method has changed the outlook on the sequencing concept. Then, the announcement of completion of HGP opened an era of HT and fast sequencing platforms that are called next-generation sequencing technologies. First NGS approach, pyrosequencing, was followed by terminator chemistry-based sequencing and ligation-based sequencing methodologies. Before long, third-generation sequencing systems depending on newfangled perspectives were introduced. As third-generation sequencing methods, single molecule fluorescent sequencing, single molecule real-time sequencing, and semiconductor sequencing have provided new opportunities to the users. Also, new technologies like nanopore sequencing present opportunities to perform analyses on DNA, RNA, or proteins. Then, the fourth-generation sequencing has become known as a new and very specific application field which depends on in situ sequencing in fixed cells and tissues and is expected to provide major contributions in key areas. A summary of sequencing generations was given in Table 1.

Table 1 Characteristics of NGS technologies

Generation	Platform	Maximum read length (bp)	Technology	Accuracy (%)	Maximum output per run	Advantages	Disadvantages	Main applications
SGS	FLX 454+	700	Sequencing by synthesis	99.9	700 Mb	Long read length	Homopolymer errors High costs	Resequencing RNA seq
SGS	GS Junior	400	Sequencing by synthesis	99	35 Mb	Long read length	Homopolymer errors High costs	Resequencing RNA seq
SGS	MiSeq	2×300	Sequencing by reversible termination	99.9	15 Gb	Long read length High accuracy Low cost	GC bias Low output	Resequencing RNA seq Small RNA Seq
SGS	NextSeq 500	2×250	Sequencing by reversible termination	99.9	120 Gb	High accuracy Low cost High throughput	GC Bias	Exome seq RNA seq Resequencing
SGS	HiSeq	2×125	Sequencing by reversible termination	99.9	1000 Gb	High throughput High accuracy	GC bias Short reads	Whole genome seq Exome seq RNA seq
SGS	SOLiD	2×50	Sequencing by ligation	99.9	320 Gb	High accuracy	Short reads	Resequencing
TGS	HeliScope	35	Single molecule flourescent sequencing	97	35 Gb	No amplification bias	High error rate Short reads	RNA seq DNA seq
TGS	PacBio RS II	10,000	Single molecule real-time sequencing	95	400 Mb	No amplification bias	High error rate	De novo seq

(continued)

Table 1 (continued)

Generation	Platform	Maximum read length (bp)	Technology	Accuracy (%)	Maximum output per run	Advantages	Disadvantages	Main applications
TGS	Ion PGM	400	Semiconducter sequencing	99	2 Gb	Short run time	Homopolymer errors	Resequencing RNA seq
TGS	Ion Proton	200	Semiconducter sequencing	99	10 Gb	Short run time	Homopolymer errors	Exome seq RNA seq Resequencing
TGS	MinION	6000	Nanopore sequencing	98	500 Mb	No amplification bias Long read length	High error rate	DNA seq RNA seq
FGS	SOLiD	2×50	In situ sequencing	99.9	320 Gb	High accuracy	Short reads	RNA seq

SGS second-generation sequencing, *TGS* third-generation sequencing, *FGS* fourth-generation sequencing, *bp* base pair, *Mb* megabase, *Gb* gigabase

Next-generation systems, in RNA- and DNA-sequencing fields, have been evolving to meet almost every need in genomics field. Whole genome, exome, transcriptome, methylome, metagenomics, ChIP, small RNA, de novo, resequencing applications have started to be used extensively in life sciences and outcomes have changed a variety of concepts in both research and clinic. Even though currently second-generation sequencing technologies respond most of the needs in genomics field, third-, and fourth-generation platforms carry a potential for accurate and practical solutions with a broader application spectrum.

Today, DNA-sequencing field is very close to the gold standards stated by NHGRI 10 years ago but there are still important issues waiting to be solved. $1000 threshold for genome sequencing is about to be overcome and it has become easier to obtain HT with current NGS systems in much shorter time (in hours). Also, it can be said that long reading problem of sequencing technologies will reach a better solution in near future with technologies like nanopore and single molecule real-time sequencing methodologies. On the other hand, accuracy problems stand as the most important issue for all newly developed technologies and a revolutionary advancement is required to make a significant change in this regard also.

DNA sequencing has been developing at a rattling rate and promising major innovations in terms of both research and routine applications. Besides the important convenience brought by new sequencing technologies for application, storage, and bioinformatics of a wide range of fields, platform specific—currently unsolved—issues stand in front of us. Considering the pace of development for sequencing technologies so far, it can be said that new platforms, generations, and solutions, will appear much faster than before in the near future as alternatives for solution of current obstacles.

References

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28(20):e87

Bahassi EM, Stambrook PJ (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis 29(5):303–310

Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ (2011) Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56:406–414

Beck S, O'Keeffe T, Coull JM, Köster H (1989) Chemiluminescent detection of DNA: application for DNA sequencing and hybridization. Nucleic Acids Res 17(13):5115–5123

Bennett S (2004) Solexa Ltd. Pharmacogenomics 5:433–438

Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the 1,000 dollars human genome. Pharmacogenomics 6:373–382

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X,

Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Biggin MD, Gibson TJ, Hong GF (1983) Buffer gradient gels and 35S label as an aid to rapid DNA sequence determination. Proc Natl Acad Sci U S A 80(13):3963–3965

Bowers J, Mitchell J, Beer E, Buzby PR, Causey M, Efcavitch JW, Jarosz M, Krzymanska-Olejnik E, Kung L, Lipson D, Lowman GM, Marappan S, McInerney P, Platt A, Roy A, Siddiqi SM, Steinmann K, Thompson JF (2009) Virtual terminator nucleotides for next-generation DNA sequencing. Nat Methods 6(8):593–595

Buermans HPJ, den Dunnen JT (2014) Next generation sequencing technology: advances and applications. Biochim Biophys Acta 1842(10):1932–1941

Chen EY (1994) The efficiency of automated DNA sequencing. In: Adams MD, Fields C, Venter JC (eds) Automated DNA sequencing and analysis. Academic, San Diego, pp 3–9

Chen CY (2014) DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front Microbiol 5:305

Chen F, Dong M, Ge M, Zhu L, Ren L, Liu G, Mu R (2013) The history and advances of reversible terminators used in new generations of sequencing technology. Genomics Proteomics Bioinformatics 11(1):34–40

Crick FHC (1958) On protein synthesis. Symp Soc Exp Biol 7:138–163

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH (2010) Nanopore DNA sequencing with MspA. Proc Natl Acad Sci U S A 107(37): 16060–16065

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138

Fischbein MD, Drndić M (2008) Electron beam nanosculpting of suspended graphene sheets. Appl Phys Lett 93(11):113107

Flanagan JH, Owens CV, Romero SE, Waddell E, Kahn SH, Hammer RP, Soper SA (1998) Near-infrared heavy-atom-modified fluorescent dyes for base-calling in DNA-sequencing applications using temporal discrimination. Anal Chem 70(13):2676–2684

França LT, Carrilho E, Kist TB (2002) A review of DNA sequencing techniques. Q Rev Biophys 35(02):169–200

Gardner AF, Wang J, Wu W, Karouby J, Li H, Stupi BP, Jack WE, Hersh MN, Metzker ML (2012) Rapid incorporation kinetics and improved fidelity of a novel class of 3′-OH unblocked reversible terminators. Nucleic Acids Res 40(15):7404–7415

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz M, McCombie WR (2015) Oxford nanopore sequencing and de novo assembly of a eukaryotic genome. Genome Res 25(11): 1750–1756

Heng JB, Ho C, Kim T, Timp R, Aksimentiev A, Grinkova YV, Sligar S, Schulten K, Timp G (2004) Sizing DNA using a nanometer-diameter pore. Biophys J 87(4):2905–2911

Huang YF, Chen SC, Chiang YS, Chen TH, Chiu KP (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6(Suppl 2):S10

Hui P (2014) Next generation sequencing: chemistry, technology and applications. Top Curr Chem 336:1–18

Hutchison CA (2007) DNA sequencing: bench to bedside and beyond. Nucleic Acids Res 35(18): 6227–6237

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. Nat Methods. doi:10.1038/nmeth.3290, Epub ahead of print

Kan CW, Fredlake CP, Doherty EA, Barron AE (2004) DNA sequencing and genotyping in miniaturized electrophoresis systems. Electrophoresis 25:3564–3588

Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. Proc Natl Acad Sci U S A 93:13770–13773

Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, Nilsson M (2013) In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods 10:857–860

Kircher M, Kelso J (2010) High-throughput DNA sequencing-concepts and limitations. Bioessays 32(6):524–536

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Grey JJ, Holden D, Saxena R, Wegener J, Turner SW (2010) Real-time DNA sequencing from single polymerase molecules. Methods Enzymol 472:431–455

Kumar S, Sood A, Wegener J, Finn PJ, Nampalli S, Nelson JR, Sekher A, Mitsis P, Macklin J, Fuller CW (2005) Terminal phosphate labeled nucleotides: synthesis, applications, and linker effect on incorporation by DNA polymerases. Nucleosides Nucleotides Nucleic Acids 24(5–7):401–408

Larkin J, Henley R, Bell DC, Cohen-Karni T, Rosenstein JK, Wanunu M (2013) Slow DNA transport through nanopores in hafnium oxide membranes. ACS Nano 7(11):10121–10128

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SS, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM (2014) Highly multiplexed subcellular RNA sequencing in situ. Science 343:1360–1363

Li Z, Bai X, Ruparel H, Kim S, Turro NJ, Ju J (2003) A photocleavable fluorescent nucleotide for DNA sequencing and analysis. Proc Natl Acad Sci U S A 100(2):414–419

Li W, Bell NA, Hernandez-Ainsa S, Thacker VV, Thackray AM, Bujdoso R, Keyser UF (2013) Single protein molecule detection by glass nanopores. ACS Nano 7:4129–4134

Litosh VA, Wu W, Stupi BP, Wang J, Morris SE, Hersh MN, Metzker ML (2011) Improved nucleotide selectivity and termination of 30-OH unblocked reversible terminators by molecular tuning of 2-nitrobenzyl alkylated HOMedU triphosphates. Nucleic Acids Res 39:e39

Liu S, Lu B, Zhao Q, Li J, Gao T, Chen Y, Zhang Y, Liu Z, Fan Z, Yang F, You L, Yu D (2013a) Boron nitride nanopores: highly sensitive DNA single-molecule detectors. Adv Mater 25: 4549–4554

Liu L, Yang C, Zhao K, Li J, Wu HC (2013b) Ultrashort single-walled carbon nanotubes in a lipid bilayer as a new nanopore sensor. Nat Commun 4:2989

Liu K, Feng J, Kis A, Radenoviç A (2014) Atomically thin molybdenum disulfide nanopores with high sensitivity for DNA translocation. ACS Nano 8:2504–2511

Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12(8):733–735
Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM (1990) High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res 18(15):4417–4421
Mardis ER (2013) Next-generation sequencing platforms. Annu Rev Anal Chem 6:287–303
Metzker ML (2005) Emerging technologies in DNA sequencing. Genome Res 15(12):1767–1776
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46
Mignardi M, Nilsson M (2014) Fourth-generation sequencing in the cell and the clinic. Genome Med 6(4):31
Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, Cocho JA (2013) A glimpse into past, present, and future DNA sequencing. Mol Genet Metab 110(1–2):3–24
Müller R, Herten DP, Lieberwirth U, Neumann M, Sauer M, Schulz A, Siebert S, Drexhage KH, Wolfrum J (1997) Efficient DNA sequencing with a pulsed semiconductor laser and a new fluorescent dye set. Chem Phys Lett 279(5):282–288
Mullis KB, Faloona FA, Scharf SJ, Saiki RK, Horn GT, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51:263–273
Novais RC, Thorstenson YR (2011) The evolution of Pyrosequencing® for microbiology: from genes to genomes. J Microbiol Methods 86(1):1–7
Nyrén P (2007) The history of pyrosequencing. Methods Mol Biol 373:1–14
Paegel BM, Blazej RG, Mathies RA (2003) Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis. Curr Opin Biotechnol 14(1):42–50
Porreca GJ, Shendure J, Church GM (2006) Polony DNA sequencing. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds) Current protocols in molecular biology. Greene and John Wiley, New York, pp 1–22, Unit 7.8
Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza JA, Jensen MA, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238(4825):336–341
Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. Nat Biotechnol 27:847–850
Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13(1):341
Roberts RJ, Carneiro MO, Schatz MC (2013) The advantages of SMRT sequencing. Genome Biol 14:405
Ronaghi M (2000) Improved performance of pyrosequencing using single stranded DNA-binding protein. Anal Biochem 286:282–288
Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 242(1):84–89
Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. Science 281(5375):363–365
Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352
Rusk N (2014) Genomics: nanopores read long genomic DNA. Nat Methods 11(9):887
Sanger F (1988) Sequences, sequences, and sequences. Ann Rev Biochem 57:1–28
Sanger F, Coulson A (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94(3):441–448

Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem J 49(4):463–481
Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467
Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hom Mol Genet 19(R2):R227–R240
Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145
Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741):1728–1732
Siwy Z, Fuliński A (2002) Fabrication of a synthetic nanopore ion pump. Phys Rev Lett 89:198103
Smith L, Sanders J, Kaiser R, Hughes P (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321(6071):674–679
Song CX, Clark TA, Lu XY, Kislyuk A, Dai Q, Turner SW, He C, Korlach J (2012) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat Methods 9(1):75–77
Soon WW, Hariharen M, Snyder MP (2013) High-throughput sequencing for biology and medicine. Mol Syst Biol 9(640):1–14
Srinivasan S, Batra J (2014) Four generations of sequencing—is it ready for the clinic yet? Next Generat Sequenc Applic 1:107
Stranneheim H, Lundeberg J (2012) Stepping stones in DNA sequencing. Biotechnol J 7(9): 1063–1073
Thompson JF, Steinmann KE (2010) Single molecule sequencing with a HeliScope genetic analysis system. Curr Protoc Mol Biol. Chapter 7, Unit 7.10
Wang Y, Yang Q, Wang Z (2014) The evolution of nanopore sequencing. Frontiers Genet 5:449
Watson JD, Crick F (1953) A structure for deoxyribonucleic acid. Nature 171:737–738
Watts D, MacBeath J (2001) Automated fluorescent DNA sequencing on the ABI PRISM 310 genetic analyzer. Meth Mol Biol 167:153–170
Wetterstrand KA (2014) DNA sequencing costs: data from the NHGRI genome sequencing program (GSP). www.genome.gov/sequencingcosts. Accessed Dec 2014
Wiemann S, Schilke A, Rechmann S, Zimmermann J, Voss H, Ansorge W (1996) Reducing "double sequences" in automated DNA sequencing with T7 DNA polymerase and internal labeling. Biotechniques 20(5):791–792
Wu R, Kaiser AD (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 35(3):523–537
Wu W, Stupi BP, Litosh VA, Mansouri D, Farley D, Morris S, Mtezker S, Metzker ML (2007) Termination of DNA synthesis by N6-alkylated, not 3′-O-alkylated, photocleavable 2′-deoxyadenosine triphosphates. Nucleic Acids Res 35:6339–6349

Molecular Markers and Their Applications

Elif Karlik and Hüseyin Tombuloğlu

Contents

Abstract The development and use of molecular markers for the detection and exploitation of polymorphism have been playing a significant role in plant breeding studies. The earliest type of the DNA marker was RFLP. However, the development of PCR allowed designing other approaches like RAPD, SSR, AFLP, and SNPs. Also, new powerful technologies as microarray and RNA-sequencing have made possible to use RNA as a molecular marker. The presence of various types of molecular markers provides to combine some desirable properties. Selecting one or more of such techniques that are based on different principles, methodologies, and application should be considered carefully. Molecular markers can be a useful tool for biotechnological developments.

Keywords Molecular markers • Markers • Cereals

E. Karlik (✉)
Institute of Sciences and Engineering, Fatih University, Buyukcekmece, Istanbul 34500, Turkey
e-mail: eliffkarlik@gmail.com

H. Tombuloğlu
Department of Biology, Fatih University, Buyukcekmece, Istanbul 34500, Turkey

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_6

1 Introduction

Since the late nineteenth century plant breeders use phenotypic traits such as plant habits, disease resistances, yield, or quality to develop new cultivars. Plant breeding can be divided into two major strategies as classical breeding and molecular breeding. Closely related individuals are used to produce new varieties with desirable characteristics in classical plant breeding. Long life cycles and several generations are need for selection and evaluation of useful genotypes (Tester and Langridge 2010). On the other hand, molecular plant breeding uses two major approaches, marker-assisted selection (MAS) and genetic transformation, to improve new cultivars (Moose and Mumm 2008; Rajib et al. 2013). MAS is a process that utilizes molecular markers rather than the phenotype of the trait to select desirable trait of plants to enhance crop yield, quality, and tolerance to biotic or abiotic stresses. DNA markers are defined as sections of the genome containing mutations/variations. These variations are used to recognize polymorphism between different genotypes or alleles of a gene in a populations or gene pool. DNA markers can be associated with particular DNA fragments within the gene of interest or be linked to a trait of interest (Foolad and Sharma 2005).

Firstly, MAS has been started with selection of visible traits, such as leaf shape, flower color, pubescence color, pod color, seed color, seed shape, hilum color, awn type and length, fruit shape, rind (exocarp) color and stripe, flesh color, stem length in 1932 (Sax), and by the early 1980s it has been proceeded with usage of allozyme as a marker (Tanksley 1993). However, workings with phenotypic markers impose restriction to determine desirable characteristics. The development of DNA markers has been changed the situation, because more polymorphisms can be revealed at the DNA level (Ruane and Sonnino 2007). The first DNA-based genetic marker, used in human linkage mapping, was restriction fragment length polymorphisms (RFLPs) (Botstein et al. 1980). Improvement of molecular techniques helps to find new markers associated with interest traits on large scale, and DNA markers have been applied in plant genetics and breeding, germplasm characterization, and management. With further advance of biotechnology, new DNA markers have been applied such as random amplification of polymorphic DNA (RAPD; Williams et al. 1990), sequence characterized amplified region (SCAR; Paran and Michelmore 1993), cleaved amplified polymorphic sequences (CAPS; Konieczny and Ausubel 1993), simple sequence repeats (SSRs; Litt and Luty 1986; Salimath et al. 1995), amplified fragment length polymorphisms (AFLPs; Vos et al. 1995), direct amplification of length polymorphisms (DALP; Desmarais et al. 1998), and single nucleotide polymorphism (SNP) (Gupta et al. 2001). Also, it is possible to make a discrimination based on similarity of SNPs with some of these marker systems which are grouped into SNPs (due to sequence variation, e.g., RFLP) and non-SNPs (due to length variation, e.g., SSR) (Gupta et al. 2001). Ideal DNA markers should have some properties: (1) high level of polymorphism, (2) even distribution across the whole genome (not clustered in certain regions), (3) codominance in expression (so that heterozygotes can be distinguished from homozygotes), (4) clear distinct allelic features (so that the different alleles can be easily identified), (5) single copy and no

pleiotropic effect, (6) low cost to use (or cost-efficient marker development and genotyping), (7) easy assay/detection and automation, (8) high availability (unrestricted use) and suitability to be duplicated/multiplexed (so that the data can be accumulated and shared between laboratories), (9) genome specific in nature (especially with polyploids), and (10) no detrimental effect on phenotype.

Based on ideal markers definition, RFLP, SSR, AFLP, and SNPs are most effective marker systems for determination of polymorphism. Application of these markers on plant breeding has been so much various such as assessment of genetic variability and characterization of germplasm, identification and fingerprinting of genotypes, estimation of genetic distances between population, inbreeds and breeding material, detection of monogenic and qualitative trait loci (QTL), MAS, identification of sequences of useful candidate genes, etc. To detect the polymorphism, two basic methods have been developed; Southern blotting, a nuclear acid hybridization technique (Southern 1975), and PCR, a polymerase chain reaction technique (Mullis 1990). These techniques have been followed by PAGE, polyacrylamide gel electrophoresis; AGE, agarose gel electrophoresis; and CE, capillary electrophoresis based on the product features, such as band size and mobility. For a decade, several new methods have been applied such as array chip and sequencing technologies to determine polymorphism.

The features of the widely used marker techniques discussed below are compared in Table 1. The choice and use of one or combination of marker methods in research and breeding is still a challenge to improve cereals for plant breeders. In this chapter, an overview over the details about the technical methods for the development of several types of molecular markers are routinely being used and also new developed in plant breeding. The major reports related to "classical" and new "high-throughput" molecular markers are given to briefly introduce the methodologies of marker systems.

2 Classical Markers

2.1 RFLP Markers

RFLP markers, the first marker system, have been used for plant genome mapping. RFLPs are detected by cutting genomic DNA with restriction enzymes which have a specific recognition sequences. The restriction fragments of certain length are separated based on size with agarose gel electrophoresis and analyzed by Southern Blotting. RFLP markers are also known as Southern-Blotting-based markers. Mutation events (point mutations, insertions, or deletions) within the restriction enzymes recognition sequences result in differences in size of length and molecular weights of restriction fragments. Most RFLPs are codominant, locus specific, high reproducibility, and have simple methodology which makes them powerful tools for comparative and synteny mapping. Another advantage of RFLPs is that no prior information of the sequence is needed. The requirement of relatively large amounts

Table 1 Comparison of most widely used marker systems in plant breeding; adapted from Semagn et al. (2006a), Xu (2010), and others

Feature and description	RFLP	RAPD	AFLP	SSR	SNP	DArT	RAD	TAM	KASPar	CNVs and PAVs
Genomic abundance	High	High	High	Moderate to high	Very high	High	High	Moderate to high	Very high	Moderate to high
Genomic coverage	Low copy coding region	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome	Whole genome
Expression/ inheritance	Codominant	Dominant	Dominant/ codominant	Codominant	Codominant	Dominant/ codominant	Codominant	Codominant	Dominant/ codominant	Dominant/ codominant
Sequence information	Not required	Not required	Not required	Required	Required	Not required	Not required	Required	Required	Required
Number of loci	Small (<1000)	Small (<1000)	Moderate (1000 s)	High (1000–10,000 s)	Very high (>100,000)	Very high (>100,000)	Very high (>100,000)	Very high (>100,000)	Very high (>100,000)	Very high (>100,000)
Level of polymorphism	Moderate	High	High	High	High	High	High	High	High	High
Type of polymorphism	Single base changes, indels	Single base changes, indels	Single base changes, indels	Changes in length of repeats	Single base changes, indels	Single base changes, indels	Single base changes, indels	Single base changes, indels	Single base changes, indels	Single base changes, indels
Type of probes/ primers	Low copy DNA or cDNA clones	10 bp random nucleotides	Specific sequence	Specific sequence	Allele-specific PCR primers	Specific sequence and adapters	Specific sequence and adapters	Allele-specific primers	Allele-specific primers and locus-specific primers	Specific sequence
Cloning and/or sequencing	Yes	No	No	Yes	Yes	Yes	Yes	No	No	Yes
PCR requiriment	Usually no	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reproducibility/ reliability	High	Low	High	High	High	High	High	High	High	High
Effective multiplex ratio	Low	Moderate	High	High	High	High	High	High	High	Moderate

Marker index	Low	Moderate	Moderate to high	High	High	High	High	High	High	High
Genotyping throughput	Low	Low	High	High	High	High	High	High	High	High
Amount of DNA required	Large (5–50 µg)	Small (0.01–0.1 µg)	Moderate (0.5–1.0 µg)	Small (0.05–0.12 µg)	Small (≥0.05 µg)	Small (0.05–0.1 µg)	Moderate (0.1–1.0 µg)	Small (0.05–0.1 µg)	Small (0.02–0.05 µg)	Small to large (0.02–5.0 µg)
Quality of DNA required	High	Moderate	High	Moderate to high	High	High	High	High	High	High
Technically demanding	Moderate	Low	Moderate	Low	High	High	High	High	High	High
Time demanding	High	Low	Moderate	Low	Low	Low	Low	Low	Low	Low
Ease of automation	Low	Moderate	Moderate to high	High	High	High	High	High	High	High
Development/ start-up cost	Moderate to high	Low	Moderate	Low	High	High	High	High	High	High
Cost per analysis	High	Low	Moderate	Low	Low	Low	Low	Low	Low	Low to high
Number of polymorphic loci per analysis	1.0–3.0	1.5–5.0	20–100	0.61–0.70	0.41–0.50	0.21–0.30	1.0	1.0	1.0	0–10.0
Primary application	Genetics	Diversity	Diversity and genetics	All purposes	All purposes	All purposes	All purposes	All purposes	All purposes	All purposes

of pure and intact DNA and the tedious experimental procedure may be counted as disadvantages of RFLPs. Also, it is very difficult to automate which limits its use and share between laboratories (Botstein et al. 1980; Weising et al. 2005; Edwards and Mccouch 2007). In 1990s, RFLP technique has been improved that this method is called cleaved amplified polymorphism sequence (CAPS), also known as PCR-RFLP. It is high-throughput markers that have been developed by using RFLP probe sequences. To detect the polymorphism by CAPS, PCR-amplified fragments are digested with restriction enzymes. The presence/absence of restriction sites in amplified fragments determines the polymorphism (Konieczny and Ausubel 1993). The last few decades, use of RFLP markers has been increased and less research been reported in genetic research and plant breeding.

2.2 RAPD Markers

RAPD markers are a PCR-based marker system based on amplification of random DNA segments with single, short (usually about ten nucleotides), arbitrary primers (Williams et al. 1990). The primer binds to many different loci based on complementary to DNA template (maybe consisting of a limited number of mismatches) and where two primers bind to DNA in close proximity for amplifying DNA sequence, PCR is successfully concluded. The PCR products depend on the primer length and size and the target genome. The amplified fragments are visualized by agarose gel electrophoresis. The primers, are nonspecies specific and can be universal, are arbitrarily chosen because prior knowledge of the DNA sequence is not to be required. Mutations or rearrangements either at or between the primer-binding sites are resulted as polymorphism which is determined by the presence or absence of a particular RAPD band with a certain molecular weight, with no information on heterozygosity in the electrophoresis. Besides providing dominance and high level of polymorphism, RAPD markers show some difficulty of reproducibility of data. However, RAPDs is technically simple, quick, efficient, and easy to be conducted in laboratories. Also, the procedure does not involve any blotting or hybridization steps. Small amounts of DNA (about 10 ng per reaction) are needed compared with RFLP and the technique can be automated. Concerned RAPD markers can be cloned and sequenced, then used to apply other types PCR-based markers such as SNP, etc. (Weising et al. 2005; Edwards and Mccouch 2007; Khan et al. 2013).

2.3 AFLP Markers

AFLPs are PCR-based markers and combines elements of RFLP and RAPD. The technique consisting three steps is based on the selective restriction fragments amplification by PCR. Firstly, genomic DNA (about 500 ng) is double digested with a rare cutter (6-bp recognition site, EcoRI, PtsI or HindIII) and a frequent cutter (4-bp

recognition site, MseI or TaqI), and then short oligonucleotide adapters are ligated to both ends of restriction fragments to provide known sequences for amplification. Secondly, the selective fragments are amplified with the primers (17–21 nucleotides in length) which are the combination of restriction sites and an arbitrary, nondegenerate "selective" sequence (1–3 nucleotides). The 3′ ends of the selective primers whose ends have a small number of nucleotides adjacent to the restriction sites provide to anneal perfectly to their target sequences. In the last step, PCR products are separated on electrophoresis and can be visualized by autoradiography, silver staining, or fluorescence. AFLPs are very reliable, robust, and produce a high marker density compared with RFLPs. It is also can tolerate small variations in PCR such as thermal cycles and template concentrations. Relatively small DNA (1–100 ng per individual) is required for AFLP. Sequence variations in a restriction sites, insertions, or deletions within products and differences of selective nucleotides in the primers can cause polymorphism for AFLP. Fifty to hundred amplified PCR products can be obtained in a typical AFLP procedure and up to 80 % of amplified fragments may serve as markers. AFLPs also have high multiplex ratio and genotyping throughput and better reproducibility than RAPDs. However, it is also does not require sequence information or any prior probe collection before applying AFLP procedure like RAPDs. This means that a set of primers can be used to detect polymorphism for different species. Like the other marker systems, AFLPs also have some limitations. One of the limitations is AFLPs require greater technical skill, larger investments in equipment, and high quality DNA is needed for digestion of restriction enzymes. For instance, biallelic markers is low (the maximum is 0.5) that the most AFLP markers are dominant rather than codominant. On the other hand, specialized algorithms and software packages have been developed to find such codominant markers. High quality DNA is needed for digestion of restriction enzymes. Some AFLPs clusters intensively are located in centromeric regions in some species whose have large and complex genomes (e.g., barley and sunflower). In addition, development of AFLP markers is complicated and required high costs. AFLP technique is applied for analysis of germplasm collections, genotyping of individuals, construction of genetic DNA marker maps, construction of physical maps, gene mapping, and transcript profiling (Vos et al. 1995; Mueller and Wolfenbarger 1999; Weising et al. 2005; Edwards and Mccouch 2007; Meudt and Clark 2007).

2.4 Microsatellite Markers

Microsatellites are also known as simple sequence repeats (SSRs) or short tandem repeats (STRs) or simple sequence length polymorphisms (SSLPs) and are the smallest class of simple repetitive DNA- and PCR-based markers. SSRs are consisting of tandem repeated short nucleotide motifs (2–6 bp/nucleotides long) which can be di-, tri-, and tetranucleotide repeats, e.g., (GT)n, (AAT)n, and (GATA)n, and widely distributed throughout the genomes of plants. The differences in the numbers of repeated units vary among individuals and cause polymorphism in plants.

DNA sequences flanking the repeats are used to design the locus specific primers because these sequences are usually conserved. The amplification products are separated in high-resolution electrophoresis systems (e.g., AGE and PAGE) and visualized by fluorescent labeling or silver staining. The amplification is resulted with different band sizes which are hypervariable, codominant, reproducible, and locus specific. In addition, this method is capable of revealing the differences of allele size seven of a single nucleotide pair. Because of being codominant, makes SSRs an excellent marker system to study population genetics, mapping, MAS, genotyping of individuals, and germplasm analysis for plants. SSRs can easily be automated and shared primer sequences between laboratories. For SSR assays, only small DNA samples (~100 ng per individual) and low start-up costs are needed. However, there are some challenges to study with SSR markers. Some of the challenges are the technique requires sequence information to design primers, labor intensive, and high costs for automazation (Powell et al. 1996; Jarne and Lagoda 1996; Goldstein and Schlötterer 1999; Nybom 2004; Weising et al. 2005). Since the 1990s, SSR marker developments have been continued and over 35,000 SSR markers have been applied to use on plant breeders (Song et al. 2010).

2.5 SNP Markers

SNPs are single nucleotide substitutions in a DNA sequence with a usual alternative of two possible nucleotides at a given position (Vignal et al. 2002; Ganal et al. 2009). The nucleotide substitutions can be either as transitions (C/T or G/A) or transversions (C/G, A/T, C/A, or T/G). A single base insertions or deletions (indels) in the genome are also considered as an SNP. The most important advantage of the SNPs in genomes is numerous; therefore, SNPs provide the ultimate form of markers as the smallest unit of the DNA and have a great marker density. SNPs are very common in plants genomes that have high frequencies with the range of one SNP every 100–300 bp (Edwards et al. 2007; Xu 2010). SNPs can occur within coding or noncoding sequences of genes or in the intergenic regions in different chromosome regions. SNP marker technique is not a gel based that provides an advantage to analysis of the large-scale genotyping. Several methods of allelic discrimination and detection platforms have been developed by plant breeding laboratories. Some of these methods are allele-specific hybridization, primer extension, oligonucleotide ligation, and invasive cleavage based on the molecular mechanisms (Sobrino et al. 2005). Different analysis methods are also available to detect the products such as electrophoresis, mass spectrophotometry, chromatography, fluorescence polarization, arrays or chips, etc. In principle, the SNP techniques reveal to identify differences between a probe of known sequence and a target DNA containing the SNP site. The target DNA segments are mostly PCR products and mismatches with the probe reveal SNPs within the amplified target DNA section. The mismatching DNA segments can be analyzed with sequencing then as the most direct way to determine SNP polymorphisms (Gupta et al. 2001; Rafalski 2002; Weising et al. 2005). The use of SNPs as a marker

has increased their potential value in studies of genetic variation, construction of genetic maps, population structure analysis, association genetics, map-based gene isolation, and other plant breeding applications including QTL mapping, germplasm characterization, and molecular breeding (Kumar et al. 2012; Abdel-Haleem et al. 2013). SNP discovery and SNP validation are the most important two parts of development of SNP markers. For this reason, over the past few decades a great deal of effort has been devoted to developing accurate, rapid, and cost-efficient technologies for SNP genotyping platforms including Invader® assay, single base extension (SBE), oligonucleotide ligation assay (OLA) SNPlex™ system, and the Illumina GoldenGate™ and Infinium™ assays (Appleby et al. 2009) have been improved and widely used to genotype crop plants with a fixed set of SNP markers (Close et al. 2009; Yang et al. 2012; Deulvot et al. 2010; McCouch et al. 2010; Hyten et al. 2010; Bachlava et al. 2012; Chao et al. 2010). The higher levels of genome complexity and the lack of reference genome sequences for some crops have been made difficult to discover SNPs in crops. The rapid genome-wide identification of a large number of SNPs at a much lower price tag have been provided by the next generation sequencing (NGS) technologies such as 454 Life Sciences (Roche Applied Science, Indianapolis, IN), Hiseq (Illumina, San Diego, CA), SOLiD and Ion Torrent (Life Technologies Corporation, Carlsbad, CA) (Mardis 2008; Kumar et al. 2012).

3 High-Throughput Markers

The choice of marker systems in research and breeding is still challenging part of plant breeding. Therefore, numerous factors need to be considered when one or more molecular marker types are chosen by breeders (Semagn et al. 2006a). The choice and used marker techniques has shifted from the first and second generations DNA markers including RFLPs, RAPDs, microsatellite, and AFLPs to the third and fourth generation markers such as DArTs, TAMs, RADs, and CNVs/PAVs (Gupta et al. 2008; Potokina et al. 2008; Springer et al. 2009; Belo et al. 2010). A breeder should make an appropriate choice that best suits the requirements according to the conditions and resources. This chapter presents information about newly developed high-developed genotyping platforms and examines their principles of genotype determination of plants.

3.1 Diversity Array Technology (DArT)

Crop improvement is based on the productive use of molecular marker technologies. Several kind of molecular markers have been applied over the last 25 years. Diversity array technologies (DArT) is one of the molecular markers is a high-throughput microarray hybridization-based technique. DArT are relatively new and has been developed in early 2000 that enables simultaneous typing of several

hundred polymorphic loci spread over the entire genome without any previous sequence information about these loci (Jaccoud et al. 2001; Wenzl et al. 2004). DArT is reproducible, high throughput and cost effective genome-wide technique to assay thousands of presence/absence polymorphisms in a single assay. Fifty to hundred nanograms of genomic DNA is required for genotyping almost 5000–8000 genomic loci simultaneously in a single-reaction. The methylation sensitive restriction enzymes are used to diget genomic DNA that allows reducing the genome complexity and enriching the low copy sequences. No specific assay for genotyping is required to be developed because of the same platform is used for both discovery and scoring of markers (Huttner et al. 2005). DArT involves preparation of genomic representations for each individual DNA sample which is digested with restriction enzymes (e.g., PstI and TaqI) and followed by ligation of restriction fragments to adapters. Complexity of the genome is reduced by amplifying fragments by using proper primers that are complementary to the adapters and selective overhangs. PCR fragments are cloned and then inserts are also amplified by using vector-specific primers and purified. After purification of interest fragments are arrayed onto a solid support (microarray chips) resulting in a "discovery array" (www.diversityarrays.com) (Jaccoud et al. 2001). The "discovery array" is developed from labeled genomic representations included in the pool which is also called metagenome which is a pool of genomes representing the diverse germplasm of interest is used to decrease the level of repetitive DNA (Killian et al. 2005). Molecular basis of DArT markers is based on single base-pair changes (SNPs) and also InDels/rearrangements within the restriction sites (Jaccoud et al. 2001). Polymorphic DArT markers are obtained from different samples show variable hybridization signal intensities for different individuals. These clones are subsequently assembled into a "genotyping array" that only polymorphic markers are routinely used for genotyping. These DArT markers are biallelic and dominant (presence vs. absence) or codominant (two doses vs. one dose vs. absent), and were successfully used in crop improvement such as rice, barley, wheat and maize (Huttner et al. 2005). DArT technique has numerous advantages that it is no need prior sequence information, high throughput, quick, and highly reproducible. The method is also cost effective and estimated cost is tenfold lower than SSR markers (Xia et al. 2005). The analysis of genetic content is indicated by the user and easily expandable. It is designed for open-sources and shared improvements in this way the users do not deal with patent rights. However, this technique has also limitations that DArT is an array-based method that involves several steps such as preparation of representation for the target species, cloning, data management and analysis. The analysis of DArT markers requires special software's including DArTsoft and DArTdb. DArT markers are primarily dominant (present or absent) or differences in intensity which causes some limitations in some applications (Semagn et al. 2006b). DArT markers have been already been applied for several number of plant species, including important and also orphan crops. However, there was no prior sequence information for orphan crops that the technique has successfully been used to develop DArT markers (Huttner et al. 2006). Jaccoud and friends (2001) first described the technique by using rice which is one of the major cereal crops.

First DArT markers have been developed by using rice (Jaccoud et al. 2001) and have also been used in large number of plant species such as *Arabidopsis thaliana*, sorghum, triticale, barley, and wheat (Wittenberg et al. 2005; Wenzl et al. 2006; Mace et al. 2009; Alheit et al. 2011; Alsop et al. 2011; Trebbi et al. 2011; Marone et al. 2012). However, DArT markers are suitable for genetic studies in nonmodel organisms and wild species for which no molecular information is available (James et al. 2008; Alsop et al. 2011). Discovery of new genes are possible with MAS which are provided by molecular markers that are linked to genes. A number of linkage maps have been produced by using DArT markers for wheat and durum wheat (Zhang et al. 2008; Peleg et al. 2008; Mantovani et al. 2008; Francki et al. 2009; Gadaleta et al. 2009; Blanco et al. 2011, 2012). Although, the integration of data derived from individual datasets in consensus maps have carried out in many species, such as sorghum, triticale, barley, and durum wheat (Wenzl et al. 2006; Mace et al. 2009; Alheit et al. 2011; Alsop et al. 2011; Trebbi et al. 2011; Marone et al. 2012).

3.2 Tagged Array Markers (TAMs)

The Tagged Array Markers (TAMs) were firstly developed to score retrotransposon-based insertion polymorphism (RBIP) markers, but is well-suited to score SNPs and insertion-deletion (indel) alleles at genomic loci in hundreds of individuals or multiple individuals (Flavell et al. 2003). In TAMs method, the initial step is PCR which is conducted by using one biotin labeled primer shared by both alleles and two unique primers carrying allele-specific oligonucleotide tags. In the second step, the Biotin-labeled target sequences in the form of PCR products are spotted directly onto streptavidin-coated glass microarray slides. The target sequences (allele-specific PCR products carrying unique tags) are differentiated by hybridization to fluorescent-labeled detector probes to identify alleles represented by each of the targeted sequences that are arrayed. The main attractions of this method are its high throughput that thousands of PCRs are analyzed per slide, making it useful for screening large population for new markers. Also TAMs provide flexibility of scoring (any combination, from a single marker in thousands of samples to thousands of markers in a single sample, can be analyzed) and flexibility of scale (any experimental scale, from a small lab setting up to a large project) (Flavell et al. 2003; Gupta et al. 2013).

3.3 Restriction-Site Associated DNA (RAD) Markers

Restriction-site associated DNA (RAD) marker system has been developed for study of genome-wide SNP variations associated with restriction recognition sites for individual restriction enzymes. More recently, a variety of microarrays (including tiling/cDNA/oligonucleotide arrays) and sequencing have also been used to

develop the RAD markers (Baird et al. 2008; Yang et al. 2012; Gupta et al. 2013; Pegadaraju et al. 2013; Matsumura et al. 2014).

For microarray assay, a genome-wide library of RAD tags is developed from genomic DNA. Then, this RAD tags have been used for hybridization on to the chosen microarray to detect all restriction-site SNP variations in a single assay. The first step initiates with digestion of genomic DNA with a specific restriction enzyme. The next steps involve ligation of biotinylated linkers to the digested DNA and random shearing of ligated DNA into fragments smaller than the average distance between restriction sites, leaving small fragments with restriction sites attached to the biotinylated linkers. After immobilization of these fragments on streptavidin-coated beads, DNA tags are released from the beads by digestion at the original restriction sites. In this way, DNA tags flanking the restriction sites are specifically isolated throughout the genome. The RAD tags from each of a number of samples provide high-throughput identification and/or typing of differential hybridization patterns when hybridized on to a microarray (Gupta et al. 2008).

The RAD-sequencing (RAD-Seq) combines two simple methods with sequencing; digestion of DNA with a particular restriction enzymes and the use of molecular identifiers (MID) to associate sequence reads to particular individuals. RAD-seq marker system is based on sequencing short fragments from MIDs, flanking restriction recognition sites in the genome, and counting their frequency (Baird et al. 2008; Davey and Blaxter 2010). Polymorphisms among cultivars or segregating individuals are defined by the presence or absence of these short sequences (tags) or SNPs and indels in the sequence flanking the restriction site that will be uniquely identify the individual. For this purpose, DNA firstly is cut with the chosen restriction enzymes to produce sticky-ended fragments. Then for sequencing, these restriction digestion products must be ligated to adapters which enable the binding and amplification of restriction site fragments only. The adapters contain a matching sticky-end and a MID. In this processes, individuals or cultivars can also be pooled before restriction digestion and first ligation step. The shared fragments are ligated to a second adapter and PCR amplification is conducted by using primers which are complementary to first and second adapters. Second adapter's structure is "Y" that is not able to bind to the second unless it has been completed by amplification by the first adapter. It provides that all PCR products have the first and second adapters, MID, the partial restriction site and a few hundred bases of flanking sequence. Approximately 200–500 base pairs fragments are selected and this RAD-seq library is sequenced. Sequencing processes is generated from the MID in the first adapter and across the restriction enzyme site, generating a data set of RAD tags (sequences downstream of restriction sites) that derive from a much-reduced part of the original genome. Two RAD tags will be produced from each site when the restriction site is symmetric. For Illumina platform, approximately 150–300 bases flanking each restriction recognition sites can be screened for polymorphisms (Baird et al. 2008; Davey and Blaxter 2010).

The RAD marker system has clear advantage over RFLPs, AFLPs and DArT markers that could assay only SNPs which disrupt restriction recognition sites. Polymorphisms among cultivars or segregating individuals are based on the presence

or absence of tags. However, the full sequences of the RAD tag and its paired contig can be screened for SNPs and indels (Davey and Blaxter 2010). RAD-seq using next-generation sequencing technology is suitable to perform for genetically closely related plant resources. Furthermore, alleles of tags linked to agronomically important traits are applicable to codominant DNA markers for plant breeding including genetic mapping and QTL analysis (Rowe et al. 2011). RAD marker system were successfully applied in a number of organisms including fruit fly, zebrafish, threespine stickleback, Neurospora (Lewis et al. 2007; Miller et al. 2007a, b), and also developed for construction of high-density genetic map in barley and wheat (Chutimanitsakun et al. 2011; Poland et al. 2012; Zhou et al. 2015).

3.4 Competitive Allele Specific PCR (KASPar) Assays

GoldenGate (GG) and/or Infinium assays are widely utilized for genotyping of a large number of SNP markers in all major crop species. GoldenGate assays are not cost-effective for genotyping a population for few SNPs (Chen et al. 2010). However, KASPar (KBioScience AlleleSpecific Polymorphism, KBioscience, UK) system provides a promising alternative method. The KASPar method involves competitive allele-specific PCR which is a modification of TAM technology, followed by SNP detection via Fluorescence Resonance Energy Transfer (FRET) (McCouch et al. 2010). The assay involves real-time detection of the products instead of a hybridization step. Therefore, KASPar is a simple, cost-effective and flexible for determining SNPs and InDels. Also, the system can be performed as 48, 96, 384 and 1536-well plate forms according to needs (Gupta et al. 2013).

Four primers, two allele specific and two locus specific primers, are used in a unique form of allele specific PCR which is a different form of the conventional amplification refractory mutation system (ARMS). The KASPar® assay system is based on the discrimination power of competitive allele specific PCR to detect the alleles at a specific locus. KBioscience improved this method by incorporating (1) a 5′–3′ exonuclease cleaved Taq DNA polymerase (the engineered Taq increases its discrimination power) and (2) a homogeneous Fluorescence Resonance Energy Transfer (FRET) detection system. The two allele-specific primers which are designed based on SNPs that they contain a unique 18 bp tail to respective allele specific products. Then, following cycles allow integration of allele specific fluorescent labels to the PCR products (with the corresponding labeled primers). The presence of dual emission modules of the detection system provides great advantage to read internal standard (ROX) and the allele specific dyes (FAM and VIC) together (Gupta et al. 2013). KASPar chemistry can be applied to small and large-scale projects that can be also combined with the Fluidigm integrated nano-fluidic circuit (IFC) and EP1 endpoint fluorescence reader. The utilization of Fluidigm IFC and EP1 systems reduce a data point cost to $0.05 per data point which is substantially much cheaper than traditional markers systems (e.g., AFLPs or SSRs) (Maughan et al. 2011). If a Fluidigm EP1 endpoint fluorescence reader cannot be used or for

small-scale project, KASPar assays can be read on a standard fluorescence resonance energy transfer (FRET) plate reader.

KASPar system provides a promising alternative for breeders to analyze a small number of targeted SNPs in a large number of samples. Therefore, KASPar genotyping technique can be used for a variety of purposes such as: (1) genetic mapping, (2) genetic diversity studies, (3) detection of SNPs within a subset of germplasm, (4) fine-mapping of QTLs, (5) marker-assisted breeding, and (6) retaining target regions in NIL development (see McCouch et al. 2010). This system has already been utilized for a numerous number of species including cereals like rice, maize, and wheat. In wheat, the assay has been utilized for rapid generation of a linkage map containing several hundred SNPs (Allen et al. 2011). Similarly, in maize, a set of 695 highly polymorphic SNPs among a total of 13,882 GG-validated SNPs were selected and diverted into KASPar genotyping assay with a 98 % success rate (Mammadov et al. 2012).

3.5 CNVs and PAVs as Markers

Copy number variations (CNVs) and presence-absence variations (PAVs) that have been used in human genome analyzing are the latest markers developed recently and now also being used in plants. These markers are generally characterized as structural variations that are detected through the use of microarrays, are then performed for comparative genomic hybridization (CGH). The CGH method was the first efficient approach to scanning the entire genome variations of DNA copy numbers and presence-absence (Kallioniemi et al. 1992; du Manoir et al. 1993). The assay of CGH involves competitive hybridization of two differentially labeled genomic DNA samples (a test and a control as a reference) on to metaphase chromosomes. To detect CNVs, PAVs, and InDels, the ratio of the fluorescent signal density of the labeled test DNA to that of the reference DNA is utilized (Schrider and Hahn. 2010). The CGH assays facilitate identification of CNVs, PAVs, InDels, and other genetic alterations. The resolution of the CGH technique has been improved by designed microarrays and termed as array-comparative genomic hybridization (aCGH). The resolution range has been improved from 5 to 10 Mb (Kirchhoff et al. 1998; Lichter et al. 2000) to 1 kb–3 Mb (Lucito et al. 2003). The spotted probs in aCGH generally represent the genomic regions of interest and then, digital imaging systems capture and quantify the relative fluorescence densities of the labeled DNA which is hybridized to each target. The fluorescence ratio of the test and reference hybridization signals is detected at different positions through the genome. The ratio of the test and reference provides information for CNVs, PAVs, and InDels in the test genome as compared to the reference genome. Roche NimbleGen and Agilent Technologies are currently the major suppliers for whole-genome aCGH platforms (Alkan et al. 2011).

CNVs, PAVs, and InDels are now being increasingly used in cereals such as rice, maize, barley, wheat and are likely to be preferred over other marker systems in future. A high-density oligonucleotide aCGH microarray was utilized in rice to estimate the

number of CNVs between two cultivars, Nipponbare and Guang-lu-ai4. These CNVs involved known genes, and may be linked to variation among rice varieties (Yu et al. 2011). The role of PAVs in definition of plant phenotype has been demonstrated in opium (*Papaver somniferum*), where a 221 kb genomic region containing a cluster of 10 genes were found to be related to synthesis of noscapine (Winzer et al. 2012). Whole-genome aCGH was also used for the analysis of CNVs and PAVs in maize were found to be associated with domestication (Swanson-Wagner et al. 2010; Chia et al. 2012). The results showed that modern breeding has advertised highly dynamic genetic variations in the form of SNPs, InDels and CNVs, and was found affected a number of genic and nongenic regions in the maize genome (Jiao et al. 2012). In soybean, 31 kb repeat segment as CNV at Rhg1 locus was observed in different haplotypes that product of this gene provide resistance for cyst nematode (SCN). One copy of the 31 kb segment per haploid genome was present in SCN-susceptible varieties. Therefore, SCN resistance was found to be linked with increased expression of the CNV-related genes (Cook et al. 2012). In addition to rice and maize, the recent association of CNVs and large InDel polymorphisms were demonstrated in wheat. CNV in the *Ppd-B1* gene was found to contribute to photoperiod sensitivity and another CNV in *Vrn-A1* was also found to be associated with intermediate or late flowering phenotypes (Díaz et al. 2012). For InDel polymorphism, 50 bp upstream region of the *Ppd-1* gene was shown to be related with heading time of wheat cultivars (Nishida et al. 2013). A comparison of CNVs in 14 barley genotypes including eight cultivars and six wild barleys relative to annotated genes identified a total of 5629 CNVs affecting exons (9.5 % of the exon sequences on the array) (Muñoz-Amatriaín et al. 2013). Trait-associated CNV in *Bot1* gene, which this gene is the boron efflux carrier and plays an significant role in boron tolerance, is another example for barley (Sutton et al. 2007). CNVs associated with disease resistance and biotic stress responses have also been determinated in *Arabidopsis* (Lu et al. 2012), rice (Xu et al. 2011) and soybean (McHale et al. 2012). The knowledge of genes affected by CNV may be advantageous for understanding of molecular mechanisms in the face of changing environmental conditions and possible threats posed by continuously evolving pest and pathogens.

4 Conclusion and Future Perspective

In the late nineteenth and early twentieth centuries, plant breeding has become a significant part of agricultural science. Conventional breeding techniques are still commonly used for development of cultivars and germplasm. Selection of desirable characteristics from different parent plants has become possible with the development of molecular marker methods in plant breeding. In the current genomics era, molecular markers techniques are bridging the gap between these desirable traits and genome sequence information. The use of the molecular markers has provided to create novel sources of genetic variations by introducing new and desirable characteristics from landraces and related grass species. With the expansion of high-throughput technologies, there has been a rapid growth in polymorphism knowledge

and the use of markers for various applications such as characterization of germplasm, identification genotypes, estimation of genetic distances between population, inbreeds and breeding material, detection of monogenic and QTL, MAS, identification of sequences of useful candidate genes, etc. The knowledge derived from marker technologies have provided opportunity to the plant breeders to create cultivar genotypes with the desired attributes following the concept of "Breeding by Design" that can yield better through improved growth and ability to withstand biotic and abiotic stresses (Peleman and van der Voort 2003). The combination of advanced methods will lead to facilitate the development of crop cultivars with improved yield, resistance, and quality.

References

Abdel-Haleem H, Ji P, Boerma HR, Li Z (2013) An R package for SNP marker-based parent-offspring tests. Plant Methods 9:44. doi:10.1186/1746-4811-9-44

Alheit KV, Reif JC, Maurer HP, Hahn V, Weissmann EA, Miedaner T, Wurschum T (2011) Detection of segregation distortion loci in triticale (x Triticosecale Wittmack) based on a high-density DArT marker consensus genetic linkage map. BMC Genomics 12:380

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Allen AM, Barker GLA, Berry ST, Coghill JA, Gwilliam R, Kirby S, Robinson P, Brenchley RC, D'Amore R, McKenzie N, Waite D, Hall A, Bevan M, Hall N, Edwards KJ (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). Plant Biotechnol J 9:1–14

Alsop BP, Farre A, Wenzl P, Wang JM, Zhou MX, Romagosa I, Kilian A, Steffenson BJ (2011) Development of wild barley-derived DArT markers and their integration into a barley consensus map. Mol Breed 27:77–92

Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. In: Somers DJ et al (eds) Methods in molecular biology. Plant genomics. Humana Press, New York, pp 19–39

Bachlava E, Taylor CA, Tang S, 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

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:e3376. doi:10.1371/journal.pone.0003376

Belo A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120:355–367

Blanco A, Colasuonno P, Gadaleta A, Mangini G, Schiavulli A, Simeone R, Digesù AM, De Vita P, Mastrangelo AM, Cattivelli L (2011) Quantitative trait loci for yellow pigment concentration and individual carotenoid compounds in durum wheat. J Cereal Sci 54:255–264

Blanco A, Mangini G, Giancaspro A, Giove S, Colasuonno P, Simeone R, Signorile A, De Vita P, Mastrangelo AM, Cattivelli L, Gadaleta A (2012) Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars. Mol Breed 30:79–92

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Chao S, Dubcovsky J, Dvorak J, Luo M-C, Baenziger S, Matnyazov R, Clark D, Talbert L, Anderson J, Dreisigacker S (2010) Population- and genome-specific patterns of linkage dis-

equilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). BMC Genomics 11(1):727

Chen W, Mingus J, Mammadov J, Backlund JE, Greene T, Thompson S, Kumpatla S (2010) KASPar: a simple and cost-effective system for SNP genotyping. In: Proceedings of plant and animal genome XVIII conference, San Diego, USA, p 194

Chia JM, Song C, Bradbury PJ et al (2012) Maize hapmap2 identifies extant variation from a genome in flux. Nat Genet 44:803–807

Chutimanitsakun Y, Nipper R, 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

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 SM, 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

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338:1206–1209

Davey JW, Blaxter ML (2010) RADSeq: nextgeneration population genetics. Brief Funct Genomics 9:416–423. doi:10.1093/bfgp/elq031

Desmarais E, Lanneluc I, Lagnel J (1998) Direct amplification of length polymorphisms (DALP), or how to get and characterize new genetic markers in many species. Nucleic Acids Res 26:1458–1465. doi:10.1093/nar/26.6.1458

Deulvot C, Charrel H, Marty A, Jacquin F, Donnadieu 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

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS One 7:e33234

du Manoir S, Speicher MR, Joos S, Schröck E, Popp S, Döhner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T (1993) Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Hum Genet 90:590–610

Edwards J, Mccouch S (2007) Molecular markers for use in plant molecular breeding and germplasm evaluation. In: Guimaraes EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection—current status and future perspectives in crops, livestock, forestry and fish. Food and Agriculture Organization of the United Nations (FAO), Rome, pp 29–49

Edwards D, Forster JW, Chagne D, Batley J (2007) What is SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silva HN (eds) Association mapping in plants. Springer, Berlin, pp 41–52

Flavell AJ, Bolshakov VN, Booth A, Jing AR, Russell J, Ellis THN, Isaac P (2003) A microarray-based high throughput molecular marker genotyping method: the tagged microarray marker (TAM) approach. Nucl Acids Res 31:e115

Foolad MR, Sharma A (2005) Molecular markers as selection tools in tomato breeding. Acta Hort 695:225–240

Francki MG, Walker E, Crawford AC, Broughton S, Ohm HW, Barclay I, Wilson RE, McLean R (2009) Comparison of genetic and cytogenetic maps of hexaploid wheat (*Triticum aestivum* L.) using SSR and DArT markers. Mol Genet Genomics 281:181–191

Gadaleta A, Giancaspro A, Giove SL, Zacheo S, Mangini G, Simeone R, Signorile A, Blanco A (2009) Genetic and physical mapping of new EST-derived SSRs on the A and B genome chromosomes of wheat. Theor Appl Genet 118:1015–1025

Ganal MW, Altmann T, Röder MS (2009) SNP identification in crop plants. Curr Opin Plant Biol 12:211–217

Goldstein D, Schlötterer C (1999) Microsatellites: evolution and applications. Oxford University Press, Oxford, UK

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity (Edinb) 101:5–18. doi:10.1038/hdy.2008.35

Gupta PK, Rustgi S, Mir RR (2013) Array-based high-throughput DNA markers and genotyping platforms for cereal genetics and genomics. In: Gupta PK, Varshney RK (eds) Cereal genomics II, doi: 10.1007/978-94-007-6401-9_2

Huttner E, Wenzl P, Akbari M, Caig V, Carling J, Cayla C et al (2005) Diversity arrays technology: a novel tool for harnessing the genetic potential of orphan crops. In: Serageldin I, Persley GJ (eds) Discovery to delivery: BioVision Alexandria 2004, Proceedings of the 2004 conference of the world biological forum. CABI Publishing, UK, pp 145–155

Huttner E, Caig V, Carling J, Evers M, Howes N, Uszynski G, Wenzl P, Xia L, Yang S, Risterucci A-M, Killian A (2006) New plant breeding strategies using an affordable and effective whole-genome profiling method. BioVision Alexandria. 73. Accessed 26–29 Apr 2006

Hyten DL, Cannon SB, Song Q et al (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11(1):38

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucl Acids Res 29, e25

James KE, Schneider H, Ansell SW, Evers M, Robba L, Uszynski G, Pedersen N, Newton AE, Russell SJ, Vogel JC, Killian A (2008) Diversity arrays technology (DArT) for pan-genomic evolutionary studies of non-model organisms. Plos One 3:e1682

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11:424–429

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J (2012) Genome-wide genetic change during modern breeding of maize. Nat Genet 44:812–815

Kallioniemi A, Kallioniemi O-P, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258:818–821

Khan F, Hakeem KR, Sidiqqui TO, Ahmad A (2013) RAPD markers associated with salt tolerance in Soybean genotypes under salt stress. Appl Biochem Biotech 170(2):257–272

Killian A, Huttner E, Wenzl P, Jaccoud D, Carling J, Caig V et al (2005) The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In: Tuberosa R, Phillips RL, Gale M (eds) Proceedings of the international congress in the wake of the double helix: from the green revolution to the gene revolution. Avenue Media, Bologna, Italy. pp 443–461, Accessed 27–31 May 2003

Kirchhoff M, Gerdes T, Rose H, Maahr J, Ottesen AM, Lundsteen C (1998) Detection of chromosomal gains and losses in comparative genomic hybridization analysis based on standard reference intervals. Cytometry 31:163–173

Konieczny A, Ausubel FM (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J 4:403–410

Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics 2012:831460. doi:10.1155/2012/831460

Lewis JA, Shiver AL, Stiffler N, Miller MR, Johnson EA, Selker EU (2007) High density detection of restriction site associated DNA (RAD) markers for rapid mapping of mutated loci in *Neurospora*. Genetics 177:1163–1171

Lichter P, Joos S, Bentz M, Lampel S (2000) Comparative genomic hybridization: uses and limitations. Semin Hematol 37:348–357

Litt M, Luty JA (1986) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Lu P, Han X, Qi J, Yang J, Wijeratne AJ, Li T, Ma H (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing *Landsberg*

erecta and all four products of a single meiosis. Genome Res 22:508–518. doi:10.1101/gr.127522.111

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 13:2291–2305

Mace ES, Rami JF, Bouchet S, Klein PE, Klein RR, Kilian A, Wenzl P, Xia L, Halloran K, Jordan DR (2009) A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers. BMC Plant Biol 9:13

Mammadov J, Chen W, Mingus J, Thompson S, Kumpatla S (2012) Development of versatile gene-based SNP assays in maize (*Zea mays* L.). Mol Breed 29:779–790

Mantovani P, Maccaferri M, Sanguineti MC, Tuberosa R, Catione I, Wenzl P, Thomson B, Carling J, Huttner E, De Ambrogio E, Kilian A (2008) An integrated DArT-SSR linkage map of durum wheat. Mol Breed 22:629–648

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

Marone D, Panio G, Ficco DB, Russo MA, De Vita P, Papa R, Rubiales D, Cattivelli L, Mastrangelo AM (2012) Characterization of wheat DArT markers: genetic and functional features. Mol Genet Genomics 287(9):741–753. doi:10.1007/s00438-012-0714-8

Matsumura H, Miyagi N, Taniai N, Fukushima M, Tarora K, Shudo A, Urasaki N (2014) Mapping of the gynoecy in bitter gourd (*Momordica charantia*) using RAD-seq analysis. Plos One 9:e87138

Maughan PJ, Smith S, Fairbanks D, Jellen E (2011) Development, characterization, and linkage mapping of single nucleotide polymorphisms in the grain amaranths (*Amaranthus* sp.). Plant Genome 4:92–101

McCouch SR, Zhao K, Wright M, Tung C, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML, McClung A, Eizenga G, Bustamante C (2010) Development of genome-wide SNP assays for rice. Breed Sci 60:524–535

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159:1295–1308. doi:10.1104/pp.112.194605

Meudt HM, Clark AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci 12:106–117

Miller MR, Atwood TS, Eames BF, Eberhart JK, Yan YL, Postlethwait JH, Johnson EA (2007a) RAD marker microarrays enable rapid mapping of zebrafish mutations. Genome Biol 8:R105

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007b) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248

Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21 century crop improvement. Plant Physiol 147:969–977. doi:10.1104/pp.108.118232

Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. Trends Ecol Evol 14:389–394

Mullis K (1990) The unusual origin of the polymerase chain reaction. Sci Am 262(4):56–61, 64–65

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14(6):R58. doi:10.1186/gb-2013-14-6-r58

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. Mol Breed 31:27–37

Nybom H (2004) Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol 13:1143–1155
Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor Appl Genet 85:985–993
Pegadaraju V, Nipper R, Hulke B, Qi L, Schultz Q (2013) De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. BMC Genomics 14:556. doi:10.1186/1471-2164-14-556
Peleg Z, Saranga Y, Suprunova T, Ronin Y, Röder MS, Kilian A, Korol AB, Fahima T (2008) High-density genetic map of durum wheat 9 wild emmer wheat based on SSR and DArT markers. Theor Appl Genet 117:103–115
Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 8:330–334
Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. Plos One 7:e32253. doi:10.1371/journal.pone.0032253
Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M (2008) Gene expression quantitative trait locus analysis of 16000 barley genes reveals a complex pattern of genomewide transcriptional regulation. Plant J 53:90–101
Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222
Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100. doi:10.1016/S1369-5266(02)00240-6
Rajib R, Abdelmoumen T, Hakeem KR, Mohamed RAG, Tah J (2013) Molecular marker-assisted technologies for crop improvement. In: Roychowdhury R (ed) Crop improvement in the era of climate change. I.K. International Publication House Pvt. Ltd, Delhi, India, pp 241–258. doi:10.13140/RG.2.1.2822.2560
Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. Mol Ecol 20:3499–3502
Ruane J, Sonnino A (2007) Marker-assisted selection as a tool for genetic improvement of crops, livestock, forestry and fish in developing countries: an overview of the issues. In: Guimaraes EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection current status and future perspectives in crops, livestock, forestry and fish. Food and Agriculture Organization of the United Nations (FAO), Rome, pp 3–13
Salimath SS, deOliveira AC, Bennetzen J, Godwin ID (1995) Assessment of genomic origin and genetic diversity in the genus Eleusine with DNA markers. Genome 38:757–763. doi:10.1139/g95–096
Sax K (1932) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560
Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. Proc Roy Soc B Biol Sci 277:3213–3221
Semagn K, Bjornstad A, Ndjiondjop MN (2006a) An overview of molecular marker methods for plants. Afr J Biotech 5:2540–2568
Semagn K, Bjornstad A, Skinnes H, Maroy AG, Tarkegne Y, William M (2006b) Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a doubled-haploid hexaploid wheat population. Genome 49:545–555
Sobrino B, Briona M, Carracedoa A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154:181–194
Song Q, Jia G, Zhu Y, Grant D, Nelson RT, Hwang EY, Hyten DL, Cregan P (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. Crop Sci 50(5):1950–1960
Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503–517
Springer NM, Ying K, Fu Y, Ji TJ, Yeh C-T, Jia Y, Wu W, Richmond TA, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable P (2009) Maize inbreds

exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. Plos Genet 5:11
Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318:1446–1449
Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689–1699
Tanksley SD (1993) Mapping polygenes. Annu Rev Genet 27:205–233
Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822. doi:10.1126/science.1183700
Trebbi D, Maccaferri M, de Heer P, Sorensen A, Giuliani S, Salvi S, Sanguineti MC, Massi A, van der Vossen EAG, Tuberosa R (2011) High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theor Appl Genet 123:555–569
Vignal A, Milan D, SanCristobal M et al (2002) A review of SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol 34:275–305
Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
Weising K, Nybom H, Wolff K, Kahl G (2005) DNA fingerprinting in plants—principles, methods, and applications. CRC Press, Boca Raton, FL, USA
Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Killian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci U S A 101:9915–9920
Wenzl P, Li H, Carling J, Zhou M, Raman H, Paul E, Hearnden P, Mailer C, Xia L, Caig V, Ovesná J, Cakir M, Poulsen D, Wang J, Raman R, Smith KP, Muehlbauer GJ, Chalmers KJ, Kleinhofs A, Huttner E, Killian A (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC Genomics 7:206
Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535
Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, Li Y, Meade F, Teodor R, Vaistij FE, Walker C, Bowser TA, Graham IA (2012) A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noncapine. Science 336:1704–1708
Wittenberg AHJ, van der Lee T, Cayla C, Kilian A, Visser RGF, Schouten HJ (2005) Validation of the high throughput marker technology DArT using the model plant Arabidopsis thaliana. Mol Genet Genomics 274:30–39
Xia L, Peng K, Yang S, Wenzl P, De Vicente MC, Fregene M, Killian A (2005) DArT for high-throughput genotyping of cassava (Manihot esculenta) and its wild relatives. Theor Appl Genet 110:1092–1098
Xu Y (2010) Molecular plant breeding. CAB Int. ISBN: 978 1 84593 392 0, p 734
Xu X, Liu X, Ge S et al (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111
Yang H, Tao Y, Zheng Z, Li C, Sweetingham MW, Howieson JG (2012) Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in *Lupinus angustifolius* L. BMC Genomics 13:318. doi:10.1186/1471-2164-13-318
Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H, Wang Y, Tang S, Wei X (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. BMC Genomics 12:372
Zhang W, Chao S, Manthey F, Chicaiza O, Brevis JC, Echenique V, Dubcovsky J (2008) QTL analysis of pasta quality using a composite microsatellite and SNP map of durum wheat. Theor Appl Genet 117:1361–1377
Zhou G, Zhang Q, Zhang XQ, Tan C, Li C (2015) Construction of high-density genetic map in Barley through restriction-site associated DNA sequencing. Plos One 10(7):e0133161. doi:10.1371/journal.pone.0133161

Transcriptomic Responses of Barley (*Hordeum vulgare* L.) to Drought and Salinity

Filiz Gürel, Neslihan Z. Öztürk, and Cüneyt Uçarlı

Contents

Abstract Drought and salinity are the main factors limiting plant growth and productivity. With the effects of global warming, severe drought episodes are expected to be widespread, which will certainly lead to decrease in crop production. Therefore, understanding plants' response to drought and salinity stresses is more urgent than ever to reveal molecular mechanisms behind the natural tolerance which, then, can be used in the generation of stress-tolerant crop species. Barley stands out as the most salinity and drought-tolerant crop in *Poaceae* family with its wide range of wild genotypes. Due to its higher tolerance to abiotic and biotic stresses among other crops, it was studied to understand the mechanisms behind the natural tolerance via generation of various genetic resources and databases created by extensive sequence data, microarray studies, next-generation sequencing (NGS), and genetic maps. Large-scale transcriptomic analyses in barley showed that ROS-scavenging enzymes, transcription factors, LEA group proteins, and enzymes coding for osmoprotectants are the prominent groups of genes differentially expressed under salinity and drought stresses. Quantitative real-time PCR was efficiently used to measure transcript levels of stress-related genes under high salt or limited water conditions,

F. Gürel (✉) • C. Uçarlı
Department of Molecular Biology and Genetics, Faculty of Science, Istanbul University, Vezneciler, Istanbul 34134, Turkey
e-mail: filiz@istanbul.edu.tr1

N.Z. Öztürk
Department of Agricultural Genetic Engineering, Faculty of Agricultural Sciences and Technologies, Nigde University, Central Campus, Nigde 51240, Turkey

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_7

allowing the prediction of functional characteristics of these genes according to their expression patterns. Small-scale expression studies also revealed the importance of cell and tissue type expression and mode of the stress treatment. However, although there are numerous candidate barley genes that can be used to develop transgenic crops with higher tolerance to salinity and drought, there are only limited isolation and cloning studies with these genes. We highly recommend more detailed studies on this naturally tolerant crop to be able to generate more drought or salt tolerance species via genetic transformation.

Keywords Barley • Transcriptomics • Drought • Salinity • Transgenics

1 Introduction

Environmental drought and soil salinity in arid and semi-arid regions are the major problems for agricultural productivity and threat for plant biodiversity (Godfree 2012; Trenberth et al. 2014). In addition to local drought, model-based analyses of soil moisture, drought indices, and precipitation-minus-evaporation studies are showing an increased risk of global drought in the twenty-first century (Dai 2013). Agricultural drought is a period with dry soils resulting from insufficient precipitation, intense but less frequent rainfalls, high evaporation or agricultural practices preventing water infiltration in the soil (Bot and Benites 2005; Dai 2011). Salinity also negatively affects the quality of soil and limits plant growth and production. Saline soils are formed by the basins with limited or no access to the rivers due to diverse soil types or unsuitable irrigation, poor drainage, and high evaporation. Soil salinity can be in different types as "irrigation-induced" and "transient" dry-land salinity (Läuchli and Grattan 2007). It is estimated that approximately 7 % of total land area and 20 % of the irrigated agriculture land is affected by salinity (Rozema and Flowers 2008; Agarwal et al. 2013).

Climatic changes all around the world with decreasing regional annual rainfall and increasing temperatures, in addition to obligation of agriculture in saline drylands, emphasize the importance of development of stress-tolerant cultivars, especially for important crop species (Cattivelli et al. 2008; Gosal et al. 2009; Mir et al. 2012). Tolerance to abiotic stresses such as drought and salinity is possessed by rare plant species including halophytes and xerophytes. Barley is known to have a significant potential tolerance to both drought and salinity stresses as described by physiological, morphological, and genomics tools; and, therefore, selected as an excellent model organism for stress response studies (Knüpffer et al. 2003; Nevo and Chen 2010; Mian et al. 2011; Roy et al. 2014). Other commonly used model organisms such as *Arabidopsis* and rice cannot complete their life cycle when exposed to 100 mM NaCl, while barley keeps its shoot growth at even higher concentrations of NaCl (200–300 mM) for at least 3 weeks (Munns and Tester

2008). Barley also has wild relatives such as *Hordeum marinum* which can grow in salty marshes and seasides, showing high salinity resistance of an unknown mechanism.

The transcriptome means the complete set of all transcripts in a cell and their quantity, for a physiological circumstance or specific developmental stage (Wang et al. 2009). Today it is known that understanding the transcriptome is important for interpreting the functional elements of the genome (Jain 2011). The key aims of the transcriptomics are: (1) to catalogue all transcripts, including mRNAs, non-coding RNAs, and small RNAs; (2) to determine the transcriptional structure of genes; and (3) to quantify the changing expression levels of each transcript during development and under different conditions (Wang et al. 2009).

In this chapter, we will provide the present knowledge on barley's response to salinity and drought stress as characterized by microarray, RNA sequencing (RNA-Seq) and more specifically quantitative real-time PCR. As many crop plants are subjected to combined stresses of drought and saline conditions in the field, pathways and gene networks overlapping at the molecular level in these two stress factors will also be discussed for a more realistic and useful approach.

2 Barley (*Hordeum vulgare* L.) as a Model

Among the cereal species, barley (*Hordeum vulgare* L.) is one of the oldest in the world and was first cultivated in Neolithic times in Fertile Crescent (Zohary and Hopf 1993; Badr et al. 2000; Morrell and Clegg 2007), from where it spread to the other parts of the world including South Africa, Europe, Near East, North Asia, China, and Japan. Barley is virtually found worldwide since it has a high adaptation to broad range of environments including steppes, savannas, mountains, as well as temperate zones in subtropics and subarctics (Nevo 1992; Hayes et al. 2003; Ullrich 2011). Based on the center of diversity concept of Vavilov, Anatolia (present-day Turkey) is one of the thirty-six agro-ecological groups where cultivated two- and six-rowed barley distributes along with the dry central and wet coastal regions (Knüpffer et al. 2003).

Three gene pools within the *H. vulgare* can be defined as "wild barley," adapted to diverse environments (mostly semi-arid); "landraces," adapted to marginal low-input agricultural regions; and "elite cultivars," grown in high-yielding agricultural regions. Wild progenitor of the barley, *Hordeum vulgare* spp. *spontaneum* C. Koch, which is also known as *Hordeum spontaneum*, has distribution in Fertile Crescent and Irano-Turanian as its primary and Mediterranean and Central Asia as secondary habitats (Zohary and Hopf 1993; Badr et al. 2000). As an example, distribution of wild barley covers from Çanakkale, North of Turkey alongside with the western coastal region to Southeastern Turkey. The comparison of the summer rainfall measurements in Çanakkale (13.4 kg/m^2) and Ceylanpınar (1.76 kg/m^2) shows the extreme climatic differences throughout the area. In addition, agricultural fields of

barley and wheat in Southeastern regions were frequently occupied by *H. spontaneum* during severe drought seasons (A. Karagöz, personal communication). It is well known that genetic diversity of cultivated barley was greatly reduced by domestication in elite cultivars (Morrell and Clegg 2007; Kilian et al. 2010) and different allele frequencies were occurred by the adaptation to different eco-geographical environments in three kinds of gene pools (Badr et al. 2000; Morrell et al. 2014). The observations summarized above imply the importance of wild barley populations for studying abiotic stress tolerance mechanisms.

Full genome sequence information is significant for understanding genome structure, evolutionary relatedness and variation, and thus development of modern strategies for crop breeding programs. Barley is diploid ($2n = 2x = 14$) and has seven chromosome pairs, designated as 1H to 7H with approximately 5.1 Gbp in size (von Bothmer 1992; Dolezel et al. 1998). An oligonucleotide array containing 350,000 high-quality ESTs and an SNP-genotyping platform were previously made available for barley (Close et al. 2009, 2004). High-throughput sequencing of barley genome was initiated on 2009 and resulted in the development of a 4.98 Gbp physical map with expressed barley genes in "Morex" cultivar (Wicker et al. 2009; Steuernagel et al. 2009; Mayer et al. 2011; The International Barley Genome Sequencing Consortium 2012). Results indicated that 84 % of barley genome comprises of mobile elements and repetitive DNA. Based on the homology comparisons with the genomes of close relatives (*Sorghum*, rice, *Brachypodium*) and *Arabidopsis*, barley has 26,159 genes of which 75 % have a multi-exon structure. These results proved the high abundance of alternative splicing in barley. Extensive SNP variation and transcriptionally active regions which are homolog to rice and *Brachypodium* have also been identified in the barley genome (The International Barley Genome Sequencing Consortium 2012).

In addition to a wide range of genomic studies proving large genomic diversity, there are other factors making barley a preferred model plant for abiotic stress research. For example, a high resistance to drought, salinity, and fungal diseases were proven in barley (Knüpffer et al. 2003; Bonman et al. 2005). Barley is also more alkaline tolerant than other cereal species (van Gool and Vernon 2006), and salt tolerance of barley cultivars is higher than bread wheat and other cultivated *Triticeae* (Garthwaite et al. 2005). However, as will be summarized in the following sections, barley was mostly studied on its tolerance to salt stress, and more studies are still needed to understand the mechanisms behind this natural abiotic and biotic tolerance.

3 Large-Scale Transcriptomics for Drought

Plants change their metabolism to reduce adverse effects of cellular dehydration caused by drought through well-conserved molecular and biochemical changes in cellular level. The complex network of metabolic changes can simply be divided into two: (1) changes in single function genes and enzymes; such as accumulation

of osmolytes, radical oxygen scavenging proteins and enzymes, ion transporters, channel proteins, and enzymes involved in lipid biosynthesis; and (2) regulatory proteins including transcription factors, protein kinases/phosphatases and proteinases responsible from the reprogramming of the metabolism in response to dehydration (Cominelli et al. 2013). A diverse family of transcriptional factors (DREB/CBF, ABF, AP2/ERF, bZIP, NAC, MYB, MYC, HD-ZIP, bHLH, NF-Y, EAR, and WRKY) is responsible for changes in gene expression under drought conditions and almost all of them belong to large family of proteins involved in regulation of several plant functions (Bhargava and Sawant 2013; Cominelli et al. 2013; Osakabe et al. 2014).

Transcriptional profiling through microarray technology made a breakthrough on understanding molecular responses of plants to abiotic stress conditions in gene expression level (Kilian et al. 2012). The very first microarray study on drought stress response of barley cv. Tokak plants on transcriptome level was performed by Ozturk et al. (2002) using an in-house printed cDNA array with 1,463 DNA elements from 6 to 10 h shock-drought-stressed leaf and root cDNA libraries constructed by the authors. The shock-drought treatment used in this study was simply done by removing 3-week-old barley plants from pots and leaving them on bench under growth conditions for 6 h (RWC 70 %) and 10 h (RWC 64 %). Hybridization was carried out with Cy3 and Cy5-dUTP labeled drought stressed leaf and root RNA with their appropriate controls. The majority of the drought stress cDNA libraries belonged to no hit (9.5 %) and unclassified protein (27.5 %) categories due to limited protein and gene information available on databanks at that time. Nevertheless, Ozturk et al. (2002) in their study was able to show the differential up-regulation of several stress-responsive genes in leaf and root tissues, including several jasmonate-induced proteins (jasmonate biosynthesis), metallothionein-like proteins, dehydrins (cellular protection), late embryogenesis abundant proteins (cellular protection), Δ 1-pyrroline-5-carboxylate synthetase (proline metabolism), wheat aluminum-induced proteins, several protein phosphatases (signal transduction), actin binding protein (cellular protection), auxin-induced protein, cytochrome P450 homologs, and cell death suppressor proteins. Down-regulated genes mostly belonged to photosynthesis metabolism such as ribulose-biphosphate carboxylase activase, rubisco small-chain precursor, and chlorophyll a/b-binding proteins. As conclusion, the authors cautioned the researchers that shock-stress treatment was effective to clone large number of drought stress-responsive genes but not comparable to field situation; therefore might not be effective to understand time-dependent changes in transcriptome in response to slow-developing drought. Indeed, Talame et al. (2007) hybridized a cDNA array containing 1,654 DNA elements from Ozturk et al. (2002) and RNA from control leaves and roots with 7 days (RWC 91 %) and 11 days (RWC 81 %) slow drought-treated barley Er/Apm variety from ICARDA (well adapted to dry environment) leaves and showed that a lower number of differentially regulated transcripts were obtained by gradual stress compared to shock-like stress treatment performed by Ozturk et al. (2002). Talame et al. (2007) also indicated that although only about 10 % of the transcripts shared similar changes in these two approaches, a considerable number of transcripts showed similar

regulation in both cases, and even concluded that a shock-like treatment can be an effective way of identifying and characterizing alleles involved in adaptive response to dehydration. Indeed, Diab et al. (2004) used cDNA libraries constructed by Ozturk et al. (2002) for QTL analysis in a population of 167 F_8 recombinant inbred lines and identified two candidate genes and ten differentially expressed sequences that were associated with QTLs in drought-tolerant traits.

Another study with cDNA microarray consisting more than 300 DNA elements derived from cold, dehydration, salinity, high light, and copper-treated barley cv. Nure leaves was published on 2004 (Atienza et al. 2004). The authors performed a shock-like dehydration treatment; however, they spotted only 20 genes from 5 h (RWC 89 %) to 10 h (RWC 80 %) drought-treated leaf samples and, therefore, revealed limited information on changes in barley transcriptome in response to dehydration.

On 2004, Affymetrix 22K Barley1 GeneChip Array made available to research (Close et al. 2004). The array contained 21,439 probes derived from 350,000 ESTs from 84 cDNA libraries and 1,145 barley gene sequences from National Center for Biotechnology Information database (NCBI, http://www.ncbi.nlm.nih.gov). Although barley Affymetrix array is available for a while, there is surprisingly low number of publications using this effective platform to identify transcriptomic changes in response to drought. Guo et al. (2009) used 22K Affymetrix Barley 1 microarray to compare transcriptomes of two drought-tolerant barley genotypes (cv. Martin and *Hordeum spontaneum* 41-1) and one drought-sensitive genotype (cv. Moroc9-75) in response to drought during reproductive stage (AWC 70 % in control and 10 % in drought; hybridizations with total RNA from leaf tissues collected at 0, 1, 3, and 5 days on 10 % AWC). The main aim of the study was to identify drought stress tolerance-related genes, and the authors stated identification of 17 genes specifically induced in drought-tolerant genotypes that can be related with tolerance through controlling stomatal closure via carbon metabolism (NADP malic enzyme and pyruvate dehydrogenase), glycine-betaine synthesis (C-4 sterol methyl oxidase), reactive oxygen scavenging (aldehyde dehydrogenase, ascorbate-dependent oxidoreductase), and membrane and protein stabilization (heat shock proteins and dehydrin). In addition, calcium-dependent protein kinase (signal transduction), membrane steroid binding protein (signal transduction), G2 pea dark accumulated protein (anti-senescence), and glutathione S-transferase (detoxification)-related genes were reported to be constitutively expressed in drought-tolerant genotypes, and the authors indicated the possibility of the direct role of these proteins in genotypic differences in drought tolerance. It is important to note that in their studies, Guo et al. (2009) observed differential regulation in a total of 263 genes (collective from all genotypes) which stands for just over 1 % of whole array; whereas Ozturk et al. (2002) in their study reported change in gene expression about 50 % of all cDNA array probes; which indicates that only a certain percentage of plant genome is transcriptionally regulated in response to drought and probes on the microarray are an important point to consider in a study aiming to detect complete transcriptomic changes in response to stress conditions.

The second comprehensive use of Affymetrix 22K Barley1 GeneChip Array was with barley cultivar Morex, where the authors investigated gene expression in the spike organs (lemma, palea, awn, and seed) exposed to drought treatment for 4 days during the grain-filling stage (RWC dropped from 85 to 60 % in lemma, palea, and awn; and from 89 to 81 % in seed) (Abebe et al. 2010). There was almost no change in transcript abundance in seed tissue; whereas the comparison suggested better drought tolerance in lemma and palea than awn. Among the stress defense-related genes, the expression of NADPH oxidase, ribosome inactivating proteins, chitinases, protease inhibitors, and amylases induced upon dehydration. The accumulation of transcripts belonging to late embryogenesis related proteins (LEA) in the lemma, palea, and awn was an indicative of protective roles of LEA proteins during dehydration via retention of water, sequestration of ions, and stabilization of proteins and chaperons of protein folding.

As indicated before, although effective, the information that can be gathered from microarray studies is limited to the extent of DNA elements on the array. Use of new technologies, mainly next-generation sequencing (NGS) via RNA-Seq, on the other hand, makes studies on transcriptome level more feasible and creates larger datasets compared to microarray approaches. There is only one publication using this new approach for the identification of drought stress-responsive transcripts in barley (Bedada et al. 2014). In this study, Bedada et al. (2014) used two wild barley ecotypes (B1K2, desert and B1K30, Mediterranean) and compared transcriptomes of 5 days of drought-stressed leaf tissues (SWC 80 % in control and 30 % in drought stress). Normalized cDNA libraries were sequenced by 454 platform and the authors identified over 800 unique transcripts from each ecotype. The majority of these unique transcripts were homologs of transcription factors (bZIP, bHLH, MYB), heat shock proteins, aquaporins and ERD, LEA, and ABC transporter proteins. As conclusion, Bedada et al. (2014) pointed out the genomic differentiation between the desert and Mediterranean wild barley ecotypes.

4 Large-Scale Transcriptomics for Salinity

Analyzing the functions of stress-inducible genes is important to understand the molecular mechanism of stress tolerance and to improve stress tolerance of crops by genetic manipulation (Seki et al. 2009). During the last decade, DNA microarrays have been the technology of choice for large-scale studies of salt regulated gene expression levels in model plants and crops, such as *Arabidopsis* (Richards et al. 2012), rice (Walia et al. 2005), barley (Walia et al. 2006; Guo et al. 2009), wheat (Kumar et al. 2014), maize (Allardyce et al. 2013), and tobacco (Edwards et al. 2010). Many genes that respond to salinity stress at the transcription level in barley were identified using microarray technology (Ozturk et al. 2002; Ueda et al. 2004; Walia et al. 2006; Gao et al. 2013) (Table 1).

First microarray analysis of barley transcripts under salinity stress was reported by Ozturk et al. (2002). Responses to salinity were investigated by microarray

Table 1 Summary of large-scale transcriptomic analyses for salt stress response in barley

Transcriptome analysis method	Cultivar and stress treatment	No. of up-regulated/ down-regulated genes in shoot tissue	No. of up-regulated/ down-regulated genes in root tissue	Reference
cDNA microarray (1463 DNA elements)	cv. Tokak; 150 mM NaCl; 24 h	29/27	10/9	Ozturk et al. (2002)
cDNA microarray (460 salt-responsive genes)	cv. Haruna-nijyo; 200 mM NaCl; 24 h	62/30	49/14	Ueda et al. (2004)
cDNA micro array (22K barley chip 1)	cv. Morex; 100 mM NaCl and 10 mM $CaCl_2$; 3, 8, and 24 h	339/311	N/A	Walia et al. (2006)
cDNA micro array (22K barley chip 1)	cv. Hua 30 (salt sensitive) cv. Hua 11 (salt tolerant); 300 mM NaCl, 30 mM $CaCl_2$; 6 h	1090/763	864/609	Gao et al. (2013)
mRNA-Seq (Illumina)	cv. Hindmarsh;150 mM NaCl; 12 h	48/62	N/A	Ziemann et al. (2013)

N/A not available

hybridization of 1463 DNA elements derived from cDNA libraries of 6 and 10 h drought-stressed plants. Salt stress was carried out on barley cv. Tokak plants in hydroponic tanks containing one-third strength Hoagland's solution (Hoagland and Arnon 1950) supplied with 150 mM NaCl for 24 h. Equal or greater than 2.5-fold changes in expression of genes were considered significant. There were 37 and 36 transcripts that showed increased or decreased expression level in salt-stressed leaf and root tissues, respectively. A limited number of the overlapping genes were found among up- or down-regulated transcripts in root and leaf tissues such as 60S acidic ribosomal protein P0 and ubiquitins (10 and 11). Metallothionein-like protein type 2 (*MT2*) was the most significantly up-regulated transcript (36-fold) in leaf tissues; however, its expression was not changed in root tissues. While the genes encode allene oxide synthase, basic proline-rich protein (PRB1L) precursor, lipid transfer protein cw 18, glutathione S-transferase (auxin induced), early responsive to dehydration 1 (ERD1), and Δ 1-pyrroline-5-carboxylate synthetase (P5CS) were induced in leaf tissues of barley, fructose-bisphosphate aldolase, elongation factor 1-alpha, ubiquitin 4 and 10, and germin-like protein were the highly induced transcripts in root tissues. The number of the down-regulated transcripts in salt-stressed leaf was threefold higher than the transcripts of root. Auxin-induced protein, late embryogenesis abundant protein LEA14-A, aminopeptidase N were the significantly down-regulated transcripts in leaf tissues.

Ueda et al. (2004) also performed a customized cDNA microarray study using 460 salt-responsive genes obtained by a previous differential display analysis (Ueda et al. 2002). The 15-day-old seedlings of barley cv. Haruna-nijyo were treated with 200 mM NaCl in half-strength Hoagland's solution. Salt-stressed plants were harvested after 1 and 24 h of salt stress. In barley roots, a total of 49 genes were found to be induced during the initial phase under salt stress. After 1 h salt stress, 13 genes showed up-regulation. These genes encode signaling elements such as receptor-like protein and serine/threonine protein kinase as well as stress tolerance genes like plasma membrane protein 3 (PMP3). After 24 h, the genes involved in the biosynthesis and transport of amino acids and genes of the cytochrome P450 family were up-regulated (16-fold) in barley roots. A total of 14 genes were found to be repressed under salt stress in barley roots. The transcript levels of water channel protein 2, phospholipase C, and salt-responsive protein (SalT) were down-regulated within 1 h, while the mRNA abundance of water channel protein 1 and DNA-binding protein CCA1 showed significant down-regulation only by 24 h under stress. The number of up-regulated genes was less in leaves than in roots. Twenty-six of these 460 genes showed increased level of expression. The calcium-dependent protein kinase, phosphatidylinositol-4-phosphate-5-kinase, pleiotropic drug resistance 5-like ATP binding cassette transporter, and *SET1* involved in signal transduction were up-regulated in barley leaves within 1 h under salt stress. After 24 h, stress tolerance genes encoding P5CS, PMP3, aldehyde dehydrogenase, betaine aldehyde dehydrogenase were highly induced in barley leaves. Besides, expression level of the genes encoding enzymes involved in the biosynthesis of amino acids such as methionine synthase and asparagine synthetase, and lipoxygenase related to jasmonic acid biosynthesis were increased significantly. Out of 460 genes, 21 genes showed down-regulation in barley leaves. The genes encoding phosphoenolpyruvate carboxylase and omega-3 fatty acid desaturase were repressed significantly at 24 h time point during salt stress. The authors indicated that the barley root cells responded to salt stress signals earlier than leaf cells based on the observation that larger numbers of the differentially expressed genes existed in roots than in the leaf cells after 1-h salt stress. While expression of PMP3 gene was induced in root cells at 1 h, expression of it was not changed in leaf cells. Although the expression of cytochrome P450 family was highly up-regulated (16-fold) in barley roots after 24 h, they were suppressed in leaf cells. There were also common genes differentially expressed in root and leaf cells after 24 h salt stress including up-regulation of P5CS and aldehyde dehydrogenase, and down-regulation of water channel 2.

Walia et al. (2006) reported the first use of the Affymetrix 22K Barley1 GeneChip Array in identification of transcripts under salt stress. A gradual salt stress was applied to 14-day-old barley cv. Morex plants. The NaCl concentrations were elevated to 100 mM by increments of 25 mM NaCl per day. The plants were harvested at 3, 8, and 27 h after reaching a final concentration of 100 mM NaCl. In the study, three time points based on the physiological status of plants in response to addition of salt were chosen to investigate the reducing growth rate stage (3 h), growth rate recovery stage (8 h), and ion-specific responses (27 h). A significant difference in Na^+ level was found between root (639 ± 124 mmol/kg) and shoot samples

(897 ± 41 mmol/kg). There were 261 probe sets corresponding to 339 unigenes and 234 probe sets (311 unigene) that showed up-regulation and down-regulation at three different time points (3, 8 and 27 h), respectively. A fold change of 1.5 was considered as an indication of significant modulation in gene expression. While less than 10 % of the up-regulated probe sets (25) were overlapped between three distinct time points (3, 8, 27 h), this ratio was less than 6 % among down-regulated probe sets (total number of down-regulated probes were 14). The relatively higher number of induced and repressed genes was observed at time points of 27 and 8 h, respectively. The nodulin MtN3 family protein, P5CS, C-4 sterol methyl oxidase, arginine/serine-rich protein were the common up-regulated probe sets at all three time points. Expression level of the photosystem II 10 kDa polypeptide was increased at all time points under salt stress and reached maximum at 27 h. Peroxidase was the highest down-regulated transcript among the repressed genes. Some of the genes involved in jasmonic acid biosynthesis and jasmonic acid-responsive genes including phospholipase, lipoxygenase, and allene oxide synthase were up-regulated especially at the 3 h time point under salt stress. However, 12-oxophytodienoate reductase, involved in the biosynthesis or metabolism of oxylipin signaling molecules, was significantly down-regulated at the 8 h time point. Walia et al. (2006) also found a number of known jasmonic acid-responsive genes with altered expression under salt stress. These were O-methyltransferase, jasmonic acid-induced proteins, glutathione S-transferase, selenium binding protein, and hordothionins. The expression level of some genes known to respond to other abiotic stress conditions such as osmotic stress and cold were found to be increased under salt stress. Well-known dehydration-responsive genes, *Dhn5* and *RD22* were induced at 3 and 27 h time points. Walia et al. (2006) showed that several well-known sodium transporters and anti-porters such as *HvNHX1* were differentially expressed (at least 1.5-fold) under salt stress at any time point. The osmoprotectants such as proline and glycine betaine are known to be accumulated in plant cells in response to the salt and drought stress in many plants (Kavi Kishor et al. 1995). The gene encoding P5CS, the key enzyme for proline biosynthesis, was the only gene found to be commonly induced in all three studies (Ozturk et al. 2002; Ueda et al. 2004; Walia et al. 2006). Gao et al. (2013) was the first comparing salt-tolerant barley (cv. Hua11) with a salt-sensitive cultivar (cv. Hua30) to identify salt stress-regulated genes on a large scale using Affymetrix 22K Barley1 GeneChip Array. At 7th day of germination, seedlings were suspended in a half-strength Hoagland solution. Salt stress (300 mM NaCl) was applied to 10-day-old seedlings grown in Hoagland's solution supplemented with 30 mM $CaCl_2$ for 6 h. Salt stress was shown to increase the production of reactive oxygen species (ROS) and cause ROS-associated injury (Miller et al. 2010). Therefore, it was concluded that the activities of ROS-scavenging enzymes like superoxide dismutase (SOD) and peroxidase (POD) play important roles in the protection of barley plants under salt stress. Gao et al. (2013) reported that the SOD and POD activities in cv. Hua11 were higher than cv. Hua30 in root and shoot tissues. A total of 1853 (1090 up-regulated and 793 down-regulated) and 1473 (864 up-regulated and 609 down-regulated) differentially expressed probe sets were found in the shoot and root tissues, respectively,

after 6 h salt treatment. When differently expressed genes were compared between Hua11 and Hua30, it was revealed that 916 (62 %) and 842 (45 %) genes were co-regulated in root and shoot tissues, respectively. Gao et al. (2013) concluded that these genes might be responsible for barley intrinsic tolerance to salt stress. Furthermore, the number of salt-responsive genes identified from shoot tissues (1853) was more than that of root tissues (1473). Number of the up-regulated probe sets in shoot tissues of tolerant cultivar (cv. Hua11) was higher than sensitive cultivar (cv. Hua30) with 906 and 685 probe sets, respectively. On the contrary, there were higher number of down-regulated probe sets in root and shoot tissues of cv. Hua30 than cv. Hua11. The number of the significantly changed probe sets related to signal transduction in salt-tolerant genotype was higher than the sensitive genotype in root and shoot. Although 14 receptor-like kinases and 3 mitogen-activated protein kinases were induced in the shoot tissue, expression level of these was not changed in root tissues. In the study, a set of hormone-related genes including jasmonic acid and gibberellins, ethylene, and cytokinin were identified in root and shoot tissues under salt stress. Among them, three lipoxygenases were up-regulated in shoot of cv. Hua11 after salt treatment. In addition, one cytokinin dehydrogenase transcript was significantly up-regulated in root tissues of salt-tolerant genotype. In the salt-tolerant genotype, 30 genes encoding transcription factors (ZIM, WRKY, MYB, CBF, NAC, bHLH) were characterized in response to salt stress in shoot, in addition to 6 genes in root tissue. Most of these genes were up-regulated except 4 from bZIP and AP2 families. Among these genes, especially one C-repeat binding factor and one bHLH transcription factor showed a special expression pattern with respect to the other transcription factors. They were both induced in the shoot tissues of salt-tolerant genotype and down-regulated in the shoot tissues of salt-sensitive genotype. Organic solutes, known as compatible solutes, including amino acids, sugars, and other low molecular weight metabolites, protect the plants from biotic and abiotic stress with osmotic adjustment, protection of membrane integrity, and stabilization of enzymes or proteins (Ashraf and Foolad 2007). After 6 h salt treatment, 15 genes in shoot, and 11 genes in root associated with compatible solutes and secondary metabolites were found to be differentially regulated. All of the genes encoding the compatible solutes such as one proline-degraded enzyme and four sugar-related genes were induced in shoot and root tissues. One of the terpenoid-related genes was down-regulated in roots.

Next-generation sequencing (NGS) technologies have become faster, more accurate, and less expensive in recent years, leading to their widespread application in diverse fields (Hawkins et al. 2010). This technology has been applied mainly in human genetics and medicine, and is now emerging as the most popular high-throughput sequencing technique in plants (Ziemann et al. 2013; Xu et al. 2012; Camilios-Neto et al. 2014). Ziemann et al. (2013) used RNA-seq technology to analyze barley genes involved in salt stress. Two-week-old barley (cv. Hindmarsh) seedlings were treated with 150 mM NaCl in Hoagland's solution (Hoagland and Arnon 1950) for 12 h. Plant leaves were harvested and analyzed by RNA-Seq. The total number of RNA sequence reads, generated in the RNA-Seq experiment from sequencing of salt-stressed and control plants were 26.7 million and 23.7 million,

respectively. The 16.4 million (63 %) reads of salt-stressed cDNAs were uniquely aligned using burrows-wheeler aligner (BWA) (Li and Durbin 2009) according to barley Unigene database. Differential gene expression between the control and salt-stressed samples was analyzed using DESeq software (Anders and Huber 2010). When controlling the false discovery rate (FDR) at 5 %, 110 genes were found differentially expressed under acute salt stress. While 48 of them were significantly up-regulated, 62 of them were significantly down-regulated. Late embryogenesis abundant protein was strongly up-regulated (over 32-fold) under salinity stress. Similarly Ueda et al. (2004) and Walia et al. (2006) were reported that this gene confers salinity tolerance in barley. Other significantly induced genes were cellulose synthase-like protein, lipoxygenase 2.1, protein phosphatase 2C, calcium/calmodulin-dependent protein kinase, as well as those encoding membrane bound proteins such as a peptide transporter, two plasma membrane ATPases and a novel wall-associated receptor kinase. Down-regulated transcripts include those in the jumonji, pumilio RNA binding, and MYB transcription factor classes, and also several transcripts of unknown function.

5 Small-Scale Expression Analyses of Stress-Responsive Genes

Tolerance of barley to drought stress was found to be related with several major classes of genes involved in signaling and protection in the plant cells. In addition to large-scale transcriptome analyses, there are also reports revealing the expressional change in this group of genes in response to abiotic stress conditions. Barley cultivars with contrasting phenotypes for their tolerance to salinity and drought were occasionally used to investigate transcriptional profiles of specific genes involving in stress tolerance. In this part of the chapter, we will try to emphasize the ones that are mostly related to drought and salt stress response and/or tolerance.

The most important group of genes activating by abiotic stresses are transcription factors (TFs) that modulate the expression of stress-related genes. The APETALA 2/ethylene-responsive element binding factors (AP2/ERF) is a large family of TFs in plants and regulates gene expression during abiotic stress responses (Mizoi et al. 2012). In barley, an AP2/ERF gene, dehydration-responsive factor 1 (*HvDRF1*), was shown to be involved in the regulation of stress-responsive genes by an abscisic acid (ABA)-mediated pathway (Xue and Loveridge 2004). Binding motif of *HvDRF1* was identified as T(T/A)ACCGCCTT. Increase in *HvDRF1* by drought, salinity, and ABA treatment was shown to activate *HVA1* gene in barley (Xue and Loveridge 2004). *HVA1* (*Hordeum vulgare* aleurone 1) is a well-known group 3 LEA protein which was previously characterized in aleurone layers, and identified as a responsive gene to drought, salinity, heat, and cold stresses (Hong et al. 1992). Another TF, namely *Hordeum vulgare* dehydration-responsive element binding protein 1 (*HvDREB1*) was characterized as an A-2 group member of the DREB family (Xu et al. 2009). Transient strong induction of this gene was observed

upon high saline concentrations (2 % salt as 60 % NaCl and 40 % Na_2SO_4) and 200 μM ABA treatment suggested a significant role of *HvDREB1* on plant stress response; however, genes associated with *HvDREB1* are still remain to be explored (Xu et al. 2009).

WRKY transcription factors in barley were identified and classified as subgroups 1–3, having the conserved promoter motif of TTGACCT, so-called W-box (Mangelsen et al. 2008). Among this family, *HvWRKY38* was reported to be expressed continuously by freezing and drought stress. *HvWRKY38* was shown to carry W-box motif and induced by dehydration after 30 min of treatment, but slightly reduced in prolonged durations of drought. Drought-response induction of this gene was shown to be ABA-independent (Rushton et al. 2010; Mare et al. 2004). In a comparative study of Tibetan hulless barley, *HvvWRKY2, HvvWRKY5, HvvWRKY19,* and *HvvWRKY46* genes were cloned from a tolerant genotype and their enhanced expression relative to a sensitive wild genotype were demonstrated under artificial drought stress (PEG-6000) (Li et al. 2014). *HvvWRKY2* was also induced by salt stress (250 mM NaCl) proving that this gene is involved in salinity tolerance in barley (Li et al. 2014).

A prominent group of genes that induced by drought stress are dehydrins (*Dhns*) in plants. *Dhn* gene family is well-characterized in barley with 13 members (Choi et al. 1999; Choi and Close 2000; Rodriguez et al. 2005). They are hydrophilic and intrinsically disordered proteins, serve as cryoprotectants in the cell during stressful conditions by yet unknown mechanisms. The interactions of dehydrins with membrane phospholipids, metal ions, water, and other unknown ligands were found to be associated with their cryoprotective functions in the plant cell (Hincha and Thalhammer 2012; Graether and Boddington 2014). Changes in dehydrin gene expression in response to drought were extensively studied in barley. Ten of the barley *Dhn* gene family (*Dhn1-11*) was shown to be up-regulated by dehydration and ABA treatment while *Dhn5* and *Dhn8* were also shown to be induced by cold stress (5 °C) (Choi et al. 1999). The expression of *Dhn13* was significantly increased upon drought, freezing, and chilling stress treatments in cv. Morex (Rodriguez et al. 2005). Similarly, after 8 h of drought stress in Tibetan hull-less barley, *Dhn13* transcript accumulations were observed, but significantly higher in drought-tolerant genotypes (TR1 and TR2) than in sensitive genotypes (TS1 and TS2) (Qian et al. 2008). Tommasini et al. (2008) proved up-regulation of the group of dehydrins including *Dhn1, Dhn2, Dhn3, Dhn4, Dhn7, Dhn9,* and *Dhn10* by drought stress and depending on the developmental stages of coleoptile, embryo, mesocotyl, and seminal roots of barley. In the same study, *Dhn6, Dhn11,* and *Dhn12* were found to be the genes that were not up-regulated upon drought or low temperature, but only expressed in barley embryo. Surprisingly, in wild barley (*H. spontaneum*), *Dhn6* showed differential expression between tolerant and sensitive genotypes and increased transcript accumulation after 12 and 24 h of dehydration (Suprunova et al. 2004). Recently, expression of *Dhn3* and *Dhn9* in flag leaves of a drought-tolerant barley (cv. Yousef) has found to be correlated with physiological traits such as chlorophyll content, osmotic adjustment, stomatal conductance, biomass, and grain yield (Karami et al. 2013). Presence or absence of *cis*-acting regulatory elements in

5′ flanking regions of dehydrins were found to be correlated with the expression patterns and most of the members of dehydrin family were shown to contain abscisic acid-responsive elements (ABRE) and dehydration-responsive elements (DRE) in their promoter regions (Choi et al. 1999; Maruyama et al. 2012). Among the *Dhn* genes, only *Dhn4, Dhn5, Dhn8,* and *Dhn10* are known to be induced by salinity stress in barley (Walia et al. 2005; Du et al. 2011).

Maintenance of ionic and oxidative homeostasis under high saline conditions is an important indication of salinity tolerance in plants (Munns and Tester 2008; Bose et al. 2014; Adem et al. 2014). There is a consensus on the idea that cellular K^+ to Na^+ ratio is a key determinant of salinity tolerance in *Arabidopsis* and monocots (Chen et al. 2007; Hauser and Horie 2010; Shabala et al. 2010; Qiu et al. 2011; Wu et al. 2013). In this context, different genotypes of barley having varying tolerance levels to salinity were studied to characterize the role of ion homeostasis components of tolerance (Chen et al. 2005; Boscari et al. 2009; Qiu et al. 2011; Roslyakova et al. 2011; Adem et al. 2014). Tolerant genotypes of barley were shown to have better control of membrane voltage (a more negative membrane), better ability to pump Na^+ from cytosol to outside, and high antioxidant capacity (Chen et al. 2007; Witzel et al. 2009). Comparative studies with three different barley cultivars showed that the most salt-tolerant genotype accumulated less Na^+ and maintained a stable K^+ concentration in the root during the salinity treatment (Adem et al. 2014). Similarly, wild *Hordeum* species were proven to have better Na^+ and Cl^- exclusion and maintained higher K^+ in the leaf compared to *H. vulgare* L. (Garthwaite et al. 2005). For example, in wild barley, K^+/Na^+ ratio were 5.2 in the leaf in response to 150 mol/m^3 salts, while it was 0.8 in *H. vulgare*. Maintenance of low Na^+ or limiting the Na^+ uptake from roots can be provided by non-selective cation channels (NSCC), Na^+/H^+ antiporters, and tonoplast-located Na^+/H^+ antiporters (Chen et al. 2007). There are several studies conducted for genetic characterization of antiporters and channels in barley. Genes coding K^+ channels, namely *HvAKT1*, *HvAKT2,* and *HvKCO1* were cloned and showed to be differentially expressed in the leaves and roots (Boscari et al. 2009). *HvAKT1* was dominantly expressed in roots whereas *HvAKT2* was expressed in leaves about 20 times higher than in roots. A high affinity *HvHAK4* K^+ transporter is expressed only in the growing leaf tissue. Salinity and K^+ treatments did not affect the expressional profile of the tested genes in a significant manner (Boscari et al. 2009). The "High-Affinity K^+ Transporter" (HKT) gene family was found to be related to salinity tolerance by regulating Na^+ transport within the plant (Huang et al. 2008; Hauser and Horie 2010; Mian et al. 2011; Benito et al. 2014). Studies with HKT family genes particularly showed their involvement in Na^+ transport by mediating Na^+ exclusion from xylem vessels in order to protect shoots from high amounts of Na^+ (Hauser and Horie 2010; Roy et al. 2014). A gene from subfamily 2, *HvHKT2;1* (also known as *HvHKT1*) showed high homology to wheat *TaHKT2;1* and its expression was dominant in roots and then in leaf blades and sheaths under normal conditions (Haro et al. 2005; Huang et al. 2008; Mian et al. 2011). Transcript levels of *HvHKT2;1* increased by 50 mM NaCl and in the absence of K^+ ions, suggest that its expression is regulated by K+ in the growth environment (Haro et al. 2005; Mian et al. 2011). Differential expressions of *HvHKT1* and

HvHKT2 genes were also demonstrated in the roots of Tibetan wild barley genotypes. *HvHKT1* was induced by salinity (150 mM NaCl) in the roots of plants nearly 50 times higher than controls, while the expression of *HvHKT2* was reduced by exposure to salinity as fivefold after 6 h (Qiu et al. 2011). Polymorphisms within the genes also explained the higher tolerance of wild accessions than cultivated barley as proven by SNP-associated studies (Qiu et al. 2011). More data is still required on expression profiles of *HKT* genes in barley in terms of cell types, growth stages, and stress conditions.

NHXs are Na^+/H^+ exchangers located on tonoplast and endosomes in addition to plasma membrane, and involved in ion and pH regulation in plant cells (Rodríguez-Rosales et al. 2009; Bassil et al. 2012). In barley, four *NHX* genes were identified for coding NHX isoforms, namely *NHX1-4*, while there are six homologous genes (*AtNHX1-6*) in *Arabidopsis* (Eckardt and Berkowitz 2011; Roslyakova et al. 2011). Overexpression of *AtNHX1* in *Arabidopsis* plants was resulted in enhanced tolerance to salinity and sustained growth under saline conditions (Apse et al. 2003). Besides ionic homeostasis, intracellular NHX antiporters are associated with significant indirect cellular functions including osmotic adjustment, vesicular traffic, cell volume regulation, and stress response. Roslyakova et al. (2011) reported that content of NHX1-3 isoforms on tonoplast increased by salinity stress (150 mM NaCl) in both leaf and root tissues. Expressions of *NHX1* and *NHX3* increased in early hours of stress while the expression of *HvNHX2* was not changed. *NHX1* expression was shown to be the same in the roots of both tolerant (cv. Elo) and sensitive cultivar (cv. Belogorskii). The comparison indicated early induction of *NHX3* expression in tolerant cultivar. As a result, authors concluded that there is a direct correlation between enhanced *NHX1* and *NHX3* expression and their protein content. In contrast, Adem et al. (2014) in their study showed that *NHX1* expression was increased twofold in a sensitive barley cultivar (cv. Naso Nijo) under salinity compared to control while there was no significant difference in other two tolerant cultivars (cvs. Numar and Golden Promise).

There are also studies reporting the expressional change of one gene in response to salinity. For example, *HvSRG6* (Stress-responsive gene 6) is characterized by differential display (DDRT-PCR) studies, which was mapped in chromosome 7H, within a region that previously was linked to osmotic adaptation in barley (Malatrasi et al. 2002). Accumulation of *HvSGR6* mRNA was shown to be relatively higher in drought-tolerant barley (*H. distichon*) cultivars upon dehydration conditions (Rapacz et al. 2010; Wojcik-Jagla et al. 2012). Upon drought, expression of *HvSRG6* was highly influenced by ABA accumulation, negatively correlated with PSII photochemical activity, and reduced by light, emphasizing the interactions between these factors and water deficit (Malatrasi et al. 2002; Rapacz et al. 2010; Wojcik-Jagla et al. 2012). In contrast, *HvSGR6* expression was not found significantly related to ABA treatment in some barley genotypes (No.62-Unitan, No.3-Keystone, No.53 GK Rezi, No. 73a Compana, No.8 Hazen) (Cseri et al. 2011).

Regulations against oxidative stress and osmotic capacity were reported as important factors for higher salinity tolerance of barley compared to other cereals (Witzel et al. 2009; Roy et al. 2014; Adem et al. 2014). However, information on

transcript levels of genes associated with oxidative and osmotic pathways is still limited for different experimental models. Adem et al. (2014) recently reported that transcriptional changes of antioxidant genes in one specific time point were not informative for evaluating salinity tolerance in contrasting genotypes. Comparison of dehydration treatments also indicated that barley responds differently upon rapid- or slow-developing water stress with respect to antioxidant mechanisms (Talame et al. 2007). Similarly, a study comparing the changes in antioxidant mechanism genes between contrasting barley cultivars (tolerant cv. Martı and sensitive cv. Erginel90) subjected to shock-like and slow-developing dehydration demonstrated a rapid induction of a 2-Cys peroxiredoxin gene, *HvBAS1,* and metallothionein-like protein type 2 (*HvMT2)* in both cultivars upon shock-like dehydration while their expression was relatively low in slow-developing water stress (Fig. 1) (Gürel et al. 2016).

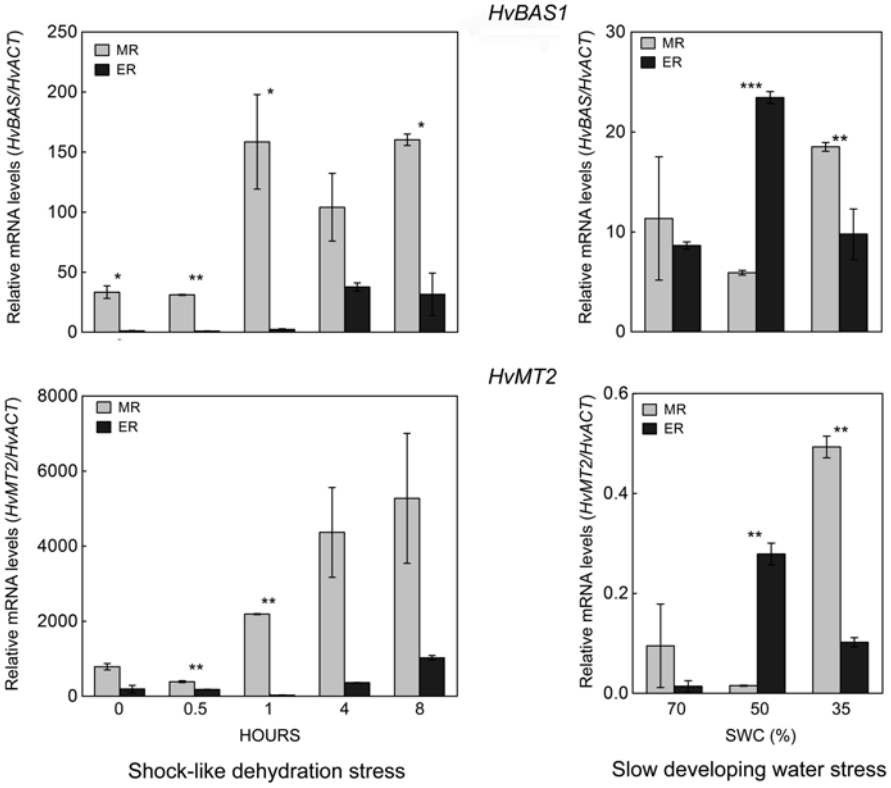


Fig. 1 Relative transcription levels of water stress-responsive genes, *HvBAS1* and *HvMT2* in barley cultivars cv. Martı (MR) and cv. Erginel90 (ER) under shock-like dehydration and slow-developing water stress. Transcript levels were measured by qRT-PCR and normalized to *HvACT*. Asterisks indicate statistical significance at * $P \leq 0.05$, ** $P \leq 0.01$, and *** $P \leq 0.001$ of the differences between cultivars according to *t*-test

Although there are a number of studies on expressional changes of stress-responsive or tolerance-related genes, the functions of the majority of the genes proven to have altered expression upon stress treatments still remain unknown and there are probably more genes to be discovered.

6 Combined Effects of Drought and Salinity on Transcriptome

Drought and salinity stress combination in nature is the main factor for yield loss in agriculture (Ahmed et al. 2013a). Cellular osmotic stress causing abiotic stress (drought, salinity, heat, cold) response pathways have common components (Hu et al. 2010; Dolferus 2014). The effect of low water and high salt in cellular level especially is quite similar, since both cause reduction in water potential and reprogramming of the metabolism to reduce the adverse effects of dehydration due to low water potential (Ahmed et al. 2013a). Most of the studies on transcriptomal changes, on the other hand, were mainly on the effect of drought or salt stress alone. The combination of stresses is obviously more detrimental to plant and, therefore, the understanding of abiotic stress on plant metabolism should be more focused on the different combinations of these stresses.

The literature on combined effects of abiotic stresses on barley, especially drought and salinity, is quite limited and mainly based on the physiological comparison of dehydration on barley genotypes rather than the changes in gene expression profile (Ahmed et al. 2013b). One outstanding study was published by Ozturk et al. (2002). They used the in-house printed cDNA array consisting 6 and 10 h shock-drought stressed leaf and root DNA elements for hybridization with 24 h 150 mM NaCl stressed 3-week-old barley plants (cv. Tokak). Only 5 % of the 1463 unique transcripts were responsive to salt stress; induction in, for example, metallothionein-like protein type 2, heat shock protein DnaJ homolog Pfj4, allene oxide synthase, glutathione S-transferase, ubiquitin; and reduction in auxin-induced protein, wheat aluminum-induced protein wali5, late embryogenesis abundant protein LEA14-A, and sucrose synthase. There were only a number of transcripts showing co-regulation in response to both drought and salt stresses; allene oxide synthase 1, metallothionein-like protein, ERD1 protein precursor, Δ 1-pyrroline-5-carboxylate synthetase and germin-like protein up-regulated in response to both stresses, and α-amylase, lipid transfer protein cw18, ABC transporter, vacuolar processing enzyme precursor, and an unknown protein from rice were down-regulated.

It is clear that there should be more studies on dose- and time-dependent responses of barley genotypes to drought, salt, and drought and salt combined stress treatments to understand the differential changes on metabolism depending on the these factors (Dolferus et al. 2011; De Mezer et al. 2014). Such studies are also required to create crop interactome maps which will enable researches to analyze

association between phenotypic variation and environmental stress tolerance (Skirycz and Inze 2010; Mochida et al. 2011; Shelden and Roessner 2013; Shanker et al. 2014).

7 Transgenic Approaches for Drought and Salinity Tolerance

It is expected that the human population will reach nine billion by 2050 and the increase in food production should be at least 70 % to provide the demand (Godfray et al. 2010). The crops comprise about 80 % of human food, while the cereals supply the half of the global food production (Langridge and Fleury 2011). In order to meet the growing demand, crop productivity should be increased to 44 million tones (Mt) per year; however, annual increase in global food production was calculated to be around 32 Mt on 2010 and, with changing global climate, there will be an increasing gap between food production and consumption (Tester and Langridge 2010). The main yield loss in agricultural production occurs due to abiotic stresses and in order to meet the growing demand for food, the harmful effects of stress factors on crop yield should be minimized. The best approach is the development of crop varieties with enhanced tolerance to environmental stress conditions either by conventional breeding, or molecular breeding and genetic engineering, or combination of both. There are a number of studies using the identified stress-responsive genes in improvement of barley genotypes to drought and salt stresses using genetic engineering (Table 2).

Transcription factors (TFs) control the expression of many genes, and overexpression of TF genes can help plants to better tolerate abiotic stress (Kasuga et al. 1999; Cominelli et al. 2013). Therefore, transcription factors have always been attractive targets for conferring abiotic stress tolerance to plants. A promising example is the family of drought-responsive element binding proteins (DREBs)/C-repeat binding factors (CBFs). The transgenic overexpression of *HvCBF4* in rice resulted in an increase in tolerance to drought, high salinity and low temperature stresses without growth retardation (Oh et al. 2007). *HsDREB1A* transcription factor isolated from wild barley was cloned to downstream of stress-inducible *HVA1s* promoter and introduced into the apomictic bahiagrass (*Paspalum notatum* Flugge) cultivar cv. Argentine by particle bombardment (James et al. 2008). Overexpression of this gene in bahiagrass plants improved survival and biomass under severe dehydration and salt stress in contrast to control plants. In another study, transformation of *Arabidopsis* with barley *HvDREB1* gene enhanced salinity and drought tolerance (Xu et al. 2009). Overexpression of *HvDREB1* driven by 35S promoter of cauliflower mosaic virus (CaMV) led to the accumulation of transcripts corresponding to *RD29A* in the transgenic *Arabidopsis* plants in addition to its own mRNA. The transgenic plants displayed less severe inhibition of root growth than non-transgenic plants during 100 mM NaCl treatment. Two DREB/CBFs (*TaDREB2* and *TaDREB3*)

Table 2 Summary of transformation studies to improve salinity and drought tolerance using genes encoded by barley and other plant species

Transgene	Origin of the gene	Transformed plant	Performance of transgenic plant	Reference
HvCBF4	Barley	Rice	Drought, salinity, and cold tolerance	Oh et al. (2007)
HsDREB1A	*H. spontaneum*	Bahiagrass plants	Salinity and drought tolerance	James et al. (2008)
HvDREB1	Barley	*Arabidopsis*	Salinity tolerance	Xu et al. (2009)
TaDREB2 and TaDREB3	Wheat	Barley	Drought and frost tolerance	Morran et al. (2011)
HvRAF	Barley	*Arabidopsis*	Salinity and pathogen tolerance	Jung et al. (2007)
HvSNAC1	Barley	Barley	Drought tolerance	Al Abdallat et al. (2014)
AtCIPK16	*Arabidopsis*	Barley	Salinity tolerance	Roy et al. (2013)
HvHVA1	Barley	Rice	Drought tolerance	Babu et al. (2004)
HvHVA1	Barley	Mulberry	Salinity, drought and cold tolerance	Checker et al. (2012)
HvAPX	Barley	*Arabidopsis*	Salinity tolerance	Xu et al. (2008)
HvHKT2;1	Barley	Barley	Salinity tolerance	Mian et al. (2011)
AtAVP1	*Arabidopsis*	Barley	Salinity tolerance	Schilling et al. (2014)

from wheat were used to modulate the stress tolerance in barley and wheat (Morran et al. 2011). Transgenic barley plants constitutively expressing *TaDREB2* and *TaDREB3* genes showed enhanced survival under severe drought conditions but growth retardation and late flowering with lowered grain yield were observed in these transgenic plants (Morran et al. 2011). Alternatively, the growth retardation was eliminated by using an inducible promoter of a maize gene responsive to abscisic acid (*ZmRAB17*).

A barley *HvRAF* (*Hordeum vulgare* root abundant factor) gene encoding an ethylene response factor-type transcription factor was cloned from young seedlings of barley and characterized for its functional importance in biotic and abiotic stress responses (Jung et al. 2007). This gene was also transferred to *Arabidopsis* by floral-dip method to monitor expression patterns of several stress-responsive genes such as *Arabidopsis* putative fungal protein (*PDF1.2*), a serine/threonine protein kinase (*KIN2*), and thaumatin-like protein (*PR5*) (Jung et al. 2007). Overexpression of the full-length *HvRAF* gene in *Arabidopsis* plants led to improved resistance to bacterial soil-borne pathogen *Ralstonia solanacearum* strain GMI1000 and enhanced salinity tolerance. Recently, a gene from *NAC* TF family, namely *HvSNAC1* was

also isolated from drought-stressed barley plants and used for transgenic approach (Al Abdallat et al. 2014). The gene was introduced to barley (cv. Golden Promise) plants via *Agrobacterium tumefaciens* under the control of maize ubiquitin (*Ubi*) promoter (Al Abdallat et al. 2014). Transgenic plants constitutively expressed *HvSNAC1* showed higher drought tolerance with improved productivity and grain yield relative to control plants.

Calcium (Ca^{2+}) serves as a ubiquitous second messenger and regarded as one of the most important molecule involved in plant signaling networks (Yu et al. 2014). Ca^{2+} signaling process is assumed to be one of the earliest events in salt signaling, and may play an essential role in the ion homeostasis leading to salt tolerance in plants (Zhu 2003; Reddy and Reddy 2004). Calcineurin B-like-interacting protein kinases (CIPKs) have important functions in the Ca^{2+} regulated signaling pathways for controlling plant responses to abiotic stresses and regulation of ion homeostasis (Weinl and Kudla 2009; Kudla et al. 2010). Transgenic barley plants overexpressing *AtCIPK16* were shown to have increased salinity tolerance, which was expressed as 20–45 % higher maintenance in biomass under exposure to long-term strong salinity stress (30 days, 300 mM NaCl) compared with control plants (Roy et al. 2013).

Another approach for developing transgenic plants with higher tolerance to salinity and drought stress is transferring single genes that encode functional and structural proteins such as LEA proteins, molecular chaperones, ion transporters, proteins involved in lipid biosynthesis, detoxification enzymes, and key enzymes for osmolyte biosynthesis (Cominelli et al. 2013). Overexpression of the barley *HvHVA1* gene (group 3 LEA protein) under the constitutive control of rice *Act1* promoter led to a significant accumulation of the HVA1 protein in the roots and leaves of transgenic rice plants (Babu et al. 2004). After transformation, the leaf relative water content (RWC) of transgenic plants was higher than control plants. In addition, the transgenic plants showed less reduction in plant growth under drought stress (Babu et al. 2004). The *HvHVA1* gene under a constitutive promoter (*Act1*) was also introduced to mulberry plants and overexpression of the gene conferred tolerance to salt and cold stresses as well as drought in transgenic plants with retardation on plant growth and productivity (Checker et al. 2012). In order to eliminate these effects, authors used another recombinant vector (*pBI121:rd29A:Hva1*), which was constructed under the control of the stress-inducible *Arabidopsis rd29A* promoter. The *rd29A*-regulated transgene expression was also conferred enhanced tolerance against drought, salt, and cold stress. The performance of transgenic mulberry plants against these multi-stress factors was evaluated in the field conditions. *Rd29A:Hva1* plants showed better performance in field conditions upon mild stress than *Act1:Hva1* plants, which performed better in severe stress conditions.

A number of transgenic improvements for abiotic stress tolerance have been achieved by overexpression of detoxifying genes such as glutathione peroxidase (GPX), ascorbate peroxidase (APX), and superoxide dismutase (SOD) (Roxas et al. 1997; Wang et al. 2005; Xu et al. 2008). For example, the transgenic *Arabidopsis* plants overexpressing barley *HvAPX1* gene under the control of the CaMV 35S promoter were shown to enhance salt tolerance compared to wild-type plants (Xu et al. 2008). However, Na^+, K^+, Ca^{2+}, and Mg^{2+} contents of transgenic plants were

not found to be significantly different from wild-type plants. In addition, H_2O_2 content and lipid peroxidation level (determined by measuring the amount of malondialdehyde) upon salt stress were lower in the transgenic plants than in the wild type. As a result, it was concluded that improvement in response to salt stress was achieved by reduction in oxidative stress injury instead of maintenance of cellular ion homeostasis (Xu et al. 2008).

Transporter proteins are important candidates in genetic engineering for enhancing salt and dehydration tolerance in plants. Transgenic barley plants overexpressing a HKT transporter *HvHKT2;1* showed higher growth rates than non-transgenic plants by approximately 25–30 % in response to 50 mM NaCl and 100 mM NCl treatments supplemented with 2 mM K^+ (Mian et al. 2011).

Finally, successful transformations were performed in transgenic plants expressing genes for enzymes involved in proton (H^+) pumps that generate energy for tonoplast transport of solutes into vacuoles. Previously, *Arabidopsis* proton pumping pyrophosphatases (H^+-PPase; *AVP1*) were transferred to different plant species to confer salinity tolerances as reviewed by Agarwal et al. (2013). Very recently, the improvement of salinity tolerance of transgenic barley (cv. Golden Promise) by overexpression of *AtAVP1* was confirmed in saline field trials in addition to assessments in greenhouse conditions (Schilling et al. 2014). The transgenic barley plants were generated via *Agrobacterium*-mediated transformation using a construct including *AtAVP1* under the control of the CaMV 35S promoter (Schilling et al. 2014). In the low-salinity area where the soil electrical conductivity (SEC) was 161 ± 11 µS/cm, the transgenic barley plants had a significantly higher (17–33 %) shoot biomass and higher (23–34 %) grain yield per plant compared to non-transgenic plants and in the high-salinity area (SEC = 1231 ± 155 µS/cm), transgenic barley had higher shoot biomass (30–42 %) and showed improved survival compared to non-transgenic plants. Furthermore, in the high-salinity area, the grain yield per plant of the transgenic barley was significantly higher (79–87 %) than that of non-transgenic plants with the higher number of heads and grains.

In conclusion, several dehydration and salinity stress-related genes have been isolated and characterized in barley and transferred to different plant species to test their efficiency in tolerance mechanism. Equally, genes from wheat and *Arabidopsis* were also used to obtain transgenic barley plants (Table 2). In most cases, the performance of the transgenic plants was evaluated under greenhouse or controlled laboratory conditions. Recently, generated transgenic lines are more frequently being tested under natural field conditions to observe phenotypic responses to single or combination of stress factors. For instance, long-term field testing of transgenic barley for higher drought tolerance is known to be carried out in Australia (Mrízová et al. 2014).

8 Conclusion and Future Perspective

A better understanding of the genes in model plants with natural tolerance to water deficiency and salinity is becoming a priority due to more frequently experienced drought episodes in certain areas and global climatic changes. With its natural tolerance to several abiotic and biotic stress factors in addition to its small genome size, barley was proposed to be an excellent model system for abiotic stress research in cereals.

Transcriptomics which allows deep functional characterization of stress-inducible or stress-associated genes in plants is an important step in omics technologies. The studies using different plant systems and different stress conditions indicated that cellular signalling and metabolic mechanisms like osmotic adjustment, transcriptional regulation, antioxidant system, and single genes such as antiporters, LEA proteins, and ABA-inducible proteins are activated by water deficit and salinity stresses and contribute to survival of plants in stressful conditions. Transcriptomics was also effectively used in barley. A large amount of data was obtained by high-throughput transcriptomic technologies (microarrays and RNA-seq) and used in the identification of differentially regulated genes upon water deficit and salinity in barley (Ozturk et al. 2002; Ueda et al. 2004; Walia et al. 2006; Gao et al. 2013; Ziemann et al. 2013). New methods are still needed for interpretation of such large-scale data, for its incorporation to prior biological knowledge, and for the establishment of the relationships among variables (genes and metabolites) more precisely (Reshetova et al. 2014). Small-scale expression studies, on the other hand, focus on investigating more detailed analyses of gene inductions and supply useful data, particularly in comparison to wild-type and contrasting genotypes differing in their tolerance to drought or salinity.

Breeding is still the most popular approach to generate abiotic stress-tolerant crops. Selection of breeding lines by physiological comparison in response to drought and salt in the field conditions is still not easy and time-consuming, although there are new developments such as high-throughput phenotyping with global positioning systems, meteorological devices and so-called phenomobiles, phenotowers, and thermal imaging sensors (Honsdorf et al. 2013; Araus and Cairns 2014; Chen et al. 2014). Therefore, studies on changes in gene expression patterns in response to drought and salt stress conditions are extremely important not only to understand the basis of tolerance but also to generate molecular markers for the selection of tolerant genotypes in laboratory (Forster et al. 2000). Transgenic approach is an attractive option to develop salt and drought-tolerant crops for near future. In this context, transcription factors (e.g. *HsDREB1A, HvSNAC1*) and a few single genes (*HvHVA1*, *AtAVP1* etc.) cloned from barley and *Arabidopsis were* demonstrated to be efficient to generate plants with increased drought and salt tolerance.

With the large transcriptomic data and resources, barley is being exploited to identify genetic determinants that underlie its high tolerance to abiotic stresses. Wild progenitor of cultivated barley *H. spontaneum* is known to adapt to diverse environments including deserts and cold regions like Tibet (Nevo and Chen 2010;

Zhao et al. 2010), and therefore, is a promising genetic resource for abiotic stress tolerance improvement. Current research indicates the role of epigenetic mechanisms for abiotic stress tolerance and long-term adaptation to stressful conditions. It is now known that in addition to the known genetic determinants, epigenetic mechanisms including the dynamic changes in chromatin and synthesis of small RNAs also contribute to the regulation of gene expression during the stress responses (Mirouze and Paszkowski 2011). There are only two studies with barley in this scope. Papaefthimiou and Tsaftaris (2012) identified a putative jumonji-like histone demethylase, namely *HvPKDM7-1* in barley and showed its up-regulation under drought. Another barley gene coding a putative HvDME protein related to cytosine methylation was highly induced by dehydration stress (Kapazoglou et al. 2013). By having a larger genome than *Arabidopsis* and rice, transcriptomic information from barley may lead to complete resolution of molecular and biochemical networks related to stress response, and may also provide new approaches for the generation of transgenic crops with higher abiotic stress tolerance.

Acknowledgment This research was supported by Scientific Research Projects Coordination Unit of Istanbul University, project BAP 4712.

References

Abebe T, Melmaiee K, Berg V, Wise RP (2010) Drought response in the spikes of barley: gene expression in the lemma, palea, awn, and seed. Funct Integr Genomics 10:191–205

Adem GD, Roy SJ, Zhou M, Bowman JP, Shabala S (2014) Evaluating contribution of ionic, osmotic and oxidative stress components towards salinity tolerance in barley. BMC Plant Biol 14:113

Agarwal PK, Shukla PS, Gupta K, Jha B (2013) Bioengineering for salinity tolerance in plants: state of the art. Mol Biotechnol 54:102–123

Ahmed IM, Cao F, Zhang M, Chen X, Zhang G, Wu F (2013a) Difference in yield and physiological features in response to drought and salinity combined stress during anthesis in Tibetan wild and cultivated barleys. PLoS One 8:e77869

Ahmed IM, Dai H, Zheng W, Cao F, Zhang G, Sun D, Wu F (2013b) Genotypic differences in physiological characteristics in the tolerance to drought and salinity combined stress between Tibetan wild and cultivated barley. Plant Physiol Biochem 63:49–60

Al Abdallat AM, Ayad JY, Abu Elenein JM, Al Ajlouni Z, Harwood WA (2014) Overexpression of the transcription factor *HvSNAC1* improves drought tolerance in barley (*Hordeum vulgare* L.). Mol Breed 33:401–414

Allardyce JA, Rookes JE, Hussain HI, Cahill DM (2013) Transcriptional profiling of *Zea mays* roots reveals roles for jasmonic acid and terpenoids in resistance against *Phytophthora cinnamomi*. Funct Integr Genomics 13:217–228

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106

Apse MP, Sottosanto JB, Blumwald E (2003) Vacuolar cation/H^+ exchange, ion homeostasis, and leaf development are altered in a T-DNA insertional mutant of *AtNHX1,* the *Arabidopsis* vacuolar Na^+/H^+ antiporter. Plant J 36:229–239

Araus JL, Cairns JE (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci 19:52–61

Ashraf M, Foolad MR (2007) Roles of glycine betaine and proline in improving plant abiotic stress resistance. Environ Exp Bot 59:206–216

Atienza SG, Faccioli P, Perrotta G et al (2004) Large scale analysis of transcripts abundance in barley subjected to several single and combined abiotic stress conditions. Plant Sci 167:1359–1365

Babu RC, Zhang J, Blum A, David Ho T-H, Wu R, Nguyen HT (2004) *HVA1*, a LEA gene from barley confers dehydration tolerance in transgenic rice (*Oryza sativa* L.) via cell membrane protection. Plant Sci 166:855–862

Badr A, Muller K, Schafer-Pregl R, El Rabey H, Effgen S, Ibrahim HH, Pozzi C, Rohde W, Salamini F (2000) On the origin and domestication history of barley. Mol Biol Evol 17:499–510

Bassil E, Coku A, Blumwald E (2012) Cellular ion homeostasis: emerging roles of intracellular NHX Na^+/H^+ antiporters in plant growth and development. J Exp Bot 63:5727–5740

Bedada G, Westerbergh A, Müller T, Galkin E, Bdolach E, Moshelion M, Fridman E, Schmidet KJ (2014) Transcriptome sequencing of two wild barley (*Hordeum spontaneum* L.) ecotypes differentially adapted to drought stress reveals ecotype-specific transcripts. BMC Genomics 15:1–20

Benito B, Haro R, Amtmann A, Cuin TA, Dreyer I (2014) The twins K^+ and Na^+ in plants. J Plant Physiol 171:723–731

Bhargava S, Sawant K (2013) Drought stress adaptation: metabolic adjustment and regulation of gene expression. Plant Breed 132:21–32

Bonman JM, Bockelman HE, Jackson LF, Steffenson BJ (2005) Disease and insect resistance in cultivated barley accessions from the USDA National Small Grains Collection. Crop Sci 45:1271–1280

Boscari A, Clément M, Volkov V, Golldack D, Hybiak J, Miller AJ, Amtmann A, Fricke W (2009) Potassium channels in barley: cloning, functional characterization and expression analyses in relation to leaf growth and development. Plant Cell Environ 32:1761–1777

Bose J, Rodrigo-Moreno A, Shabala S (2014) ROS homeostasis in halophytes in the context of salinity stress tolerance. J Exp Bot 65:1241–1257

Bot A, Benites J (2005) The importance of soil organic matter, key to drought-resistant soil and sustained food production. FAO Soils Bulletin, Rome

Camilios-Neto D, Bonato P, Wassem R, Brusamarello-Santos LCC, Valdameri G, Donatti L, Faoro H, Weiss VA, Chubatsu LS, OPedrosa F, Souzaet EM (2014) Dual RNA-seq transcriptional analysis of wheat roots colonized by *Azospirillum brasilense* reveals up-regulation of nutrient acquisition and cell cycle genes. BMC Genomics 15:378

Cattivelli L, Rizza F, Badeck FW, Mazzucotelli E, Mastrangelo AM, Francia E, Marè C, Tondelli A, Stanca AM (2008) Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. Field Crops Res 105:1–14

Checker VG, Chhibbar AK, Khurana P (2012) Stress-inducible expression of barley *Hva1* gene in transgenic mulberry displays enhanced tolerance against drought, salinity and cold stress. Transgenic Res 21:939–957

Chen Z, Newman I, Zhou M et al (2005) Screening plants for salt tolerance by measuring K^+ flux: a case study for barley. Plant Cell Environ 28:1230–1246

Chen Z, Pottosin II, Cuin TA, Fuglsang AT, Tester M, Jha D, Zepeda-Jazo I, Zhou M, Palmgren MG, Newman IA, Shabala S (2007) Root plasma membrane transporters controlling K^+/Na^+ homeostasis in salt-stressed barley. Plant Physiol 145:1714–1725

Chen D, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C (2014) Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. Plant Cell 26:4636–4655

Choi DW, Close TJ (2000) A newly identified barley gene, Dhn12, encoding a YSK2 DHN, is located on chromosome 6H and has embryo-specific expression. Theor Appl Genet 100:1274–1278

Choi DW, Zhu B, Close TJ (1999) The barley (*Hordeum vulgare* L.) dehydrin multigene family: sequences, allele types, chromosome assignments, and expression characteristics of 11 *Dhn* genes of cv Dicktoo. Theor Appl Genet 98:1234–1247

Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP (2004) A new resource for cereal genomics: 22K Barley GeneChip comes of age. Plant Physiol 134:960–968

Close TJ et al (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Cominelli E, Conti L, Tonelli C, Galbiati M (2013) Challenges and perspectives to improve crop drought and salinity tolerance. N Biotechnol 30:355–361

Cseri A, Cserháti M, von Korff M, Nagy B, Horvath GB, Palagyi A, Pauk J, Dudits D, Törjek O (2011) Allele mining and haplotype discovery in barley candidate genes for drought tolerance. Euphytica 181:341–356

Dai A (2011) Drought under global warming: a review. Wiley Interdiscip Rev Clim Chang 2:45–65

Dai A (2013) Increasing drought under global warming in observations and models. Nat Clim Chang 3:52–58

De Mezer M, Turska-Taraska A, Kaczmarek Z, Glowacka K, Swarcewicz B, Rorat T (2014) Differential physiological and molecular response of barley genotypes to water deficit. Plant Physiol Biochem 80:234–248

Diab AA, Teulat-Merah B, This D, Ozturk NZ, Bensher D, Sorrells ME (2004) Identification of drought-inducible genes and differentially expressed sequence tags in barley. Theor Appl Genet 109:1417–1425

Dolezel J, Greilhuber J, Lucretii S, Meister A, Lysak MA, Nardi L, Obermayer R (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. Ann Bot 82:17–26

Dolferus R (2014) To grow and not to grow: a stressful decision for plants. Plant Sci 229:247–261

Dolferus R, Ji X, Richards RA (2011) Abiotic stress and control of grain number in cereals. Plant Sci 181:331–341

Du JB, Yuan S, Chen YE, Sun X, Zhang ZW, Xu F, Yuan M, Shang J, Lin HH (2011) Comparative expression analysis of dehydrins between two barley varieties, wild barley and Tibetan hulless barley associated with different stress resistance. Acta Physiol Plant 33:567–574

Eckardt NA, Berkowitz GA (2011) Functional analysis of *Arabidopsis* NHX antiporters: the role of the vacuole in cellular turgor and growth. Plant Cell 23:3087–3088

Edwards KD, Bombarely A, Story GW, Allen F, Mueller LA, Coates SA, Jones L (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. BMC Genomics 11:142

Forster BP, Ellis RP, Thomas WT, Newton AC, Tuberose R, This D, el-Enein RA, Bahri MH, Ben Salem M (2000) The development and application of molecular markers for abiotic stress tolerance in barley. J Exp Bot 51:19–27

Gao R, Duan K, Guo G, Du Z, Chen Z, Li L, He T, Lu R, Huang J (2013) Comparative transcriptional profiling of two contrasting barley genotypes under salinity stress during the seedling stage. Int J Genomics 2013:1–19

Garthwaite AJ, von Bothmer R, Colmer TD (2005) Salt tolerance in wild *Hordeum* species is associated with restricted entry of Na^+ and Cl^- into the shoots. J Exp Bot 56:2365–2378

Godfray HC, Beddington JR, Crute IR (2010) Food security: the challenge of feeding 9 billion people. Science 327:812–818

Godfree RC (2012) The impacts of extreme drought and climate change on plant population dynamics and evolution. In: Neves DF, Sanz JD (eds) Droughts: new research. Nova, New York, pp 189–214

Gosal SS, Wani SH, Kang MS (2009) Biotechnology and drought tolerance. J Crop Improv 23:19–54

Graether SP, Boddington KF (2014) Disorder and function: a review of the dehydrin protein family. Front Plant Sci 5:576

Guo P, Baum M, Grando S, Ceccarelli S, Bai G, Li R, von Korff N, Varshney RK, Graner A, Valkonun J (2009) Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. J Exp Bot 60:3531–3544

Gürel F, Öztürk NZ, Yörük E, Uçarlı C, Poyraz N (2016) Comparison of expression patterns of selected drought-responsive genes in barley (Hordeum vulgare L.) under shock-dehydration and slow drought treatments. Plant Growth Regul 1–11. doi:10.1007/s10725-016-0156-0

Haro R, Banuelos MA, Senn ME, Barrrero-Gil J, Rodriguez-Navarro A (2005) HKT1 mediates sodium uniport in roots Pitfalls in the expression of *HKT1* in yeast. Plant Physiol 139:1495–1506

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

Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486

Hayes PM, Castro A, Marquez-Cedillo L, Corey A (2003) Genetic diversity for quantitatively inherited agronomic and malting quality traits. In: von Bothmer R, van Hintum T, Knüpffer H, Sato K (eds) Diversity in barley (*Hordeum vulgare*) developments in plant genetics and breeding. Elsevier Science, Amsterdam, pp 201–226

Hincha DK, Thalhammer A (2012) LEA proteins: IDPs with versatile functions in cellular dehydration tolerance. Biochem Soc Trans 40:1000–1003

Hoagland DR, Arnon DI (1950) The water culture method for growing plants without soil. University of California Agricultural Experiment Station, Berkley, Circular 347

Hong B, Barg R, Ho T-HD (1992) Developmental and organ-specific expression of an ABA- and stress-induced protein in barley. Plant Mol Biol 18:663–674

Honsdorf N, March TJ, Berger B, Tester M, Pillen K (2013) High-throughput phenotyping to detect drought tolerance QTL in wild barley introgression lines. PLoS One 9:e97047

Hu XJ, Zhang ZB, Xu P (2010) Multifunctional genes: the cross-talk among the regulation networks of abiotic stress responses. Biol Plant 54:213–223

Huang S, Spielmeyer W, Lagudah ES, Munns R (2008) Comparative mapping of *HKT* genes in wheat, barley, and rice, key determinants of Na^+ transport, and salt tolerance. J Exp Bot 59:927–937

Jain M (2011) A next-generation approach to the characterization of a non-model plant transcriptome. Curr Sci 101:1435–1439

James VA, Neibaur I, Altpeter F (2008) Stress inducible expression of the DREB1A transcription factor from xeric, Hordeum spontaneum L. in turf and forage grass (Paspalum notatum Flugge) enhances abiotic stress tolerance. Transgenic Res 17:93–104

Jung J, Won SY, Suh SC, Kim H, Wing R, Jeong Y, Hwang I, Kim M (2007) The barley ERF-type transcription factor HvRAF confers enhanced pathogen resistance and salt tolerance in *Arabidopsis*. Planta 225:575–588

Kapazoglou A, Drosou V, Argiriou A, Tsaftaris AS (2013) The study of a barley epigenetic regulator, HvDME, in seed development and under drought. BMC Plant Biol 13:172

Karami A, Shahbazi M, Niknam V, Shobbar ZS, Tafreshi RS, Abedini R, Mabood HE (2013) Expression analysis of dehydrin multigene family across tolerant and susceptible barley (*Hordeum vulgare* L.) genotypes in response to terminal drought stress. Acta Physiol Plant 35:2289–2297

Kasuga M, Liu Q, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1999) Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. Nat Biotechnol 17:287–291

Kavi Kishor PB, Hong Z, Miao GH, Hu CAA, Verma DPS (1995) Over-expression of-pyrroline-5-carboxylate synthetase increases proline production and confers osmotolerance in transgenic plants. Plant Physiol 25:1387–1394

Kilian B, Martin W, Salamini F (2010) Genetic diversity, evolution and domestication of wheat and barley in the Fertile Crescent. In: Glaubrecht M (ed) Evolution in action. Springer, Berlin, pp 137–166

Kilian J, Peschke F, Berendzen KW, Harter K, Wanke D (2012) Prerequisites, performance and profits of transcriptional profiling the abiotic stress response. Biochim Biophys Acta 1819:166–175

Knüpffer H, Terentyeva I, Hammer K et al (2003) Ecogeographical diversity—a Vavilovian approach. In: von Bothmer R, van Hintum T, Knüpffer H, Sato K (eds) Diversity in barley (*Hordeum vulgare*). Elsevier, Amsterdam, pp 53–76

Kudla J, Batistic O, Hashimoto K (2010) Calcium signals: the lead currency of plant information processing. Plant Cell 22:541–563

Kumar S, Wang Z, Banks TW, Jordan MC, McCallum BD, Cloutier S (2014) *Lr1*-mediated leaf rust resistance pathways of transgenic wheat lines revealed by a gene expression study using the Affymetrix GeneChip® Wheat Genome Array. Mol Breed 34:127–141

Langridge P, Fleury D (2011) Making the most of "omics" for crop breeding. Trends Biotechnol 29:33–40

Läuchli A, Grattan SR (2007) Plant growth and development under salinity stress. In: Jenks MA, Hasegawa PM, Jain SM (eds) Advances in molecular breeding toward drought and salt tolerant crops. Springer, Dordrecht, pp 285–315

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Guo Q, Lan X, Zhou Q, Wei N (2014) Comparative expression analysis of five *WRKY* genes from Tibetan hulless barley under various abiotic stresses between drought-resistant and sensitive genotype. Acta Physiol Plant 36:963–973

Malatrasi M, Close TJ, Marmirolli N (2002) Identification and mapping of putative stress response regulator gene in barley. Plant Mol Biol 50:143–152

Mangelsen E, Kilian J, Berendzen KW, Kolukisaoglu UH, Harter K, Jansson C, Wanke D (2008) Phylogenetic and comparative gene expression analysis of barley (*Hordeum vulgare*) WRKY transcription factor family reveals putatively retained functions between monocots and dicots. BMC Genomics 9:194

Mare C, Mazzucotelli E, Crosatti C, Francia E, Stanca AM, Cattivelli L (2004) Hv-WRKY38: a new transcription factor involved in cold- and drought-response in barley. Plant Mol Biol 55:399–416

Maruyama K, Todaka D, Mizoi J, Yoshida T, Kidokoro S, Matsukura S, Takasaki H, Sakurai T, Yamamoto YY, Yoshiwara K, Kojima M, Sakakibara H, Shinozaki K, Yamaguchi-Shinozaki K (2012) Identification of cis-acting promoter elements in cold- and dehydration- induced transcriptional pathways in *Arabidopsis*, rice, and soybean. DNA Res 19:37–49

Mayer KFX et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263

Mian A, Oomen RJFJ, Isayenkov S, Sentenac H, Maathuis FJM, Very AA (2011) Over-expression of an Na^+-and K^+-permeable HKT transporter in barley improves salt tolerance. Plant J 68:468–479

Miller GAD, Suzuki N, Ciftci-Yilmaz S et al (2010) Reactive oxygen species homeostasis and signalling during drought and salinity stresses. Plant Cell Environ 33:453–467

Mir RR, Zaman-Allah M, Sreenivasulu N, Trethowan R (2012) Integrated genomics, physiology and breeding approaches for improving drought tolerance in crops. Theor Appl Genet 125:625–645

Mirouze M, Paszkowski J (2011) Epigenetic contribution to stress adaptation in plants. Curr Opin Plant Biol 14:267–274

Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta 1819:86–96

Mochida K, Uehara-Yamaguchi Y, Yoshida T, Sakurai T, Shinozaki K (2011) Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. Plant Cell Physiol 52:785–803

Morran S, Eini O, Pyvovarenko T, Parent B, Singh R, Ismagul A, Eliby S, Shirley N, Langridge P, Lopato S (2011) Improvement of stress tolerance of wheat and barley by modulation of expression of DREB/CBF factors. Plant Biotechnol J 9:230–249

Morrell PL, Clegg MT (2007) Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. Proc Natl Acad Sci U S A 104:3289–3294

Morrell PL, Gonzales AM, Meyer KK, Clegg MT (2014) Resequencing data indicate a modest effect of domestication on diversity in barley: a cultigen with multiple origins. J Hered 105:253–264

Mrízová K, Holasková E, Öz MT, Jiskrová E, Frébort I, Galuszka P (2014) Transgenic barley: a prospective tool for biotechnology and agriculture. Biotechnol Adv 32:137–157

Munns R, Tester M (2008) Mechanisms of salinity tolerance. Annu Rev Plant Biol 59:651–681

Nevo E (1992) Origin, evolution, population genetics and resources for breeding of wild barley, *Hordeum spontaneum*, in the Fertile Crescent. In: Shewry PR (ed) Barley: genetics, biochemistry, molecular biology and biotechnology. CAB International, Wallingford, UK, pp 19–43

Nevo E, Chen G (2010) Drought and salt tolerances in wild relatives for wheat and barley improvement. Plant Cell Environ 33:670–685

Oh SJ, Kwon CW, Choi DW, Song SI, Kim JK (2007) Expression of barley *HvCBF4* enhances tolerance to abiotic stress in transgenic rice. Plant Biotechnol J 5:646–656

Osakabe Y, Osakabe K, Shinozaki K, Tran LSP (2014) Response of plants to water stress. Front Plant Sci 5:86

Ozturk ZN, Talame V, Deyhoyos M, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N, Tuberosa R, Bohnert HJ (2002) Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley. Plant Mol Biol 48:551–573

Papaefthimiou D, Tsaftaris AS (2012) Significant induction by drought of HvPKDM7-1, a gene encoding a jumonji-like histone demethylase homologue in barley (H. vulgare). Acta Physiol Plant 34:1187–1198

Qian G, Liu Y, Ao D, Yang F, Yu M (2008) Differential expression of dehydrin genes in hull-less barley (*Hordeum vulgare* ssp. *vulgare*) depending on duration of dehydration stress. Can J Plant Sci 88:899–906

Qiu L, Wu DZ, Ali S, Cai S, Dai F, Jin X, Wu F, Zhang G (2011) Evaluation of salinity tolerance and analysis of allelic function of *HvHKT1* and *HvHKT2* in Tibetan wild barley. Theor Appl Genet 122:695–703

Rapacz M, Koscielniak J, Jurczyk B, Adamska A, Wojcik M (2010) Different patterns of physiological and molecular response to drought in seedlings of malt and feed-type barleys (*Hordeum vulgare*). J Agron Crop Sci 196:9–19

Reddy VS, Reddy ASN (2004) Proteomics of calcium-signaling components in plants. Phytochemistry 65:1745–1776

Reshetova P, Smilde AK, van Kampen AH, Westerhuis JA (2014) Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data. BMC Syst Biol 8(Suppl 2):S2

Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD (2012) Genome-wide patterns of *Arabidopsis* gene expression in nature. PLoS Genet 8:e1002662

Rodriguez EM, Svenson JT, Malatrasi M, Choi DW, Close TJ (2005) Barley *Dhn13* encodes a KS-type dehydrin with constitutive and stress responsive expression. Theor Appl Genet 110:852–858

Rodríguez-Rosales MP, Gálvez FJ, Huertas R, Aranda MN, Baghour M, Cagnac O, Venema K (2009) Plant NHX cation/proton antiporters. Plant Signal Behav 4:265–276

Roslyakova TV, Molchan OV, Vasekina AV, Lazareva EM, Sokolik AI, Yurin VM, de Boer AH, Babakov AV (2011) Salt tolerance of barley: relations between expression of isoforms of vacuolar Na^+/H^+-antiporter and $^{22}Na^+$ accumulation. Russ J Plant Physiol 58:24–35

Roxas VP, Smith RK, Allen ER, Allen RD (1997) Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress. Nat Biotechnol 15:988–991

Roy SJ, Huang W, Wang XJ, Evrard A, Schmöckel SM, Zafar ZU, Tester M (2013) A novel protein kinase involved in Na^+ exclusion revealed from positional cloning. Plant Cell Environ 36:553–568

Roy SJ, Negrão S, Tester M (2014) Salt resistant crop plants. Curr Opin Biotechnol 26:115–124

Rozema J, Flowers T (2008) Crops for a salinized world. Science 322:1478–1480

Rushton PJ, Somssich IE, Ringler P, Shen QJ (2010) WRKY transcription factors. Trends Plant Sci 15:247–258

Schilling RK, Marschner P, Shavrukov Y, Berger B, Tester M, Roy SJ, Plett DC (2014) Expression of the *Arabidopsis* vacuolar H^+-pyrophosphatase gene (*AVP1*) improves the shoot biomass of transgenic barley and increases grain yield in a saline field. Plant Biotechnol J 12:378–386

Seki M, Okamoto M, Matsui A, Kim JM, Kurihara Y, Ishida J, Morosawa T, Kawashima M, To TK, Shinozaki K (2009) Microarray analysis for studying the abiotic stress responses in plants. In: Jain SM, Brar DS (eds) Molecular techniques in crop improvement. Springer, Amsterdam, pp 333–355

Shabala S, Shabala S, Cuin TA, Pang J, Percey W, Chen Z, Conn S, Eing C, Wegner LH (2010) Xylem ionic relations and salinity tolerance in barley. Plant J 61:839–853

Shanker AK, Maheswari M, Yadav SK, Desai S, Bhanu D, Attal NB, Venkateswarlu B (2014) Drought stress responses in crops. Funct Integr Genomics 14:11–22

Shelden MC, Roessner U (2013) Advances in functional genomics for investigating salinity stress tolerance mechanisms in cereals. Front Plant Sci 4:123

Skirycz A, Inze D (2010) More from less: plant growth under limited water. Curr Opin Biotech 21:197–203

Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KF, Platzer M, Stein N (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. BMC Genomics 10:547

Suprunova T, Krugman T, Fahima T, Chen G, Shams I, Korol A, Nevo E (2004) Differential expression of dehydrin genes in wild barley, *Hordeum spontaneum*, associated with resistance to water deficit. Plant Cell Environ 27:1297–1308

Talame V, Ozturk ZN, Bohnert HJ (2007) Barley transcript profiles under dehydration shock and drought stress treatments: a comparative analysis. J Exp Bot 58:229–240

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

Tommasini T, Svensson JT, Rodriguez EM, Wahid A, Malatrasi M, Kato K, Wanamaker S, Resnik J, Close TJ (2008) Dehydrin gene expression provides an indicator of low temperature and drought stress: transcriptome-based analysis of barley (*Hordeum vulgare* L.). Funct Integr Genomics 8:387–405

Trenberth KE, Dai A, van der Schrier G, Jones PD, Barichivich J, Briffa KR, Sheffield J (2014) Global warming and changes in drought. Nat Climate Change 4:17–22

Ueda A, Shi W, Nakamura T, Takabe T (2002) Analysis of salt-inducible genes in barley roots by differential display. J Plant Res 115:119–130

Ueda A, Kathiresan A, Inada M, Narita Y, Nakamura T, Shi W, Takabe T, Bennett J (2004) Osmotic stress in barley regulates expression of a different set of genes than salt stress does. J Exp Bot 55:2213–2218

Ullrich SE (2011) Significance, adaptation, production, and trade of barley. In: Ullrich SE (ed) Barley production, improvement, and uses. Wiley, Ames, pp 3–13

Van Gool D, Vernon L (2006) Potential impacts of climate change on agricultural land use suitability: barley. Report No. 302, Department of Agriculture, Western Australia

Von Bothmer R (1992) The wild species of *Hordeum*: relationships and potential use for improvement of cultivated barley. In: Shewry PR (ed) Barley: genetics, biochemistry, molecular biology and biotechnology. CAB International, Wallingford, Oxon, pp 3–18

Walia H, Wilson C, Condamine P, Liu X, Ismail AM, Wanamaker SI, Mandal J, Xu J, Cui X, Close TJ (2005) Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. Plant Physiol 139:822–835

Walia H, Wilson C, Wahid A, Condamine P, Cui X, Close TJ (2006) Expression analysis of barley (*Hordeum vulgare* L.) during salinity stress. Funct Integr Genomics 6:143–156

Wang FZ, Wang QB, Kwon SY, Kwak SS, Su WA (2005) Enhanced drought tolerance of transgenic rice plants expressing a pea manganese superoxide dismutase. J Plant Physiol 162:465–472

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Weinl S, Kudla J (2009) The CBL-CIPK Ca^{2+}-decoding signaling network: function and perspectives. New Phytol 184:517–528

Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. Plant J 59:712–722

Witzel K, Weidner A, Surabhi GK, Börner A, Mock HP (2009) Salt stress-induced alterations in the root proteome of barley genotypes with contrasting response towards salinity. J Exp Bot 60:3545–3557

Wojcik-Jagla M, Rapacz M, Barcik W, Janowiak F (2012) Differential regulation of barley (*Hordeum distichon*) *HVA1* and *SRG6* transcript accumulation during the induction of soil and leaf water deficit. Acta Physiol Plant 34:2069–2078

Wu D, Shen Q, Cai S, Chen ZH, Dai F, Zhang G (2013) Ionomic responses and correlations between elements and metabolites under salt stress in wild and cultivated barley. Plant Cell Physiol 54:1976–1988

Xu WF, Shi WM, Ueda A, Takabe T (2008) Mechanisms of salt tolerance in transgenic *Arabidopsis thaliana* carrying a peroxisomal ascorbate peroxidase gene from barley. Pedosphere 18:486–495

Xu ZS, Ni ZY, Li ZY, Li LC, Chen M, Gao DY, Yu XD, Liu P, Ma YZ (2009) Isolation and functional characterization of *HvDREB1*, a gene encoding a dehydration-responsive element binding protein in *Hordeum vulgare*. J Plant Res 122:121–130

Xu H, Gao Y, Wang J (2012) Transcriptomic analysis of rice (*Oryza sativa*) developing embryos using the RNA-Seq technique. PLoS One 7:e30646

Xue GP, Loveridge CW (2004) *HvDRF1* is involved in abscisic acid-mediated gene regulation in barley and produces two forms of AP2 transcriptional activators, interacting preferably with a CT-rich element. Plant J 37:326–339

Yu Q, An L, Li W (2014) The CBL-CIPK network mediates different signaling pathways in plants. Plant Cell Rep 33:203–214

Zhao J, Sun H, Dai H, Zhang G, Wu F (2010) Difference in response to drought stress among Tibet wild barley genotypes. Euphytica 172:395–403

Zhu JK (2003) Regulation of ion homeostasis under salt stress. Curr Opin Plant Biol 6:441–445

Ziemann M, Kamboj A, Hove RM, Loveridge S, El Osta A, Bhave M (2013) Analysis of the barley leaf transcriptome under salinity stress using mRNA-Seq. Acta Physiol Plant 35:1915–1924

Zohary D, Hopf M (1993) Domestication of plants in the Old World. The origin and spread of cultivated plants in West Asia, Europe and the Nile Valley. Clarendon Press, Oxford

miRNA Profiling in Plants: Current Identification and Expression Approaches

Bilgin Candar-Cakir and Ozgur Cakir

Contents

Abstract MicroRNAs (miRNAs) are the members of small noncoding RNA molecules in eukaryotes that regulate the posttranscriptional gene expression of target mRNAs positively or negatively via their degradation and/or translational inhibition. miRNAs play a crucial role in biological processes such as growth, development, maturation, cell differentiation, and response to various abiotic and biotic stress factors. Therefore, miRNA discovery and determining the functional aspects of miRNA expressions and their targets are important for breeding strategies and plant biotechnology. There are several computational and experimental approaches to identify miRNAs and reveal the expression patterns such as cloning, homology-based approaches, high-throughput deep sequencing, quantitative real-time PCR (qRT-PCR) and hybridization-based methods such as microarray and northern blot. Here, we describe these recent approaches for miRNA profiling on some model and non-model plant species following brief information about miRNA evolution and biogenesis.

B. Candar-Cakir
Programme of Molecular Biology and Genetics, Institute of Science, Istanbul University, 34134 Vezneciler, Istanbul, Turkey
e-mail: bilgincandar@gmail.com

O. Cakir (✉)
Department of Molecular Biology and Genetics, Faculty of Science, Istanbul University, Istanbul, Turkey
e-mail: ozgurckr@istanbul.edu.tr

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_8

Keywords Bioinformatics • Deep sequencing • Expression • Microarray • MicroRNA • Identification • Plant

1 Introduction

MicroRNAs have been firstly discovered in *C. elegans* while investigating *lin-4* gene function which is responsible for the controlling of larval development (Lee et al. 1993). 22 nt-length small RNA, *lin-4*, was figured out to repress *lin-14* gene expression binding to 3′ UTR region with mutant analysis (Lee et al. 1993). Then, the discovery of *let-7* in *C. elegans* followed the detection of *lin-4* (Reinhart et al. 2000). Up to date, many miRNAs have been found in eukaryotes and viruses (Bartel 2004; Carrington and Ambros 2003; Schwab et al. 2005), and computational and experimental studies showed that miRNAs are mostly conserved among species (Pasquinelli et al. 2000). MicroRNAs are described as regulatory factors of gene expression post transcriptionally. These endogenous, noncoding small RNA molecules are generally 21–24 nt-long and they play an essential role in gene expression by targeting mRNAs which have complementary sequences. Once the complementary sequence was found and annealed, the mRNA goes under degradation via cleavage or is repressed at translational level. Plant miRNAs acquire high sequence similarity with their targets, whereas animal miRNAs may work with somewhat similar sequences (Jones-Rhoades et al. 2006). MiRNAs are involved in almost every cellular mechanisms such as development (Moxon et al. 2008), abiotic and biotic stress responses (Jin and Wu 2015; Xie et al. 2015b; Yin et al. 2014), hormone signal transduction (Yin et al. 2014), etc. One miRNA can regulate not only a single gene but also a network; therefore it may affect different metabolisms at the same time indicating regulation of genetic networks.

Since the day they were discovered, miRNAs are the source of a great attraction and over the passing ten years, a great number of new bioinformatic softwares and tools are developed. By using these technologies with the combination of the next-generation deep sequencing, miRNA identification and expression studies in plants have increased dramatically. The data about the identified miRNAs are being kept in miRBase Release (v21 on June 26th, 2014) database (http://www.mirbase.org/ftp.shtml) (Kozomara and Griffiths-Jones 2014) which is highly dynamic. 6992 precursor sequences and 8496 mature miRNAs from 73 plant species were annotated in the last release.

In this chapter, we aimed to describe briefly biogenesis of plant miRNAs and different ways for discovering and expression profiling depends on computational and experimental methods and also research experiments related to these strategies.

2 Plant miRNA Biogenesis and Functional Aspects

MicroRNAs, as the member of small RNA classes, control a great number of biological and metabolic processes such as growth, development, ripening, abiotic stress response, and pathogen defense regulating target genes post-transcriptionally (Xie et al. 2012). miRNAs are recently targeted for plant biotechnology researches to improve specific crop traits including biotic stress tolerance like bacteria, fungi, and nematode; biotic stress tolerance such as drought, heat, cold, salt, oxidative stress, nutrient deprivation, and heavy metal detoxification; plant biomass and grain yield; developmental processes like seed, root, fruit, floral development, and nutritional quality (Zhang and Wang 2015; Zhou and Luo 2013).

miRNAs are encoded by endogenous *MIR* genes which are mostly conserved among plant species (Nozawa et al. 2012; Reinhart et al. 2002). These genes may found mostly at intergenic, but sometimes intragenic regions of the genome as exon or intron sequences (Eldem et al. 2013; Nozawa et al. 2012; Xie et al. 2015c). *MIR*s are transcribed like any other protein coding genes and the promoter regions of these genes resemble to other genes (Ozsolak et al. 2008). These promoters are controlled by various transcription factors, enhancers, and silencing elements. The transcription of a miRNA gene is closely associated with a specific biological or metabolomic process. MIR genes are transcribed by DNA-dependent RNA polymerase II (Pol II) into primary miRNA transcripts (pri-miRNA) (Xie et al. 2005). The functional miRNAs are produced by pri-miRNAs. Although the mature miRNAs are usually 20–22 nucleotides, the pri-miRNAs are much longer and their length varies between 100 and 1000 nucleotides (Zhang and Wang 2015). The number of miRNA genes and their length are variable among the organisms; basically they are generally longer in plants when compared to animals. In our chapter, we briefly explain the miRNA biogenesis in plants below. The detailed description about miRNA biogenesis is well reviewed in Axtell et al. (2011), Ha and Kim (2014), Krol et al. (2010), Rogers and Chen (2013).

After the pri-miRNA transcription of *MIR* genes by RNA polymerase II, they fold to form hairpin secondary structures (Chen and Rajewsky 2007) (Fig. 1). DAWDLE (DLD) is a nuclear RNA-binding protein which is necessary for miRNA stabilization until the pri-miRNA processing is completed (Yu et al. 2008). After transcription, pri-miRNAs are processed to 70 nt-length precursor miRNAs (pre-miRNAs) catalyzing by RNase III enzyme Dicer-like1 (DCL) together with HYPONASTIC LEAVES 1 (HYL 1) and SERRATE (SE) proteins (Khraiwesh et al. 2012; Reinhart et al. 2002; Voinnet 2009). Researches about miRNA biogenesis indicates that plant miRNA processing starts and ends in nucleus but for animals miRNA processing continues at cytoplasm, since there is no Drosha-like proteins in plants (Park et al. 2005). DCL has important domains like RNA-binding, cleavage and other double-strand processing domains. Most importantly, plants appear to possess more than one type of Dicer proteins to mature different kinds of miRNAs (Reinhart et al. 2002). The number of DCL enzymes in plants is different, for instance, it is known that *Arabidopsis* has four DCL enzymes (Fahlgren et al.

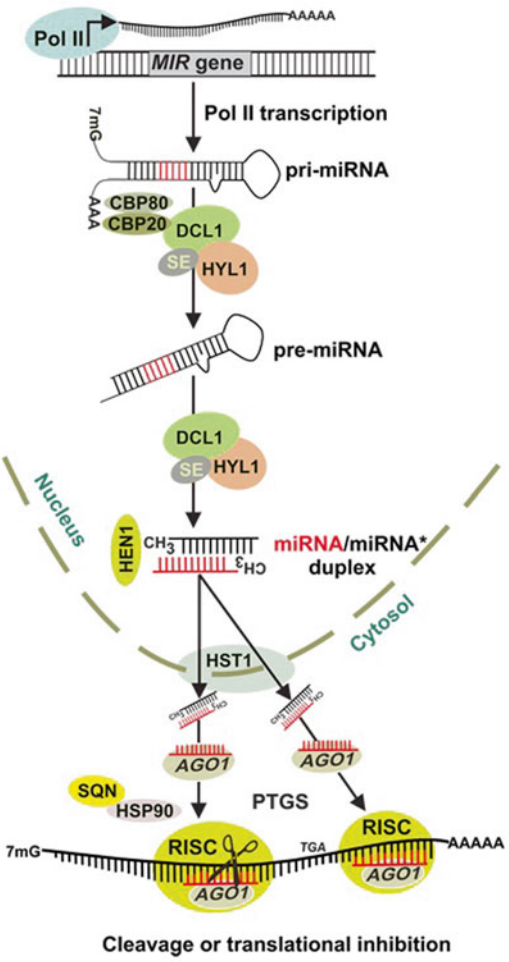


Fig. 1 Biogenesis of miRNAs in plants [from Khraiwesh et al. (2012)]

2007). This result indicates the variation of miRNA biogenesis in different plant species (Xie et al. 2015c). After the formation of DCL-HYL1-SE complex, the pre-miRNAs are converted to double-strand miRNA/miRNA* duplexes (Fig. 1). The 3′ overhang of this dsRNA is then methylated by HEN1 (HUA ENHANCER 1) and exported to cytoplasm by the protein HASTY (HST) (Eldem et al. 2013; Khraiwesh et al. 2012; Park et al. 2005). The miRNA strand is then incorporated to the complex named RISC (RNA-induced silencing complex). This complex with the catalytic component Argonaute protein (AGO) targets the mRNA which is going to be cleaved or translationally repressed (Fig. 1). AGO proteins have the activity to slice the mRNAs (Rogers and Chen 2013). The miRNA-associated gene regulation of target genes occurs in two different ways as mRNA degradation or translational repression. Mostly, the complementarity between miRNA and mRNA sequences leads to degradation of mRNA by cleavage (Bartel 2004). Degradome sequencing

studies also revealed the slicing events in plants (Addo-Quaye et al. 2008; German et al. 2009; Gregory et al. 2008). Translational repression is thought to take place when the complementarity between the sequences of miRNA and mRNA is imperfect although it should be still near perfect in plants for repression of translation (Bartel 2004).

3 Strategies for miRNA Identification and Expression

3.1 Identification Approaches Based on Conventional and Computational Methods

Strategies to identify miRNAs in different organisms are involved in conventional experimental methods and computational tools which also should be supported by experimental methods. There are some difficulties to identify miRNAs by conventional methods because of their characteristics such as short length and not having special sequences like poly(A)-tail. Also the low abundance of small RNAs is another drawback (Eldem et al. 2013). The small RNAs amount in total RNAs is as low as 0.01 % (Pritchard et al. 2012). Because of the challenges that occur naturally based on miRNAs features, some special techniques should be used for identifying and profiling of small RNAs.

Researchers used forward genetic methods to identify the first miRNAs. When compared to today's miRNA identification approaches, genetic screening is time-consuming and sometimes miRNAs cannot be detected because of redundancy (Unver et al. 2009). Another strategy, cloning and sequencing is based on ligating the miRNAs with adapter sequences to make them reverse transcribed (Lu et al. 2005). Then the reverse transcribed miRNA can be amplified by PCR with the help of adapter sequences, cloned and Sanger sequenced (Sunkar and Zhu 2004). This can be applied for identifying the miRNAs which are specific to a developmental stage or tissue of an organism. Also, blast-based EST (expressed sequenced tags) and GSS (genomic sequence survey) analyses might be used for miRNA identification based on the miRNA conservation (Zhang et al. 2005). Many miRNAs are evolutionary conserved among species. This feature allows us to compare the sequences of precursor and mature miRNA sequences among species. EST and GSS-based blast strategy may provide information about conservation and evolution of miRNAs. When the genomic sequence of a species is not known, this strategy may be applied (Zhang et al. 2006). For example, miRNA sequences of model plants, such as *Arabidopsis* and rice, contribute the detection of unknown miRNAs in other species by using EST and GSS analysis.

With the developments on computational biology, many bioinformatics tools have been developed till today for identification and analyzing of miRNAs according to their sequences (Li et al. 2012b). The sequences of the precursor miRNAs allow to reveal the secondary structures and the hairpin loops are easily predicted by miRNA-

specific databases and tools such as miRBase (Kozomara and Griffiths-Jones 2014), PNRD ((Yi et al. 2015), and Plant miRNAT (Rhee et al. 2015) (Table 1). Since the specificity and sensitivity variation of prediction softwares, researchers use different tools to predict miRNAs taking advantage of sequence conservation among species. Some softwares contain machine learning methods and use two sets of data, one with known miRNAs as a positive data and other data set with hairpin but contains no miRNA as the negative one. By processing these data sets, softwares build a prediction tool (Kang and Friedlander 2015). Also, there are softwares called target centered approaches that use target mRNAs to predict miRNAs (Gomes et al. 2013; Xie et al. 2005). All prediction approaches take advantage of miRNA features not only on sequence comparison but also on novel miRNA prediction. The identification of miRNAs is followed by determination of functional aspects which is related to possible targets. Plant miRNAs mostly have high similarity with their targets (Rhoades et al. 2002) and this property allows the researchers to generate sensible target prediction and identification tools (Table 1). For example, psRNATarget (Dai and Zhao 2011) is an online tool to predict miRNA targets allowing users to handle deep-sequencing data. TAPIR is another program that can predict miRNA targets especially for plants (Bonnet et al. 2010). mTide (Zhang et al. 2015), comTAR (Chorostecki and Palatnik 2014), and SoMART (Li et al. 2012a) are also used for miRNA target prediction (Table 1). After computational prediction, all targets need to be validated experimentally. A modified 5' RACE (Rapid Amplified cDNA Ends)-PCR that determines the cleavage sites of miRNAs pairing with target mRNAs is commonly employed for this purpose (Llave et al. 2002). High-throughput sequencing approaches, parallel analysis of RNA ends (PARE) (German et al. 2008), degradome sequencing (Addo-Quaye et al. 2008), and genome-wide mapping of uncapped transcripts (GMUCT) (Gregory et al. 2008) allow genome-wide identification and validation of miRNA targets (Hou et al. 2014; Thomson et al. 2011). After sequencing, modified 5' RACE-PCR validation is required and bioinformatics tools are also needed to analyze sequencing data such as CleaveLand used for degradome data mining (Addo-Quaye et al. 2009). Sequencing-based discovery of miRNA targets has been carried out on a large number of plants such as *Arabidopsis* (Addo-Quaye et al. 2008), soybean (Shamimuzzaman and Vodkin 2012), maize (Zhao et al. 2012), tomato (Karlova et al. 2013), and poplar (Li et al. 2013).

After miRNA and target identification, gene ontology (GO) (http://geneontology.org/) (Ashburner et al. 2000; Gene Ontology Consortium 2015) enrichment and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (http://www.kegg.jp/) (Kanehisa 2016) analyses are needed to understand the complex networks controlled by miRNAs in plant growth, development, and stress response. GO enrichment enables to determine biological processes, functional and cellular components of plant targets regulated by miRNAs. DAVID (The Database for Annotation, Visualization and Integrated Discovery) (Huang et al. 2009), GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion tool) (Eden et al. 2009), and agriGO (GO Analysis Toolkit and Database for Agricultural Community (Du et al. 2010) are commonly used online tools for enrichment and pathway analyses in plants (Table 1).

Table 1 Publicly available major plant miRNA databases and computational tools for plant target prediction and identification as well as gene enrichment and pathway analysis

Name	Website	Reference
miRNA databases		
miRBase	http://www.mirbase.org	Griffiths-Jones (2004), Griffiths-Jones et al. (2006, 2008), and Kozomara and Griffiths-Jones (2011, 2014)
miRTarBase	http://mirtarbase.mbc.nctu.edu.tw/	Chou et al. (2015)
mTide	http://bis.zju.edu.cn/MTide/	Zhang et al. (2015)
PmiRKB	http://bis.zju.edu.cn/pmirkb/	Meng et al. (2011)
Plant MPSS	http://mpss.udel.edu/	Nakano et al. (2006)
PMTED	http://pmted.agrinome.org/	Sun et al. (2013)
PNRD	http://structuralbiology.cau.edu.cn/PNRD/index.php	Yi et al. (2015)
miRNA identification and analysis tools		
C-mii	http://www.biotec.or.th/isl/c-mii	Numnark et al. (2012)
findmiRNA	http://sundarlab.ucdavis.edu/mirna/	Adai et al. (2005)
HHMMiR	http://biodev.hgen.pitt.edu/kadriAPBC2009.html	Kadri et al. (2009)
miRAlign	http://bioinfo.au.tsinghua.edu.cn/miralign/	Wang et al. (2005)
miRanalyzer	http://bioinfo5.ugr.es/miRanalyzer/miRanalyzer.php	Hackenberg et al. (2011)
miRcheck	http://bartellab.wi.mit.edu/software.html	Jones-Rhoades and Bartel (2004)
miRDeep-P	http://faculty.virginia.edu/lilab/miRDP/	Yang and Li (2011)
miRExpress	http://mirexpress.mbc.nctu.edu.tw/	Wang et al. (2009)
miRNest	http://rhesus.amu.edu.pl/mirnest/copy/	Szczesniak and Makalowska (2014)
miRPlant	http://www.australianprostatecentre.org/research/software/mirplant	An et al. (2014)
mirTools 2.0	http://122.228.158.106/mr2_dev/	Wu et al. (2013)
PasmiR	http://pcsb.ahau.edu.cn:8080/PASmiR/	Zhang et al. (2013)
Plant miRNAT	https://sites.google.com/site/biohealthinformaticslab/resources	Rhee et al. (2015)
PmiRKB	http://bis.zju.edu.cn/pmirkb/	Meng et al. (2011)
PMRD	http://bioinformatics.cau.edu.cn/PMRD/	Zhang et al. (2010) (updated version is PNRD)
SemiRNA	http://www.bioinfocabd.upo.es/semirna/	Munoz-Merida et al. (2012)
ShortStack	http://axtell-lab-psu.weebly.com/shortstack.html	Axtell (2013)
Target prediction and identification tools		
CleaveLand	http://axtell-lab-psu.weebly.com/cleaveland.html	Addo-Quaye et al. (2009) and Brousse et al. (2014)
C-mii	http://www.biotec.or.th/isl/c-mii	Numnark et al. (2012)

(continued)

Table 1 (continued)

Name	Website	Reference
comTAR	http://rnabiology.ibr-conicet.gov.ar/comtar/	Chorostecki and Palatnik (2014)
mTide	http://bis.zju.edu.cn/MTide/	Zhang et al. (2015)
psRobot	http://omicslab.genetics.ac.cn/psRobot/	Wu et al. (2012)
psRNATarget	http://plantgrn.noble.org/psRNATarget/	Dai and Zhao (2011)
p-TAREF	http://scbb.ihbt.res.in/new/p-taref/form1.html	Jha and Shankar (2011)
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid	Kruger and Rehmsmeier (2006) and Rehmsmeier et al. (2004)
SoMart	http://bakerlab.berkeley.edu/somart-webserver-mirna-sirna-analysis	Li et al. (2012a)
TAPIR	http://bioinformatics.psb.ugent.be/webtools/tapir/	Bonnet et al. (2010)
TargetFinder	http://carringtonlab.org/resources/targetfinder	Allen et al. (2005), Fahlgren and Carrington (2010), and Fahlgren et al. (2007)
Gene enrichment and pathway analyses tools		
AmiGO 2	http://amigo2.berkeleybop.org/amigo	Carbon et al. (2009)
agriGO	http://bioinfo.cau.edu.cn/agriGO/	Du et al. (2010)
BiNGO	http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html	Maere et al. (2005)
Blast2GO	https://www.blast2go.com/	Conesa and Gotz (2008) and Conesa et al. (2005)
DAVID	https://david.ncifcrf.gov/	da Huang et al. (2009a, b)
EasyGO	http://bioinformatics.cau.edu.cn/easygo/	Zhou and Su (2007)
GFSAT	http://nclab.hit.edu.cn/GFSAT/	Xu et al. (2013b)
GO	http://geneontology.org/	Ashburner et al. (2000) and Gene Ontology Consortium (2015)
GOEAST	http://omicslab.genetics.ac.cn/GOEAST/tools.php	Zheng and Wang (2008)
GOrilla	http://cbl-gorilla.cs.technion.ac.il/	Eden et al. (2007, 2009)
KEGG	http://www.kegg.jp/	Kanehisa (2016)
RGAP	http://rice.plantbiology.msu.edu/	Kawahara et al. (2013)
QuickGO	https://www.ebi.ac.uk/QuickGO/	Binns et al. (2009)
SoymiRFN	http://nclab.hit.edu.cn/SoymiRNet/	Xu et al. (2014)
UniProt-GOA	http://www.ebi.ac.uk/GOA	Huntley et al. (2015)

Table is modified from Tripathi et al. (2015) and Zhang and Wang (2015)

3.2 Expression Analysis Based on Experimental Methods

Since miRNAs regulate the expression of plant growth, development, and environmental stress response-related target mRNAs, the determination of expression levels in tissue, genotype and time-dependent manners is important in terms of improving plant yield (Zhang 2015). For this purpose, several experimental approaches are applied to characterize functionally plant miRNAs. Even today, genetic screening (Lee et al. 1993) and direct cloning (Reinhart et al. 2002) methods are not commonly used for miRNA identification and expression; the preliminary roles of miRNAs in plants were discovered with these approaches. Recently, four major technologies are mostly preferred for miRNA profiling in plants: Northern blotting, quantitative reverse transcription PCR (qRT-PCR), microarray, and high-throughput sequencing.

3.2.1 Northern Blotting

Northern blotting as one of the hybridization-based approaches is extensively used for miRNA profiling. This method is generally employed together with polyacrylamide gel electrophoresis to detect expression levels of both mature miRNAs and their precursors and also allows size determination of small RNA molecules (Valoczi et al. 2004). Besides, northern blot is the oldest and efficient standard approach to validate miRNA profiling results of novel miRNA identification and expression technologies such as miRNA microarray and small RNA sequencing (Dong et al. 2013; Zhang and Wang 2015). However, it has some disadvantages such as requirement for large amount and high quality total RNA especially when experienced with low-abundant miRNAs, low sensitivity, time-consuming in comparison with other approaches and especially enabling only known miRNA analysis (Dong et al. 2013; Valoczi et al. 2004; Zhang and Wang 2015). To resolve some drawbacks of northern blot, novel probe types were developed (Koscianska et al. 2011). Recently three different northern blotting procedures are used for miRNA expression and validation analyses. Firstly, locked nucleic acid (LNA)-modified oligonucleotide probes were developed for improving hybridization efficiency (Valoczi et al. 2004; Varallyay et al. 2007, 2008). Then, regular UV cross-linking of RNAs to nylon membranes was changed to chemical cross-linking way that increases the miRNA determination possibility with EDC [1-ethyl-3-(3-dimethylaminopropyl) carbodiimide] (Pall et al. 2007; Pall and Hamilton 2008). The third approach is combined with the properties of LNA and EDC with digoxigenin (DIG)-labeled probes (Kim et al. 2010). Despite the novelties of northern blotting method, today it is not commonly preferred in miRNA expression studies because of allowing the profiling of only predicted and known miRNAs and limited number of miRNAs in one assay. However, it is still very effective for validation of miRNA expressions after microarray and deep-sequencing analyses. For instance, selected several miRNAs were

confirmed by northern blot after identification via deep sequencing in soybean (Li et al. 2011) and *Brassica* (Zhou et al. 2012). Similarly, (Lang et al. 2011) validated the expressions of tomato miRNAs with northern blotting after microarray experiment.

3.2.2 qRT-PCR

One of the commonly used strategies to compare tissue, development, and treatment-specific miRNA expression in plants is RT-PCR (reverse transcription-polymerase chain reaction). (Chen et al. 2005) developed individual Taqman miRNA assay based on quantitative real-time PCR (qRT-PCR) for miRNA detection. This approach contains reverse transcription of miRNAs with stem-loop primers followed by standard Taqman PCR. Then, (Varkonyi-Gasic et al. 2007) reported a RT-PCR protocol for miRNA expression. In this protocol, firstly designed stem-loop RT primers (according to Chen et al. 2005) are bound to 3′ end of miRNA sequences and reverse transcription is carried out. In next step, RT-PCR reaction is employed using forward primers specific to miRNAs and universal reverse primer (Chen et al. 2005; Kramer 2011; Varkonyi-Gasic et al. 2007)(Fig. 2a). Stem-loop primers are preferred due to base stacking property of structure increasing sensitivity and specificity in comparison with linear primers (Chen et al. 2005; Varkonyi-Gasic et al. 2007). Alternatively, miRNA qRT-PCR assays might be employed with the SYBR Green I dye (Fig. 2b) or Universal Probe Library (UPL; Roche Diagnostics) probes (Fig. 2c) that lead to detection of low copy miRNAs separately or multiplex formats (Varkonyi-Gasic et al. 2007). Although qRT-PCR is more sensitive than northern blotting enabling a number of miRNA expression simultaneously, it provides only known miRNA expression profiling. However, this technology is commonly used not only in detecting miRNA expression but also in validating expression results in plant species following microarray and deep-sequencing data (Zhang and Wang 2015). qRT-PCR has been chosen in some stress-related and tissue-specific miRNA expression analyses (Korir et al. 2013; Luan et al. 2014; Sun et al. 2012, 2014; Wang et al. 2013; Xu et al. 2013a; Zhuang et al. 2014). Luan et al. (2014) investigated biotic and abiotic stress-associated miRNA expression with qRT-PCR following homology research. They applied salt and drought stress on tomato leaves for abiotic stress and infected the leaves with Phytophthora infestans for biotic stress. After homology analysis, they selected miR157, miR170, miR398, miR473, miR479, miR828, miR830, miR1446, miR2111, and miR2118 for expression analyses via RT-PCR and investigated the expression of miR398 with qRT-PCR. They figured out that the expression was decreased gradually in time-dependent manner in all stress treatments suggesting the function of miR398 in tomato stress response system. Besides, there are a great number of validation studies in plants obtained by qRT-PCR after microarray and small RNA sequencing studies (Gao et al. 2015; Inal et al. 2014; Xie et al. 2014, 2015b).

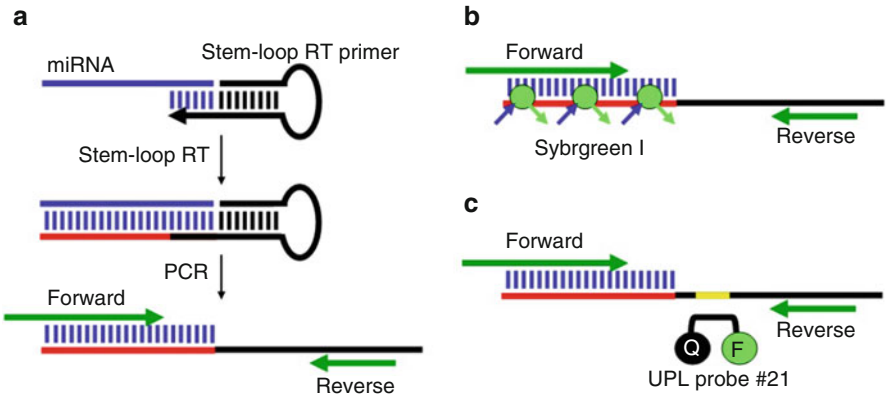


Fig. 2 Illustration of miRNA quantification by stem-loop RT-PCR. (**a**) End-point PCR after stem-loop RT-PCR. (**b**) SYBR Green I assay for real-time PCR. (**c**) Universal ProbeLibrary (UPL) probe assay [from Varkonyi-Gasic et al. (2007)]

3.2.3 Microarray

Microarray technology, one of the best known hybridization methods, can be preferred for expression analysis of large number of plant miRNA transcripts on genome-wide scale (Barrera-Figueroa et al. 2013; Pritchard et al. 2012). Mature, known miRNA sequences of plant species deposited in miRBase database are used for preparation of oligonucleotide microarray slides (Barrera-Figueroa et al. 2013) and generally at least two comparisons (for example, control versus treatment) are needed to calculate the expression of relative miRNA abundance in samples (Pritchard et al. 2012). miRNA microarray technology has the advantage of lower cost than deep-sequencing methods, whereas some restrictions make the technology unfavorable sometimes. For example, this technology shows low sensitivity, has some difficulties for perfect quantification of rare miRNA abundance in a sample, might show false positive results for some miRNAs that have similar sequences, has some challenges to adapt some tissue and developmental stage-specific miRNAs to array platform and allows only detection of conserved known miRNAs among species (Barrera-Figueroa et al. 2013; Pritchard et al. 2012). Since sensitivity is lower in miRNA microarray analysis, validation is required after experiment with qRT-PCR or northern blot (Pritchard et al. 2012). miRNA arrays are widely used today to analyze differential miRNA abundance among plant species or tissue- and stage-specific manners of same plants and investigate the miRNA roles in biological processes such as fruit development (Rosas-Cardenas Fde et al. 2015), abiotic stresses like drought (Kantar et al. 2011; Liu et al. 2008; Pasini et al. 2014; Zhao et al. 2007, 2010), salinity (Ding et al. 2009; Liu et al. 2008; Yin et al. 2012), cold (Liu et al. 2008; Lu et al. 2008; Lv et al. 2010; Zhang et al. 2012, 2014b), heat (Lu et al. 2008), heavy metal (Ding et al. 2011), and nutrient deprivation (Xu et al. 2011; Zamboni et al. 2012; Zhu et al. 2010); biotic stresses caused by agents such as fungus (Feng

et al. 2015; Inal et al. 2014; Jin et al. 2012; Wu et al. 2014), parasite (Verma et al. 2014), and virus (Lang et al. 2011).

Kantar et al. (2011) used root and leaf tissues to analyze drought effect on miRNAs of two drought-tolerant *Triticum dicoccoides* genotypes in time-dependent manner. They determined a total of 13 significantly expressed miRNAs. Three out of those conserved miRNAs belong to leaf tissues (miR528, miR896, miR1867), while miR166, miR171, miR356, miR396, miR474, miR894, miR1432, miR1450, and miR1881 were expressed only in root tissues and miR398 was common between two tissue types. In leaf tissues, miR1867 was upregulated, while miR896 was downregulated in 4-h-stage then upregulated in 8-h-stage exposure. miR528 and miR398, in case, were expressed only in later stage and miR398 was upregulated, whereas miR528 was downregulated. miR396 expression was lower only in early stage root tissues in comparison with control plants and miR166 and miR171 were downregulated in 8-h-stage root samples. The significantly expressed other miRNAs (miR156, miR356, miR474, miR894, miR1432, miR1450, miR1881) were upregulated in root tissues. These results indicated the functional role of miRNAs in wheat in response to shock drought stress. Similarly, (Pasini et al. 2014) investigated the water potential-related miRNA expression in drought-tolerant *Sorghum* genotype at different days and determined that four miRNA families (miR164, miR395, miR399, miR827) were upregulated, whereas miR528 was downregulated at different water potential levels after microarray study.

In rice (*Oryza sativa*), heavy metal-related miRNA expression was investigated via cadmium treatment on time-dependent root tissues (Ding et al. 2011). Only miR528 was upregulated with cadmium exposure, while 18 miRNAs belong to nine families (miR156, miR162, miR166, miR168, miR171, miR390, miR396, miR444, and miR1432) were downregulated. These results indicated the regulatory role of miRNAs in rice response to Cd at molecular level together with heavy metal tolerance in plants. In another study, (Inal et al. 2014) used two different fungal pathogens (*Fusarium culmorum* and *Bipolaris sorokiniana*) for stress treatment on tolerant and sensitive wheat cultivars in order to analyze biotic stress-related miRNA expression. After miRNA microarray analysis, totally 87 significantly expressed miRNAs were determined and in tolerant cultivar 21 upregulated and 41 downregulated miRNAs were detected whereas 20 upregulated and 23 downregulated miRNAs were found in sensitive one. Most of the stimulated miRNAs belonged to *F. culmorum* and conserved miRNAs such as miR169, miR869, miR2592, miR2657, miR4409, miR5208, miR5338, miR5674 were so highly upregulated, besides the expression levels of miR916 and miR3626 were decreased sharply in tolerant genotype. When compared stress-treated samples with control plants in sensitive cultivar, miR916 and miR6436 expressions were decreased almost 100-fold, while the expressions of miR319, miR390, miR1168, miR1427, and miR4402 were increased significantly. With these results, (Inal et al. 2014) suggested that the response of miRNAs to fungal stress is associated with pathogen and cultivar types.

In tomato, host pathogen-related 3 miRNAs were determined using microarray inoculating with fungal pathogen *Botrytis cinerea* and miR160 and miR171 were downregulated after inoculation, whereas miR169 expression was increased (Jin et al. 2012). Verma et al. (2014) investigated miRNA expression changes in the roots

of *Brassica napus* after inoculation with the clubroot agent *Plasmodiophora brassicae* for 10 and 20 days. They detected that 10 miRNAs were differentially expressed at 10-day inoculation, while 34 miRNAs showed different expression profile after 20-day pathogen infection. In 10-day inoculation, miR156, miR166, and miR2916 expressions were increased but miR159, miR169, miR854, and miR909 expressions were downregulated. However, when the inoculation time was extended to 20 days, miR854 and miR909 expressions were increased. Also, some miRNAs were expressed only 20-day-treated plants, while some of them (miR162, miR172, miR396) were downregulated, whereas miR169 and miR948 were upregulated indicating the regulation of disease development-associated target gene expression.

In prickly pear cactus (*Opuntia ficus indica*), the expression levels of conserved miRNAs related to stage-specific fruit development process were investigated with array platform (Rosas-Cardenas Fde et al. 2015) and 34 miRNAs were detected to be expressed significantly in different development stages like floral bud, green, young, and mature fruits. miRNAs such as miR159, miR164, miR390, miR2119, and miR3636 were expressed at all stages, whereas miR394 and miR408 had miRNA abundance in only floral buds. Similarly, miR824 and miR1916 were expressed only in young fruits, while miR169, miR846, and miR3629 had specific expression to green fruit tissues. However, there was no unique miRNA belonging to mature fruit samples. Finally, they suggested that miR164 with the significantly high expression level had a crucial role in fruit development of prickly pear cactus (Rosas-Cardenas Fde et al. 2015).

3.2.4 High-Throughput Deep Sequencing

In the last decade, next-generation sequencing (NGS) became very fast, robust, and favorite method and excessive increase in plant miRNA studies was observed. With the developments in the next-generation sequencing technologies, researchers started to reveal the small RNA repertoire of different species and with bioinformatics tools, it became much more easier to analyze enormous data with the help of high-throughput sequencing to discover, identify, and profile the expression of miRNAs in a very fast and accurate way (Pareek et al. 2011). NGS has been applied in many researches as powerful technology for miRNA identification of model (Fahlgren et al. 2007; Schreiber et al. 2011; Sunkar et al. 2008; Szittya et al. 2008) and non-model plant species (Pantaleo et al. 2010; Song et al. 2010; Zhao et al. 2010) because of its high sensitivity and dynamic range. With next-generation sequencing, very large data can be obtained, known and novel miRNAs and their precursors can be identified and their expressions may be detected (Motameny et al. 2010). The workflow of this approach can be described briefly in a few steps as library preparation, library amplification, and sequencing on different platforms (Fig. 3), although some differences can be observed according to sample types (DNA or RNA) (Knief 2014). For instance, in small RNA sequencing, firstly the RNAs are fractionized to obtain small RNAs. Then they are ligated to adapter sequences and reverse transcribed. After all, they are selected according to their size again and

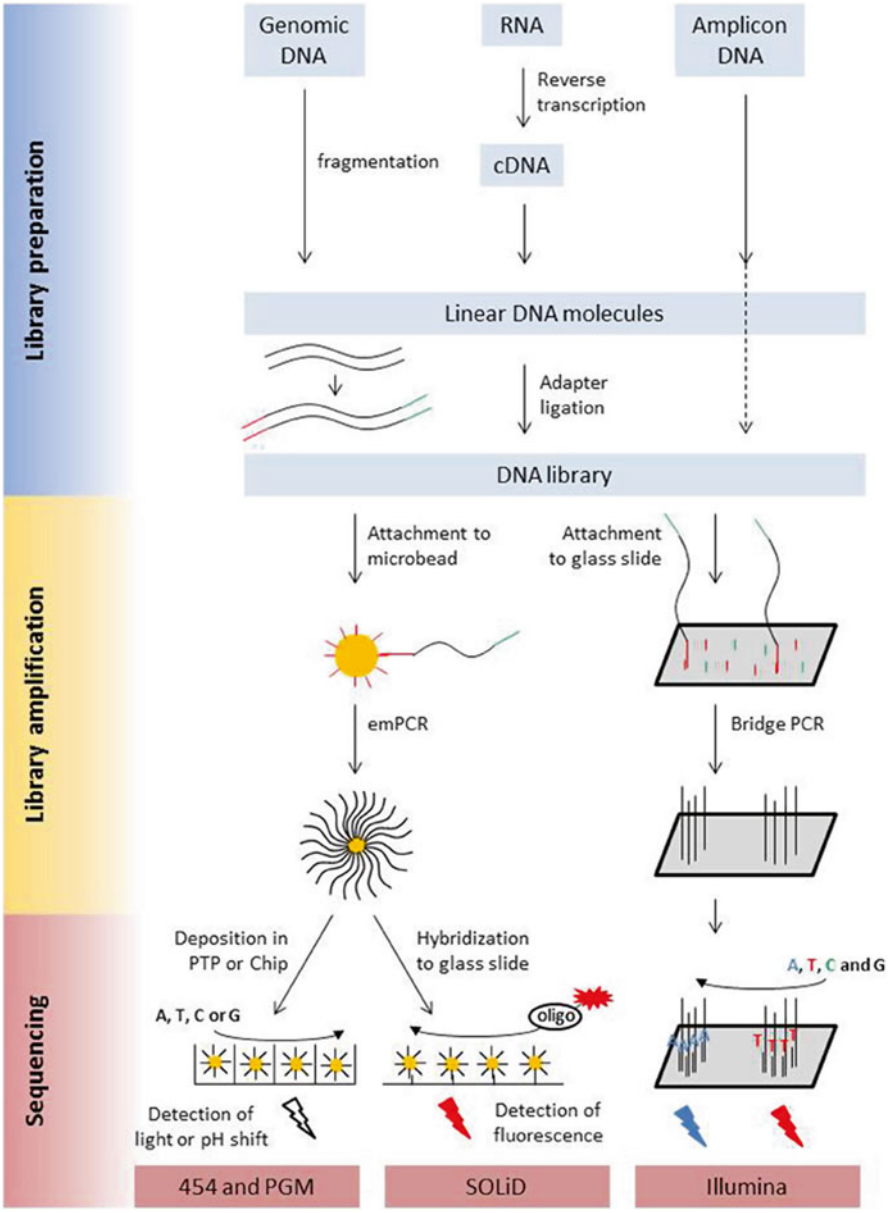


Fig. 3 The general workflow of commonly used NGS platforms [from Knief (2014)]

lastly sequenced with Solexa/Illumina platform (Chu and Corey 2012) (Fig. 3). Sequencing platform should be selected according to experimental design. For example, if low rate sequencing errors are desired, HiSeq sequencing system by Illumina/Solexa platform which is based on fluorophore cyclic reversible terminator technology or SOLiD sequencing system by Applied Biosystems based on

fluorophore cyclic sequencing-by-ligation chemistry should be preferred for miRNA sequencing. Using these technologies, millions of sequencing reactions can be performed and enormous amount of data may be produced in one sequencing even if the reaction time, data amount, and sequencing cost vary among different platforms (Myllykangas et al. 2012). As well as the advantages of next-generation sequencing in comparison with other identification and expression methods, one of the most important issues about this technology is the requirement for computational tools and bioinformatics skills to interpret the data (Zhang and Wang 2015). There are various tools and softwares used to analyze the NGS data and some of them are publicly available (Table 1).

Deep-sequencing approach is widely favored today for miRNA identification and expression analyses of plant species-related significant biological processes such as development of fruit (Gao et al. 2014; Mohorianu et al. 2011; Moxon et al. 2008), fiber (Xie et al. 2015a), ovule (Xie et al. 2015a), root (Yu et al. 2015), ear (Liu et al. 2014b), pollen (Wei et al. 2011), anther (Yan et al. 2015), periodicity stage (Yanik et al. 2013), inflorescence (Barrera-Figueroa et al. 2012), response to abiotic stress factors like nitrogen (Ren et al. 2015), cadmium (Tang et al. 2014), boron (Ozhuner et al. 2013), selenium (Cakir et al. 2015), drought (Bhardwaj et al. 2014; Candar-Cakir et al. 2016; Eldem et al. 2012; Ferreira et al. 2012; Hackenberg et al. 2015; Thiebaut et al. 2014; Xie et al. 2014, 2015b; Yin et al. 2014; Zhang et al. 2014a), salt (Bhardwaj et al. 2014; Kohli et al. 2014; Si et al. 2014; Tian et al. 2014; Xie et al. 2014, 2015b; Yin et al. 2012), heat (Bhardwaj et al. 2014; Kumar et al. 2014; Liu et al. 2014a; Yu et al. 2012), cold (Cao et al. 2014; Ding et al. 2014; Zhang et al. 2014b) and response biotic stress agents such as fungus (Jin and Wu 2015; Kohli et al. 2014), virus (Feng et al. 2014; Pradhan et al. 2015) and nematode (Ding et al. 2015).

In tomato, fruit development stages-related miRNAs were investigated with high-throughput sequencing technology (Mohorianu et al. 2011). Ten conserved and six novel miRNAs were associated with fruit development. Three (miR167, miR172, and miR390) out of ten conserved miRNAs showed differential expression on flower stage, whereas miR159, miR162, miR165 and miR166 were expressed on early fruit development phase. Besides, miR156, miR164, and miR396 showed significant expression changes on fruit ripening stage. The expressions were validated with northern blotting and the results confirmed high-throughput expression abundance. Also six novel miRNAs (miR-T/V/W*/X/Y/Z) were identified and the targets of these miRNAs were validated with 5′ RACE-PCR. Only targets of miR-W* could be determined as membrane-bound ATPase and glutamate permease. These targets play a role in ATP-dependent glutamate transport and accumulation and glutamate has a function in fruit ripening; so they suggested that miRNA expression changes and target regulation may play a role in formation of fruit taste directing glutamate accumulation. Similarly, (Gao et al. 2015) researched the expression of miRNAs which have crucial role in tomato fruit ripening compared with ripening inhibitor (RIN) mutant. Totally, 33 miRNAs were found as associated with ripening and 14 out of them expressed significantly. In control plants, most of the miRNAs were downregulated in fruit development stage, while expression of four miRNAs (miR164, miR399, miR858, miR6026) was increased. In RIN mutants, miR156,

mR164, miR 172, and miR5301 were upregulated against control plants. The expression abundance of these miRNAs and ripening-related targets (transcription factors) were validated by qRT-PCR and they suggested that RIN has effect on miRNAs indirectly and modulate fruit ripening genes in the same way binding to the promoter of miR172.

Xie et al. (2015a) examined fiber and ovule development-related miRNAs in cotton via high-throughput sequencing. They identified a total of 65 conserved miRNA families and 59 miRNAs out of 65 families expressed significantly. At the same time, eight novel miRNAs were identified and three of them showed differential expression pattern. Among conserved miRNAs, 11 and 50 miRNAs were specific to ovule and leaf, respectively, while 59 miRNAs were common between two tissues. For novel miRNAs, only one miRNA was specific to ovule, whereas 4 miRNAs expressed only leaf tissues. miRNA expression abundance was validated via qRT-PCR also. After miRNA identification and expression analysis, target prediction was performed with computational approach and 1498 miRNA-target pairs were predicted belonging to 99 miRNA families and contained 820 genes. They validated the target results with degradome sequencing and determined that 22 pairs were same with prediction results. They revealed miR828, miR164, miR160, and miR171 targets and MYB2D, NAC11, ARF10, and GRAS genes, respectively indicating the miRNA roles in cotton fiber development.

Abiotic stress-related miRNA expression was investigated with drought and salt treatments on cotton seedlings by high-throughput approach and (Xie et al. 2015b) identified 284 conserved and 48 novel miRNAs. One hundred and fifty-five out of them expressed significantly and among 77 significantly expressed conserved miRNAs, miR156, miR166, miR167, miR172, and miR396 were upregulated after salt and drought exposure, whereas expressions of miR160, miR393, miR394, and miR5340 were decreased. The expression abundance of miRNAs was validated by qRT-PCR also. Then, 1895 target genes belong to 271 conserved and 20 novel miRNAs were predicted using EST database and almost 1019 genes were determined to play a role in stress response and fiber development. Fifty five computationally predicted targets were also validated with degradome sequencing. Finally, they suggested that miR164, miR172, miR396, miR1520, miR6158, ghr-n24, ghr-n56, and ghr-n59 are drought- and salinity-responsive miRNAs in cotton and associated with fiber development.

Jin and Wu (2015) examined the miRNA abundance change on tomato leaves after fungal pathogen *Botrytis cinerea* infection. By sRNA deep sequencing, a total of 143 conserved and seven novel miRNA were identified. Out of 150 miRNAs, 57 known miRNAs belong to 24 families and one novel miRNA expressed differentially after inoculation and expression of 41 miRNAs were increased, while 16 miRNAs were downregulated. Among 24 families, seven families (miR159, miR169, miR319, miR394, miR1919, miR1446, miR5300) were upregulated, only miR2111 expression was decreased. Conserved and novel miRNA expressions were validated via qRT-PCR and northern blot. Targets of fungal infection-related miRNAs were predicted with computational tool psRNATarget and the miRNAs belonging to

upregulated seven families were predicted via CleaveLand pipeline, also. MYB transcription factor, TCP transcription factor, F-box protein, and pathogenesis-related transcription factor were target genes of miR159, miR319, miR394, and novel miRNA miRn1, respectively. Finally, they suggested that especially miR319, miR394, and miRn1 can play a crucial role in fungal infection response of tomato leaves.

4 Conclusion

MicroRNAs are essential regulators of gene expression in plants affecting almost every biological process such as fruit and root development, organ differentiation, improved biomass and yield, and abiotic and biotic stress responses (Zheng and Qu 2015). The approaches for identification of plant miRNAs and their targets vary both experimentally, such as next-generation sequencing technologies, and computationally (new tools). With the utilization of these methods, many plant miRNAs, targets, and expression profiles under certain conditions may be characterized and novel miRNAs, their targets, and interactions and discovering the complex cellular networks can be elucidated allowing to enhance plant tolerance to stress conditions, biomass and yield quality and determine the miRNA functions in cells, tissues, and organs. However, new strategies and perspectives are need to better understand miRNA functions, their regulatory mechanisms, and interactions. The first step is the identification of target genes regulated by miRNAs to figure out miRNA function and still there is a requirement for developing new target identification and functional clarification methods (Zhang and Wang 2015). Several strategies are developed to elucidate miRNA and target functions in plant organisms such as construction of vectors generating overexpressed miRNAs or downregulated target genes (Zhou and Luo 2013). Studies designed for overexpressing miRNAs in plants may provide information about their regulatory mechanisms associated with target gene expression and with the understanding the miRNA functions completely, we can take advantage of using them to improve plant features like stress tolerance (Ding et al. 2013; Zhou and Luo 2013). The another more accurate strategy for miRNA functional studies is artificial miRNAs (amiRNAs) which were designed for silencing of target gene expressions specifically and used for some plant species (Ossowski et al. 2008; Schwab et al. 2006; Zhang and Wang 2015). The plants containing amiRNAs which were controlled by constitutive, tissue-specific, or stress-inducible promoters result in downregulated genes indicating potential transgenic plants contacting improved properties related with abiotic or biotic stress tolerance, biomass or yield (Schwab et al. 2006; Zhang and Wang 2015; Zhou and Luo 2013). Recently, new genome editing technologies such as TALEN (transcription activator-like effector nucleases) and CRISPR-Cas9 (clustered regulatory interspaced short palindromic repeats/CRISPR-associated protein 9) systems were developed and they are suggested for creating knockdown/knockout miRNAs (Barrangou et al. 2015; Basak and Nithin 2015; Sprink et al. 2015; Zhang and Wang 2015). Especially,

using CRISPR-Cas9 system for knockout of miRNAs, which were encoded within promoter or hairpin, increases the chance for efficient mutation for miRNAs and their binding sites in targets (Barrangou et al. 2015; Basak and Nithin 2015; Bassett et al. 2014). All these approaches are efficient to understand miRNA and target functional aspects and interactions for plant improvement; however, it may not be sufficient and the other functional noncoding small RNA classes such as long noncoding RNAs (lncRNAs) remain to be elucidated to increase biological and metabolic processes for plants.

Acknowledgments We specially thank Khraiwesh et al. (2012), Knief (2014), and Varkonyi-Gasic et al. (2007) for figure permissions as well as the publishers BMC, Elsevier, and Frontiers. Besides, we appreciate Zhang and Wang (2015) and Tripathi et al. (2015) and also the publishers Wiley and Frontiers for allowing us to use table with modifications.

References

Adai A, Johnson C, Mlotshwa S, Archer-Evans S, Manocha V, Vance V, Sundaresan V (2005) Computational prediction of miRNAs in Arabidopsis thaliana. Genome Res 15:78–91

Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ (2008) Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. Curr Biol 18:758–762

Addo-Quaye C, Miller W, Axtell MJ (2009) CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics 25:130–131

Allen E, Xie Z, Gustafson AM, Carrington JC (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. Cell 121:207–221

An J, Lai J, Sajjanhar A, Lehman ML, Nelson CC (2014) miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. BMC Bioinformatics 15:275

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Axtell MJ (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. RNA 19:740–751

Axtell MJ, Westholm JO, Lai EC (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals. Genome Biol 12:221

Barrangou R, Birmingham A, Wiemann S, Beijersbergen RL, Hornung V, Smith A (2015) Advances in CRISPR-Cas9 genome engineering: lessons learned from RNA interference. Nucleic Acids Res 43:3407–3419

Barrera-Figueroa BE, Gao L, Wu Z, Zhou X, Zhu J, Jin H, Liu R, Zhu JK (2012) High throughput sequencing reveals novel and abiotic stress-regulated microRNAs in the inflorescences of rice. BMC Plant Biol 12:132

Barrera-Figueroa BE, Wu Z, Liu R (2013) Abiotic stress-associated microRNAs in plants: discovery, expression analysis, and evolution. Front Biol 8:189–197

Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116:281–297

Basak J, Nithin C (2015) Targeting non-coding RNAs in plants with the CRISPR-Cas technology is a challenge yet worth accepting. Front Plant Sci 6:1001

Bassett AR, Azzam G, Wheatley L, Tibbit C, Rajakumar T, McGowan S, Stanger N, Ewels PA, Taylor S, Ponting CP, Liu JL, Sauka-Spengler T, Fulga TA (2014) Understanding functional miRNA-target interactions in vivo by site-specific genome engineering. Nat Commun 5:4640

Bhardwaj AR, Joshi G, Pandey R, Kukreja B, Goel S, Jagannath A, Kumar A, Katiyar-Agarwal S, Agarwal M (2014) A genome-wide perspective of miRNAome in response to high temperature, salinity and drought stresses in Brassica juncea (Czern) L. PLoS One 9:e92456

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R (2009) QuickGO: a web-based tool for gene ontology searching. Bioinformatics 25:3045–3046

Bonnet E, He Y, Billiau K, Van de Peer Y (2010) TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. Bioinformatics 26:1566–1568

Brousse C, Liu Q, Beauclair L, Deremetz A, Axtell MJ, Bouche N (2014) A non-canonical plant microRNA target site. Nucleic Acids Res 42:5270–5279

Cakir O, Candar-Cakir B, Zhang B (2015) Small RNA and degradome sequencing reveals important microRNA function in Astragalus chrysochlorus response to selenium stimuli. Plant Biotechnol J 14:543–556

Candar-Cakir B, Arican E, Zhang B (2016) Small RNA and degradome deep sequencing reveals drought-and tissue-specific micrornas and their important roles in drought-sensitive and drought-tolerant tomato genotypes. Plant Biotechnol J doi:10.1111/pbi.12533

Cao X, Wu Z, Jiang F, Zhou R, Yang Z (2014) Identification of chilling stress-responsive tomato microRNAs and their target genes by high-throughput sequencing and degradome analysis. BMC Genomics 15:1130

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ami GOH, Web Presence Working Group (2009) AmiGO: online access to ontology and annotation data. Bioinformatics 25:288–289

Carrington JC, Ambros V (2003) Role of microRNAs in plant and animal development. Science 301:336–338

Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8:93–103

Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33:e179

Chorostecki U, Palatnik JF (2014) comTAR: a web tool for the prediction and characterization of conserved microRNA targets in plants. Bioinformatics 30:2066–2067

Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD (2015) miRTarBase 2016: updates to the experimentally validated miRNAtarget interactions database. Nucleic Acids Res 44:D239–D247

Chu Y, Corey DR (2012) RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther 22:271–274

Conesa A, Gotz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:619832

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

da Huang W, Sherman BT, Lempicki RA (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13

da Huang W, Sherman BT, Lempicki RA (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39:W155–W159

Ding D, Zhang L, Wang H, Liu Z, Zhang Z, Zheng Y (2009) Differential expression of miRNAs in response to salt stress in maize roots. Ann Bot 103:29–38

Ding Y, Chen Z, Zhu C (2011) Microarray-based analysis of cadmium-responsive microRNAs in rice (Oryza sativa). J Exp Bot 62:3563–3573

Ding Y, Tao Y, Zhu C (2013) Emerging roles of microRNAs in the mediation of drought stress response in plants. J Exp Bot 64:3077–3086

Ding Q, Zeng J, He XQ (2014) Deep sequencing on a genome-wide scale reveals diverse stage-specific microRNAs in cambium during dormancy-release induced by chilling in poplar. BMC Plant Biol 14:267

Ding X, Ye J, Wu X, Huang L, Zhu L, Lin S (2015) Deep sequencing analyses of pine wood nematode Bursaphelenchus xylophilus microRNAs reveal distinct miRNA expression patterns during the pathological process of pine wilt disease. Gene 555:346–356

Dong H, Lei J, Ding L, Wen Y, Ju H, Zhang X (2013) MicroRNA: function, detection, and bioanalysis. Chem Rev 113:6207–6233

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 38:W64–W70

Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol 3:e39

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:48

Eldem V, Celikkol Akcay U, Ozhuner E, Bakir Y, Uranbey S, Unver T (2012) Genome-wide identification of miRNAs responsive to drought in peach (Prunus persica) by high-throughput deep sequencing. PLoS One 7:e50298

Eldem V, Okay S, Unver T (2013) Plant microRNAs: new players in functional genomics. Turk J Agric For 37:1–21

Fahlgren N, Carrington JC (2010) miRNA target prediction in plants. Methods Mol Biol 592:51–57

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, Carrington JC (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS One 2:e219

Feng J, Liu S, Wang M, Lang Q, Jin C (2014) Identification of microRNAs and their targets in tomato infected with Cucumber mosaic virus based on deep sequencing. Planta 240:1335–1352

Feng H, Sun Y, Wang B, Wang X, Kang Z (2015) Microarray-based identification of conserved microRNA from wheat and their expression profiles response to Puccinia striiformis f. sp. tritici. Can J Plant Pathol 37:82–91

Ferreira TH, Gentile A, Vilela RD, Costa GG, Dias LI, Endres L, Menossi M (2012) microRNAs associated with drought response in the bioenergy crop sugarcane (Saccharum spp.). PLoS One 7:e46703

Gao C, Ju Z, Cao D, Zhai B, Qin G, Zhu H, Fu D, Luo Y, Zhu B (2014) MicroRNA profiling analysis throughout tomato fruit development and ripening reveals potential regulatory role of RIN on microRNAs accumulation. Plant Biotechnol J 13:370–382

Gao C, Ju Z, Cao D, Zhai B, Qin G, Zhu H, Fu D, Luo Y, Zhu B (2015) MicroRNA profiling analysis throughout tomato fruit development and ripening reveals potential regulatory role of RIN on microRNAs accumulation. Plant Biotechnol J 13:370–382

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–D1056

German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, Kannan V, Rymarquis LA, Nobuta K, German R, De Paoli E, Lu C, Schroth G, Meyers BC, Green PJ (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol 26:941–946

German MA, Luo S, Schroth G, Meyers BC, Green PJ (2009) Construction of parallel analysis of RNA ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc 4:356–362

Gomes CP, Cho JH, Hood L, Franco OL, Pereira RW, Wang K (2013) A review of computational tools in microRNA discovery. Front Genet 4:81

Gregory BD, O'Malley RC, Lister R, Urich MA, Tonti-Filippini J, Chen H, Millar AH, Ecker JR (2008) A link between RNA metabolism and silencing affecting *Arabidopsis* development. Dev Cell 14:854–866

Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32:D109–D111
Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158
Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15:509–524
Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res 39:W132–W138
Hackenberg M, Gustafson P, Langridge P, Shi BJ (2015) Differential expression of microRNAs and other small RNAs in barley between water and drought conditions. Plant Biotechnol J 13:2–13
Hou CY, Wu MT, Lu SH, Hsing YI, Chen HM (2014) Beyond cleaved small RNA targets: unraveling the complexity of plant RNA degradome data. BMC Genomics 15:15
Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C (2015) The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res 43:D1057–D1063
Inal B, Turktas M, Eren H, Ilhan E, Okay S, Atak M, Erayman M, Unver T (2014) Genome-wide fungal stress responsive miRNA expression in wheat. Planta 240:1287–1298
Jha A, Shankar R (2011) Employing machine learning for reliable miRNA target identification in plants. BMC Genomics 12:636
Jin W, Wu F (2015) Characterization of miRNAs associated with Botrytis cinerea infection of tomato leaves. BMC Plant Biol 15:1
Jin W, Wu F, Xiao L, Liang G, Zhen Y, Guo Z, Guo A (2012) Microarray-based analysis of tomato miRNA regulated by *Botrytis cinerea*. J Plant Growth Regul 31:38–46
Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 14:787–799
Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol 57:19–53
Kadri S, Hinman V, Benos PV (2009) HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC Bioinformatics 10(Suppl 1):S35
Kanehisa M (2016) KEGG bioinformatics resource for plant genomics and metabolomics. In: Edwards D (ed) Plant bioinformatics. Springer, New York, pp 55–77
Kang W, Friedlander MR (2015) Computational prediction of miRNA genes from small RNA sequencing data. Front Bioeng Biotechnol 3:7
Kantar M, Lucas SJ, Budak H (2011) miRNA expression patterns of Triticum dicoccoides in response to shock drought stress. Planta 233:471–484
Karlova R, van Haarst JC, Maliepaard C, van de Geest H, Bovy AG, Lammers M, Angenent GC, de Maagd RA (2013) Identification of microRNA targets in tomato fruit development using high-throughput sequencing and degradome analysis. J Exp Bot 64:1863–1878
Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6:4
Khraiwesh 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
Kim SW, Li Z, Moore PS, Monaghan AP, Chang Y, Nichols M, John B (2010) A sensitive non-radioactive northern blot method to detect small RNAs. Nucleic Acids Res 38:e98
Knief C (2014) Analysis of plant microbe interactions in the era of next generation sequencing technologies. Front Plant Sci 5:216

Kohli D, Joshi G, Deokar AA, Bhardwaj AR, Agarwal M, Katiyar-Agarwal S, Srinivasan R, Jain PK (2014) Identification and characterization of Wilt and salt stress-responsive microRNAs in chickpea through high-throughput sequencing. PLoS One 9:e108851

Korir NK, Li X, Xin S, Wang C, Changnian S, Kayesh E, Fang J (2013) Characterization and expression profiling of selected microRNAs in tomato (Solanum lycopersicon) 'Jiangshu14'. Mol Biol Rep 40:3503–3521

Koscianska E, Starega-Roslan J, Czubala K, Krzyzosiak WJ (2011) High-resolution northern blot for a reliable analysis of microRNAs and their precursors. Sci World J 11:102–117

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42:D68–D73

Kramer MF (2011) Stem-loop RT-qPCR for miRNAs. Curr Protoc Mol Biol Chapter 15:Unit 15 10

Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet 11:597–610

Kruger J, Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 34:W451–W454

Kumar RR, Pathak H, Sharma SK, Kala YK, Nirjal MK, Singh GP, Goswami S, Rai RD (2014) Novel and conserved heat-responsive microRNAs in wheat (Triticum aestivum L.). Funct Integr Genomics

Lang QL, Zhou XC, Zhang XL, Drabek R, Zuo ZX, Ren YL, Li TB, Chen JS, Gao XL (2011) Microarray-based identification of tomato microRNAs and time course analysis of their response to Cucumber mosaic virus infection. J Zhejiang Univ Sci B 12:116–125

Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843–854

Li H, Dong Y, Yin H, Wang N, Yang J, Liu X, Wang Y, Wu J, Li X (2011) Characterization of the stress associated microRNAs in Glycine max by deep sequencing. BMC Plant Biol 11:170

Li F, Orban R, Baker B (2012a) SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. Plant J 70:891–901

Li Y, Zhang Z, Liu F, Vongsangnak W, Jing Q, Shen B (2012b) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. Nucleic Acids Res 40:4298–4305

Li B, Duan H, Li J, Deng XW, Yin W, Xia X (2013) Global identification of miRNAs and targets in Populus euphratica under salt stress. Plant Mol Biol 81:525–539

Liu HH, Tian X, Li YJ, Wu CA, Zheng CC (2008) Microarray-based analysis of stress-regulated microRNAs in Arabidopsis thaliana. RNA 14:836–843

Liu H, Qin C, Chen Z, Zuo T, Yang X, Zhou H, Xu M, Cao S, Shen Y, Lin H, He X, Zhang Y, Li L, Ding H, Lubberstedt T, Zhang Z, Pan G (2014a) Identification of miRNAs and their target genes in developing maize ears by combined small RNA and degradome sequencing. BMC Genomics 15:25

Liu F, Wang W, Sun X, Liang Z, Wang F (2014b) Conserved and novel heat stress-responsive microRNAs were identified by deep sequencing in Saccharina japonica (Laminariales, Phaeophyta). Plant Cell Environ 38:1357–1367

Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297:2053–2056

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) Elucidation of the small RNA component of the transcriptome. Science 309:1567–1569

Lu S, Sun YH, Chiang VL (2008) Stress-responsive microRNAs in Populus. Plant J 55:131–151

Luan Y, Wang W, Liu P (2014) Identification and functional analysis of novel and conserved microRNAs in tomato. Mol Biol Rep 41:5385–5394

Lv DK, Bai X, Li Y, Ding XD, Ge Y, Cai H, Ji W, Wu N, Zhu YM (2010) Profiling of cold-stress-responsive miRNAs in rice by microarrays. Gene 459:39–47

Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21:3448–3449

Meng Y, Gou L, Chen D, Mao C, Jin Y, Wu P, Chen M (2011) PmiRKB: a plant microRNA knowledge base. Nucleic Acids Res 39:D181–D187

Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. Plant J 67:232–246

Motameny S, Wolters S, Nurnberg P, Schumacher B (2010) Next generation sequencing of miRNAs—strategies, resources and methods. Genes (Basel) 1:70–84

Moxon S, Jing R, Szittya G, Schwach F, Rusholme Pilcher RL, Moulton V, Dalmay T (2008) Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. Genome Res 18:1602–1609

Munoz-Merida A, Perkins JR, Viguera E, Thode G, Bejarano ER, Perez-Pulido AJ (2012) Semirna: searching for plant miRNAs using target sequences. OMICS 16:168–177

Myllykangas S, Buenrostro J, Ji HP (2012) Overview of sequencing technology platforms. In: Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM (eds) Bioinformatics for high throughput sequencing. Springer, New York

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34:D731–D735

Nozawa M, Miura S, Nei M (2012) Origins and evolution of microRNA genes in plant species. Genome Biol Evol 4:230–239

Numnark S, Mhuantong W, Ingsriswang S, Wichadakul D (2012) C-mii: a tool for plant miRNA and target identification. BMC Genomics 13(Suppl 7):S16

Ossowski S, Schwab R, Weigel D (2008) Gene silencing in plants using artificial microRNAs and other small RNAs. Plant J 53:674–690

Ozhuner E, Eldem V, Ipek A, Okay S, Sakcali S, Zhang B, Boke H, Unver T (2013) Boron stress responsive microRNAs and their targets in barley. PLoS One 8:e59543

Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE (2008) Chromatin structure analyses identify miRNA promoters. Genes Dev 22:3172–3183

Pall GS, Hamilton AJ (2008) Improved northern blot method for enhanced detection of small RNA. Nat Protoc 3:1077–1084

Pall GS, Codony-Servat C, Byrne J, Ritchie L, Hamilton A (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. Nucleic Acids Res 35:e60

Pantaleo V, Szittya G, Moxon S, Miozzi L, Moulton V, Dalmay T, Burgyan J (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. Plant J 62:960–976

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. J Appl Genet 52:413–435

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

Pasini L, Bergonti M, Fracasso A, Marocco A, Amaducci S (2014) Microarray analysis of differentially expressed mRNAs and miRNAs in young leaves of sorghum under dry-down conditions. J Plant Physiol 171:537–548

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P, Spring J, Srinivasan A, Fishman M, Finnerty J, Corbo J, Levine M, Leahy P, Davidson E, Ruvkun G (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408:86–89

Pradhan B, Naqvi AR, Saraf S, Mukherjee SK, Dey N (2015) Prediction and characterization of Tomato leaf curl New Delhi virus (ToLCNDV) responsive novel microRNAs in Solanum lycopersicum. Virus Res 195:183–195

Pritchard CC, Cheng HH, Tewari M (2012) MicroRNA profiling: approaches and considerations. Nat Rev Genet 13:358–369

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517
Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403:901–906
Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16:1616–1626
Ren Y, Sun F, Hou J, Chen L, Zhang Y, Kang X, Wang Y (2015) Differential profiling analysis of miRNAs reveals a regulatory role in low N stress response of Populus. Funct Integr Genomics 15:93–105
Rhee S, Chae H, Kim S (2015) PlantMirnaT: miRNA and mRNA integrated analysis fully utilizing characteristics of plant sequencing data. Methods 83:80–87
Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP (2002) Prediction of plant microRNA targets. Cell 110:513–520
Rogers K, Chen X (2013) Biogenesis, turnover, and mode of action of plant microRNAs. Plant Cell 25:2383–2399
Rosas-Cardenas Fde F, Caballero-Perez J, Gutierrez-Ramos X, Marsch-Martinez N, Cruz-Hernandez A, de Folter S (2015) miRNA expression during prickly pear cactus fruit development. Planta 241:435–448
Schreiber AW, Shi BJ, Huang CY, Langridge P, Baumann U (2011) Discovery of barley miRNAs through deep sequencing of short reads. BMC Genomics 12:129
Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D (2005) Specific effects of microRNAs on the plant transcriptome. Dev Cell 8:517–527
Schwab R, Ossowski S, Riester M, Warthmann N, Weigel D (2006) Highly specific gene silencing by artificial microRNAs in Arabidopsis. Plant Cell 18:1121–1133
Shamimuzzaman M, Vodkin L (2012) Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. BMC Genomics 13:310
Si J, Zhou T, Bo W, Xu F, Wu R (2014) Genome-wide analysis of salt-responsive and novel microRNAs in Populus euphratica by deep sequencing. BMC Genet 15(Suppl 1):S6
Song C, Wang C, Zhang C, Korir NK, Yu H, Ma Z, Fang J (2010) Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (Citrus trifoliata). BMC Genomics 11:431
Sprink T, Metje J, Hartung F (2015) Plant genome editing by novel tools: TALEN and other sequence specific nucleases. Curr Opin Biotechnol 32:47–53
Sun G, Stewart CN Jr, Xiao P, Zhang B (2012) MicroRNA expression analysis in the cellulosic biofuel crop switchgrass (Panicum virgatum) under abiotic stress. PLoS One 7:e32017
Sun X, Dong B, Yin L, Zhang R, Du W, Liu D, Shi N, Li A, Liang Y, Mao L (2013) PMTED: a plant microRNA target expression database. BMC Bioinformatics 14:174
Sun R, Wang Q, Ma J, He Q, Zhang B (2014) Differentiated expression of microRNAs may regulate genotype-dependent traits in cotton. Gene 547:233–238
Sunkar R, Zhu JK (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. Plant Cell 16:2001–2019
Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu JK (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol 8:25
Szczesniak MW, Makalowska I (2014) miRNEST 2.0: a database of plant and animal microRNAs. Nucleic Acids Res 42:D74–D77
Szittya G, Moxon S, Santos DM, Jing R, Fevereiro MP, Moulton V, Dalmay T (2008) High-throughput sequencing of Medicago truncatula short RNAs identifies eight new miRNA families. BMC Genomics 9:593
Tang M, Mao D, Xu L, Li D, Song S, Chen C (2014) Integrated analysis of miRNA and mRNA expression profiles in response to Cd exposure in rice seedlings. BMC Genomics 15:835
Thiebaut F, Grativol C, Tanurdzic M, Carnavale-Bottino M, Vieira T, Motta MR, Rojas C, Vincentini R, Chabregas SM, Hemerly AS, Martienssen RA, Ferreira PC (2014) Differential sRNA regulation in leaves and roots of sugarcane under water depletion. PLoS One 9:e93822

Thomson DW, Bracken CP, Goodall GJ (2011) Experimental strategies for microRNA target identification. Nucleic Acids Res 39:6845–6853

Tian Y, Tian Y, Luo X, Zhou T, Huang Z, Liu Y, Qiu Y, Hou B, Sun D, Deng H, Qian S, Yao K (2014) Identification and characterization of microRNAs related to salt stress in broccoli, using high-throughput sequencing and bioinformatics analysis. BMC Plant Biol 14:226

Tripathi A, Goswami K, Sanan-Mishra N (2015) Role of bioinformatics in establishing microRNAs as modulators of abiotic stress responses: the new revolution. Front Physiol 6:286

Unver T, Namuth-Covert DM, Budak H (2009) Review of current methodological approaches for characterizing microRNAs in plants. Int J Plant Genomics 2009:262463

Valoczi A, Hornyik C, Varga N, Burgyan J, Kauppinen S, Havelda Z (2004) Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Res 32:e175

Varallyay E, Burgyan J, Havelda Z (2007) Detection of microRNAs by Northern blot analyses using LNA probes. Methods 43:140–145

Varallyay E, Burgyan J, Havelda Z (2008) MicroRNA detection by northern blotting using locked nucleic acid probes. Nat Protoc 3:190–196

Varkonyi-Gasic E, Wu R, Wood M, Walton EF, Hellens RP (2007) Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. Plant Methods 3:12

Verma SS, Rahman MH, Deyholos MK, Basu U, Kav NN (2014) Differential expression of miRNAs in Brassica napus root following infection with Plasmodiophora brassicae. PLoS One 9:e86648

Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. Cell 136:669–687

Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y (2005) MicroRNA identification based on sequence and structure alignment. Bioinformatics 21:3610–3614

Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC Bioinformatics 10:328

Wang M, Wang Q, Zhang B (2013) Response of miRNAs and their targets to salt and drought stresses in cotton (Gossypium hirsutum L.). Gene 530:26–32

Wei LQ, Yan LF, Wang T (2011) Deep sequencing on genome-wide scale reveals the unique composition and expression patterns of microRNAs in developing pollen of Oryza sativa. Genome Biol 12:R53

Wu HJ, Ma YK, Chen T, Wang M, Wang XJ (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox. Nucleic Acids Res 40:W22–W28

Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, Sheng Sun Z, Shi Q (2013) mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. RNA Biol 10:1087–1092

Wu F, Shu J, Jin W (2014) Identification and validation of miRNAs associated with the resistance of maize (Zea mays L.) to Exserohilum turcicum. PLoS One 9:e87251

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434:338–345

Xie Z, Jia G, Ghosh A (2012) Small RNAs in plants. In: Sunkar R (ed) MicroRNAs in plant development and stress responses. Springer, Heidelberg, pp 1–28

Xie F, Stewart CN Jr, Taki FA, He Q, Liu H, Zhang B (2014) High-throughput deep sequencing shows that microRNAs play important roles in switchgrass responses to drought and salinity stress. Plant Biotechnol J 12:354–366

Xie F, Wang Q, Sun R, Zhang B (2015a) Deep sequencing reveals important roles of microRNAs in response to drought and salinity stress in cotton. J Exp Bot 66:789–804

Xie M, Zhang S, Yu B (2015b) microRNA biogenesis, degradation and activity in plants. Cell Mol Life Sci 72:87–99

Xie F, Jones DC, Wang Q, Sun R, Zhang B (2015c) Small RNA sequencing identifies miRNA roles in ovule and fibre development. Plant Biotechnol J 13(3):355–369

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

Xu D, Guo S, Liu M (2013a) Identification of miRNAs involved in long-term simulated microgravity response in Solanum lycopersicum. Plant Physiol Biochem 66:10–19

Xu Y, Guo M, Shi W, Liu X, Wang C (2013b) A novel insight into gene ontology semantic similarity. Genomics 101:368–375

Xu Y, Guo M, Liu X, Wang C, Liu Y (2014) Inferring the soybean (Glycine max) microRNA functional network based on target gene network. Bioinformatics 30:94–103

Yan J, Zhang H, Zheng Y, Ding Y (2015) Comparative expression profiling of miRNAs between the cytoplasmic male sterile line MeixiangA and its maintainer line MeixiangB during rice anther development. Planta 241:109–123

Yang X, Li L (2011) miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics 27:2614–2615

Yanik H, Turktas M, Dundar E, Hernandez P, Dorado G, Unver T (2013) Genome-wide identification of alternate bearing-associated microRNAs (miRNAs) in olive (Olea europaea L.). BMC Plant Biol 13:10

Yi X, Zhang Z, Ling Y, Xu W, Su Z (2015) PNRD: a plant non-coding RNA database. Nucleic Acids Res 43:D982–D989

Yin Z, Li Y, Yu J, Liu Y, Li C, Han X, Shen F (2012) Difference in miRNA expression profiles between two cotton cultivars with distinct salt sensitivity. Mol Biol Rep 39:4961–4970

Yin F, Gao J, Liu M, Qin C, Zhang W, Yang A, Xia M, Zhang Z, Shen Y, Lin H, Luo C, Pan G (2014) Genome-wide analysis of water-stress-responsive microRNA expression profile in tobacco roots. Funct Integr Genomics 14:319–332

Yu B, Bi L, Zheng B, Ji L, Chevalier D, Agarwal M, Ramachandran V, Li W, Lagrange T, Walker JC, Chen X (2008) The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. Proc Natl Acad Sci U S A 105:10073–10078

Yu X, Wang H, Lu Y, de Ruiter M, Cariaso M, Prins M, van Tunen A, He Y (2012) Identification of conserved and novel microRNAs that are responsive to heat stress in Brassica rapa. J Exp Bot 63:1025–1038

Yu R, Wang Y, Xu L, Zhu X, Zhang W, Wang R, Gong Y, Limera C, Liu L (2015) Transcriptome profiling of root microRNAs reveals novel insights into taproot thickening in radish (Raphanus sativus L.). BMC Plant Biol 15:30

Zamboni A, Zanin L, Tomasi N, Pezzotti M, Pinton R, Varanini Z, Cesco S (2012) Genome-wide microarray analysis of tomato roots showed defined responses to iron deficiency. BMC Genomics 13:101

Zhang B (2015) MicroRNA: a new target for improving plant tolerance to abiotic stress. J Exp Bot 66:1749–1761

Zhang B, Wang Q (2015) MicroRNA-based biotechnology for plant improvement. J Cell Physiol 230:1–15

Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA (2005) Identification and characterization of new plant microRNAs using EST analysis. Cell Res 15:336–360

Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA (2006) Conservation and divergence of plant microRNA genes. Plant J 46:243–259

Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z (2010) PMRD: plant microRNA database. Nucleic Acids Res 38:D806–D813

Zhang T, Zhao X, Wang W, Pan Y, Huang L, Liu X, Zong Y, Zhu L, Yang D, Fu B (2012) Comparative transcriptome profiling of chilling stress responsiveness in two contrasting rice genotypes. PLoS One 7:e43274

Zhang S, Yue Y, Sheng L, Wu Y, Fan G, Li A, Hu X, Shangguan M, Wei C (2013) PASmiR: a literature-curated database for miRNA molecular regulation in plant response to abiotic stress. BMC Plant Biol 13:33

Zhang N, Yang J, Wang Z, Wen Y, Wang J, He W, Liu B, Si H, Wang D (2014a) Identification of novel and conserved microRNAs related to drought stress in potato by deep sequencing. PLoS One 9:e95489

Zhang Y, Zhu X, Chen X, Song C, Zou Z, Wang Y, Wang M, Fang W, Li X (2014b) Identification and characterization of cold-responsive microRNAs in tea plant (Camellia sinensis) and their targets using high-throughput sequencing and degradome analysis. BMC Plant Biol 14:271

Zhang Z, Jiang L, Wang J, Gu P, Chen M (2015) MTide: an integrated tool for the identification of miRNA-target interaction in plants. Bioinformatics 31:290–291

Zhao B, Liang R, Ge L, Li W, Xiao H, Lin H, Ruan K, Jin Y (2007) Identification of drought-induced microRNAs in rice. Biochem Biophys Res Commun 354:585–590

Zhao CZ, Xia H, Frazier TP, Yao YY, Bi YP, Li AQ, Li MJ, Li CS, Zhang BH, Wang XJ (2010) Deep sequencing identifies novel and conserved microRNAs in peanuts (Arachis hypogaea L.). BMC Plant Biol 10:3

Zhao M, Tai H, Sun S, Zhang F, Xu Y, Li WX (2012) Cloning and characterization of maize miRNAs involved in responses to nitrogen deficiency. PLoS One 7:e29669

Zheng LL, Qu LH (2015) Application of microRNA gene resources in the improvement of agronomic traits in rice. Plant Biotechnol J 13:329–336

Zheng Q, Wang XJ (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36:W358–W363

Zhou M, Luo H (2013) MicroRNA-mediated gene regulation: potential applications for plant genetic engineering. Plant Mol Biol 83:59–75

Zhou X, Su Z (2007) EasyGO: gene ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics 8:246

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

Zhou ZS, Song JB, Yang ZM (2012) Genome-wide identification of Brassica napus microRNAs and their targets in response to cadmium. J Exp Bot 63:4597–4613

Zhu YY, Zeng HQ, Dong CX, Yin XM, Shen QR, Yang ZM (2010) microRNA expression profiles associated with phosphorus deficiency in white lupin (*Lupinus albus* L.). Plant Sci 178:23–29

Zhuang Y, Zhou X, Liu J (2014) Conserved miRNAs and their response to salt stress in wild eggplant Solanum linnaeanum roots. Int J Mol Sci 15:839–849

Identification of Gene Families Using Genomics and/or Transcriptomics Data

Sezer Okay

Contents

Abstract Thousands of putative open reading frames (ORFs) are identified via annotation of sequenced plant genomes. Classification of these ORFs into gene families has a crucial importance to understand the evolution, function, and structure of the encoded proteins such as transcription factors, and the non-coding RNAs such as microRNAs (miRNAs). Thus, molecular mechanisms underlying the metabolic processes in plants are uncovered as well. Some members of the gene families are species-specific being more dynamic during evolution whereas others are more conserved, phylogenetically sharing common features. The latter are especially important for the annotation of putative ORFs by revealing known counterparts with high sequence identity via sequence alignment to discover conserved motifs. Various bioinformatic tools are available to find out gene families in plants. The BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi) is widely used for identification of homologous sequences. Phytozome (http://www.phytozome.net) or GreenPhyl (http://www.greenphyl.org) are the web resources utilized for the functional and comparative genomics in plants to analyze gene families. TRAPID (http://bioinformatics.psb.ugent.be/webtools/trapid) offers a free of charge web

S. Okay (✉)
Department of Biology, Faculty of Science, Çankırı Karatekin University, Çankırı 18100, Turkey
e-mail: sezerokay@gmail.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_9

source for functional and comparative analyses of transcriptome data sets for identification of gene families, alignment of multiple sequences and phylogenetic tree construction. Some of the databases store specific type of gene families such as plant transcription factor databases PlantTFDB (http://planttfdb.cbi.pku.edu.cn) and PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v3.0), or miRBase (http://www.mirbase.org) for miRNAs. Molecular Evolutionary Genetics Analysis (MEGA) software is an integrated tool for the analyses such as alignment of sequence, construction of phylogenetic trees, and access to online databases. In this chapter, the bioinformatic tools for analyses of genomics and/or transcriptomics data sets to discover gene families as well as sample researches are discussed.

Keywords Bioinformatics • Gene family • Genome • Phylogeny • RNA-seq • Transcriptome

1 Introduction

The high-throughput next-generation sequencing (NGS) technologies produce a large amount of nucleotide sequences. Complete genome sequences and RNA-seq analysis of many plants utilizing NGS have been published, and a vast quantity of sequences are stored at the publicly available databases such as NCBI GenBank (http://www.ncbi.nlm.nih.gov/nuccore) or Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra). Identification of all elements in a genome or transcriptome is highly time-consuming and laborious. Therefore, the databases still contain sequences waiting to be annotated to find out their functions. In silico genome-wide analysis of these sequences using various bioinformatic tools is currently a hot research topic. Among the genome/transcriptome annotation products, one of the main groups is the gene families (Mochida and Shinozaki 2011; Martinez 2013).

A plant genome includes thousands of protein-coding genes. Although many genes posses distinct features, a large number of genes share homology in terms of sequence structure, which are classified as gene families (Frech and Chen 2010). Genome-wide comparative analysis of gene families in plants results in the identification of their distribution pattern, orthology and synteny status, and phylogenic relationships (Chanroj et al. 2012; Rawal et al. 2013; Hofberger et al. 2014; Pan et al. 2015). Moreover, comprehensive analysis of gene families may reveal a new function (Wang et al. 2015a), a new family (Saito et al. 2014), or an evolutionary history among all eukaryotes: plants, animals, and fungi (Li et al. 2014b).

Comparative analysis of the transcriptome data produced by the high-throughput sequencing of cDNA libraries (RNA-seq) from a plant exposed to different conditions and/or from differing plant organs gives important clues about the biotic/abiotic stress and hormone responses (Nawaz et al. 2014; Okay et al. 2014; Kim et al. 2015) as well as the plant development (Ha et al. 2014; Chettoor et al. 2014; Jali

et al. 2014). The RNA-seq raw data are reconstructed via mapping the reads to a complete reference genome; however, if the reference genome is partial or absent, two strategies can be utilized. In mapping-first approach, the sequence reads are mapped to an unannotated reference genome, and then the overlapping sequences are merged. On the other hand, in assembly-first approach (de novo assembly), the sequence reads are assembled directly, and then the assembly may be mapped to a reference genome, if available (Grabherr et al. 2011). For instance, Newbler software assembles the Roche 454 GS reads de novo; however, the lack of a reference genome results in the loss of some data as unassembled.

2 Bioinformatic Tools for Identification of Gene Families

Currently, a broad range of bioinformatic tools are present to process the excess amount of data produced by high-throughput sequencing for identification of gene families. First of all, the genome and/or transcriptome sequence data are required for the analyses. If previously produced sequence reads are to be used, the data can be retrieved from a database. Generally, the genome databases store information specific for an organism; however, some of the databases contain data for varying species.

The National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) portal is one of the most common sources for bioinformatics analyses. NCBI is a division of the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). NCBI server provides diverse resources for storing and analyzing information about genetics, biochemistry, and molecular biology (Fig. 1). Among these resources, the Nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore) is a collection of genome, gene, and transcript sequence data. The Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) stores raw sequencing data and alignment data from high-throughput next-generation sequencing platforms. NCBI also provides the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi) for query of nucleotide and amino acid sequences.

The Phytozome database (Goodstein et al. 2012; http://phytozome.jgi.doe.gov/pz/portal.html), a joint project of the Center for Integrative Genomics and the Joint Genome Institute (JGI), is also frequently used for retrieval of genome data belonging to green plants (Fig. 2). The current version 10.1 of Phytozome includes the genome data of 48 plants and the gene families clustered at 12 phylogenetically important nodes. Additionally, this database contains the information about the genes, gene families, diversity, and expression data for 52 plant genomes. A detailed guide for the utilization of Phytozome can be reached from http://phytozome.jgi.doe.gov/pz/QuickStart.html.

The conserved domains representing the protein families should be identified to search the genome/transcriptome sequence. The Protein families (Pfam) database (Finn et al. 2014; http://pfam.xfam.org) is widely used for the identification of

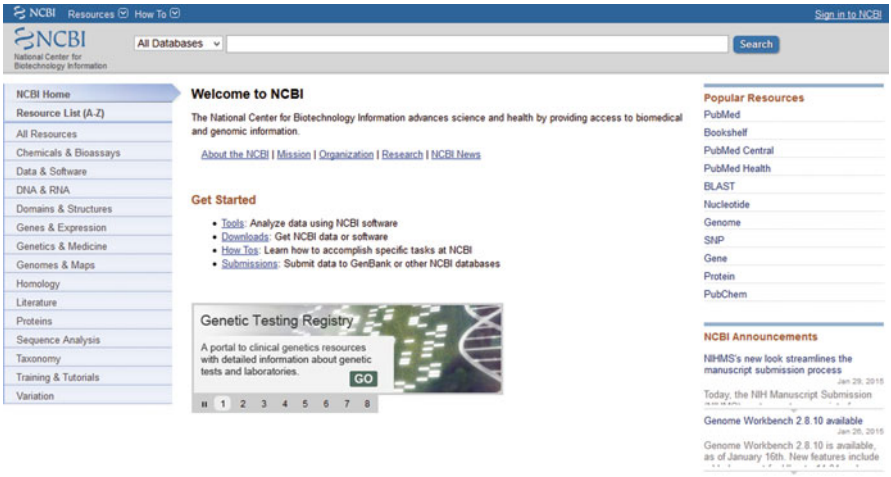


Fig. 1 The homepage of the NCBI database

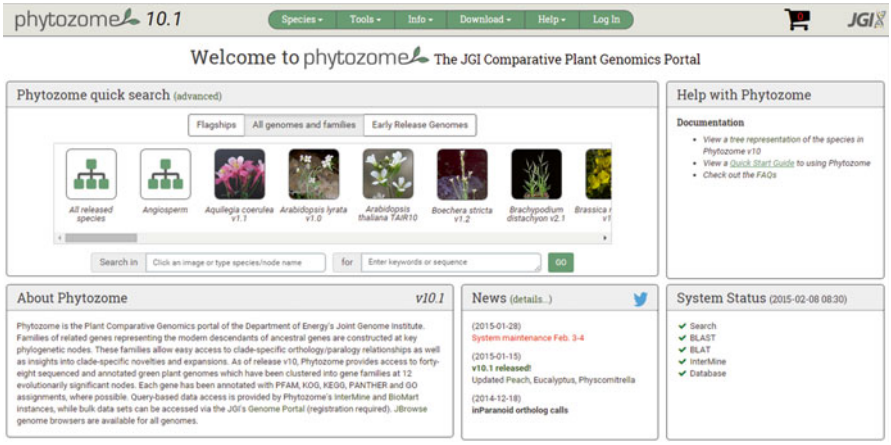


Fig. 2 The homepage of the Phytozome database v10.1

protein families and clans, the higher-level groupings of related families (Fig. 3). In Pfam database, the protein families are represented by multiple sequence alignments and hidden Markov models (HMMs). There are two components of Pfam, Pfam-A and Pfam-B. Pfam-A covers manually curated high quality data whereas Pfam-B includes automatically generated lower quality entries.

Another database for the domain analysis is the Simple Modular Architecture Research Tool, SMART (Letunic et al. 2015; http://smart.embl.de), which contains two modes according to the protein database used (Fig. 4). The Normal SMART covers SP-TrEMBL, Swiss-Prot, and stable Ensembl proteomes whereas the Genomic SMART contains Ensembl for metazoans and Swiss-Prot for the rest.

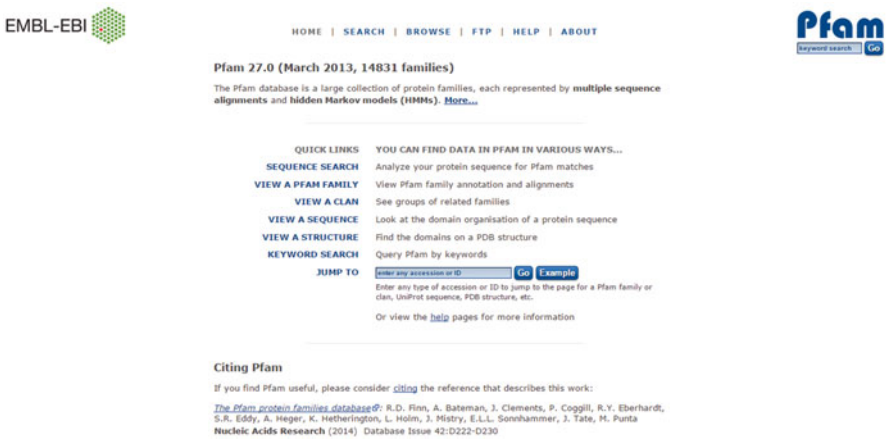


Fig. 3 The homepage of the Pfam database v27.0

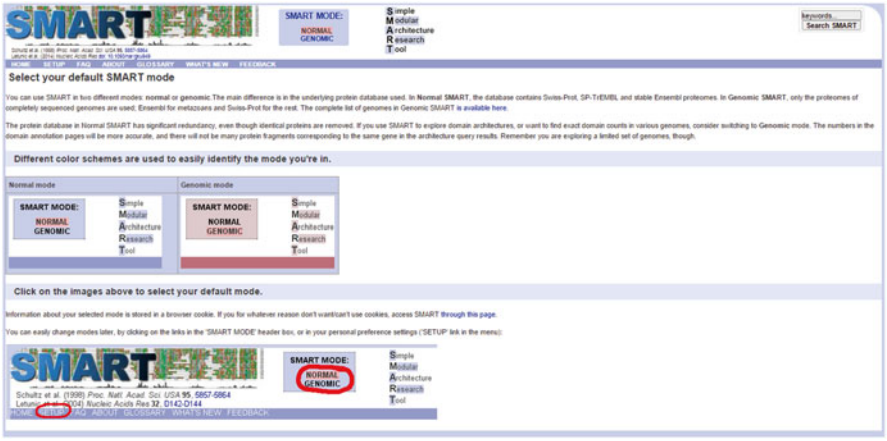


Fig. 4 The homepage of the SMART database

Following the determination of gene/protein families, multiple sequence alignment is performed. The Clustal (Sievers et al. 2011; http://www.clustal.org) is a widely used tool for multiple sequence alignment (Fig. 5). It has two versions: ClustalW2 and Clustal Omega. (1) ClustalW2 (ClustalW/X). ClustalW is the command line version and ClustalX is the graphical version. (2) Clustal Omega is the latest addition offering higher scalability, velocity, and quality. Only command line/web server of Clustal Omega is in use currently. The Clustal can be utilized online from the European Bioinformatics Institute (EMBL-EBI) website (http://www.ebi.ac.uk/Tools/msa).

The Multiple Sequence Comparison by Log-Expectation (MUSCLE) program (Edgar 2004; http://drive5.com/muscle) is also used for multiple sequence alignment.

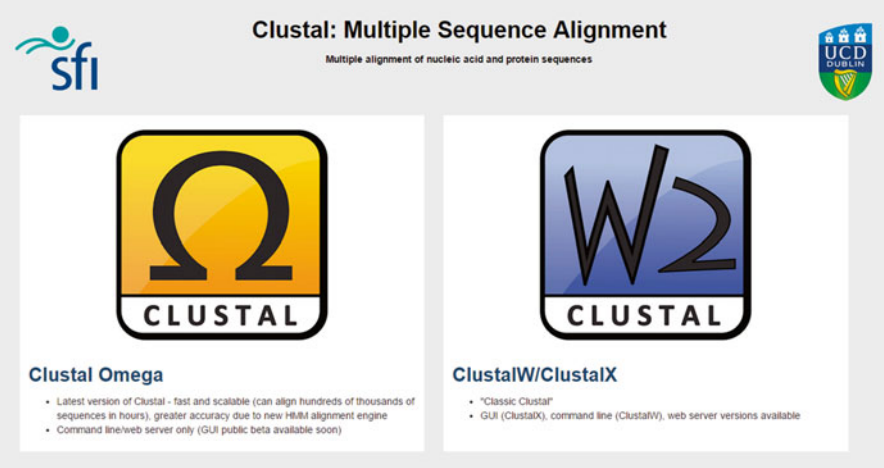


Fig. 5 The homepage of the Clustal tool

The MUSCLE is faster than Clustal and can be operated using Linux, Windows, Windows with Cygwin, and Mac OSX. Additionally, an online tool is available at the EMBL-EBI website (http://www.ebi.ac.uk/Tools/msa/muscle).

The phylogenetic relationships of the members belonging to a gene family are investigated as well. The Molecular Evolutionary Genetics Analysis (MEGA) software (Tamura et al. 2013; http://www.megasoftware.net) is a common tool for phylogenetic analyses. Currently, MEGA v6.0 is accessible (Fig. 6), and can be utilized for diverse analyses including sequence alignment, construction of phylogenetic trees, online database search, estimation of divergence time, and molecular evolution rate.

The conserved motifs representing the gene families are also identified. The motif-based sequence analysis tool package MEME Suite (Bailey et al. 2009; http://meme.nbcr.net/meme), developed and maintained by seven institutes including National Center for Research Resources and National Biomedical Computation Resource, is widely used for this purpose. Currently, version 4.9.1 of the package includes 13 different tools for discovery, alignment, comparison, enrichment, and annotation of the conserved motifs (Fig. 7).

The Expert Protein Analysis System (ExPASy) portal (Artimo et al. 2012; http://www.expasy.org), launched by the SIB (Swiss Institute of Bioinformatics), is frequently used to analyze the physicochemical properties such as molecular weight (Mw) and isoelectric point (pI) of the proteins. Although previous version of the portal was utilized only for protein analysis, the current bioinformatics resources portal includes many databases and software tools useful for the different areas of life sciences (Fig. 8).

The most widely used tools are mentioned above; however, there are various additional tools for diverse analyses to identify the gene families. The bioinformatic

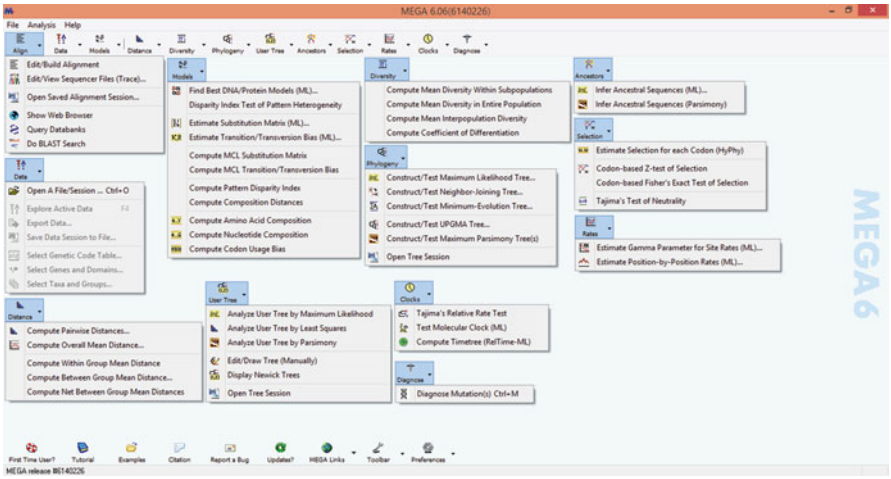


Fig. 6 The menu of the MEGA v6.06. The menu options were placed mixed due to the insufficient space

tools used in the most recent studies on the analysis of gene families in plants are mentioned under the next title.

3 Identification of Gene Families in Plants

3.1 Identification of Gene Families in Monocots

3.1.1 Poaceae (Gramineae)

Rice

Gene families in rice (*Oryza sativa*) have been identified both genome-wide to show their distribution on the chromosomes, and transcriptome-wide to determine their expression patterns related with stress conditions and developmental stages. Nguyen et al. (2014) identified 16 *Catharanthus roseus* receptor-like kinase1-like kinase (*CrRLK1L*) genes in rice genome using the Greenphyl and Rice Genome Annotation Project (RGAP; http://rice.plantbiology.msu.edu) databases. Moreover, the protein, genomic, and cDNA accession numbers were retrieved from NCBI, the clone name from Knowledge-based Oryza Molecular biological Encyclopedia (KOME; http://cdna01.dna.affrc.go.jp/cDNA), and GO terms from the Rice Oligonucleotide Array Database (http://www.ricearray.org). Additionally, the ortholog names were collected from The Arabidopsis Information Resource database (TAIR; http://www.Arabidopsis.org).

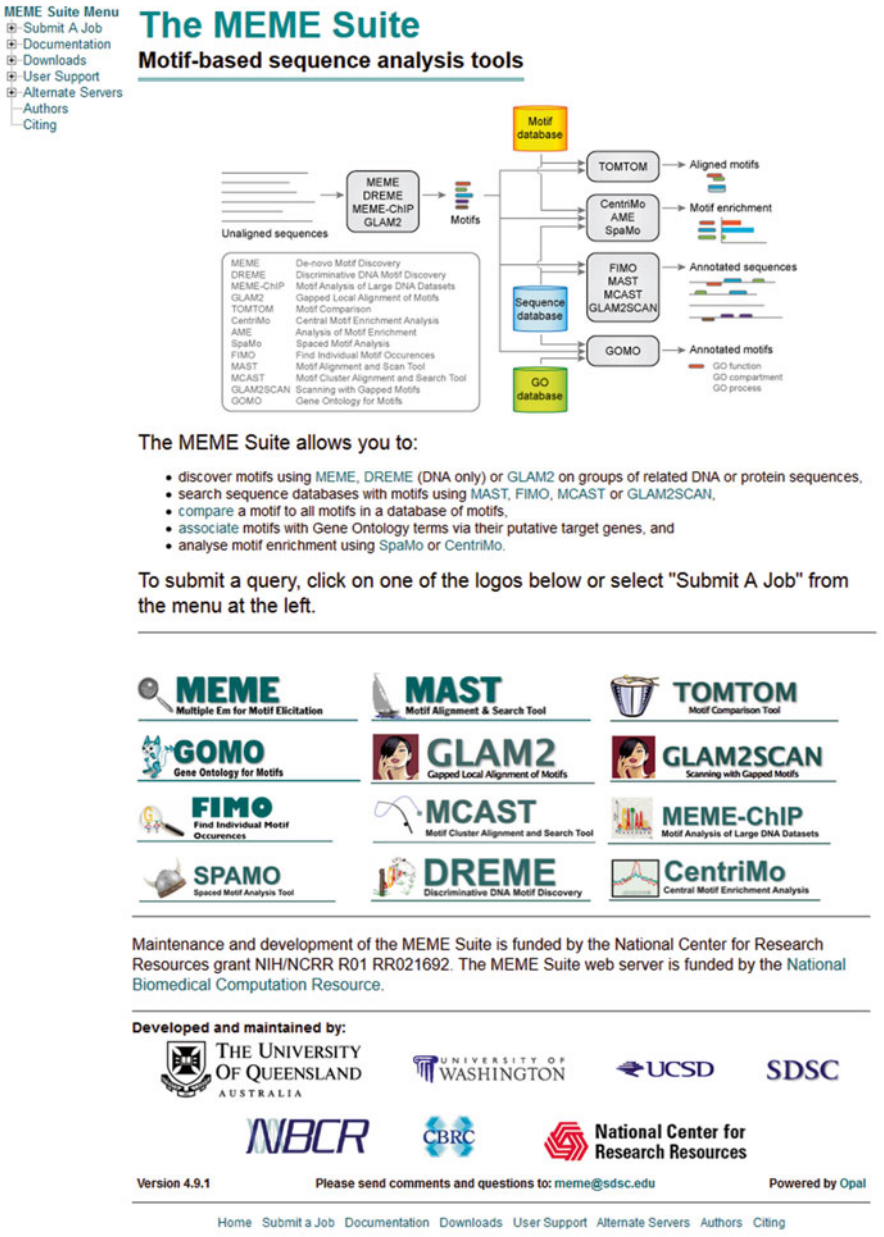


Fig. 7 The homepage of the MEME Suite tool

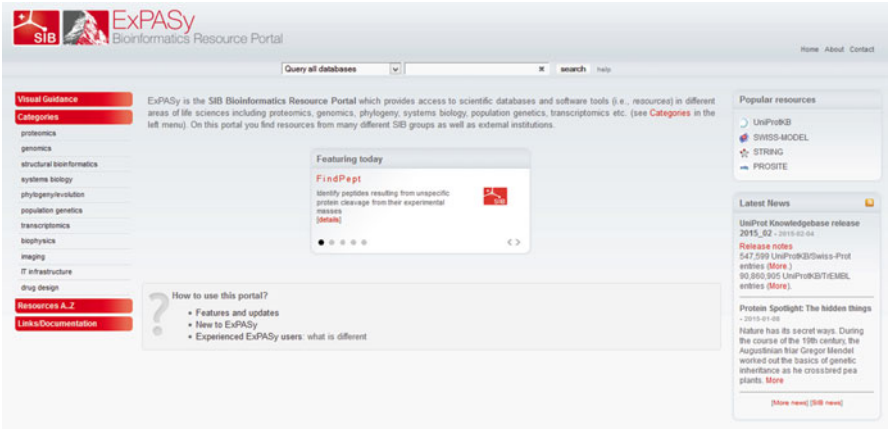


Fig. 8 The homepage of the ExPASy bioinformatics resource portal

Ma et al. (2014c) identified 21 xylogen-like arabinogalactan protein (*XYLP*) genes from the rice genome using the RGAP database. The presence of nonspecific lipid transfer protein-like (nsLTP) domains, N-terminal signal peptide, GPI-anchored signal, and N-glycosylation sites were predicted via InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan5), SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP), Big-PI Plant Predictor (http://mendel.imp.ac.at/gpi/plant_server.html), and NetNGlyc 1.0 Server (http://www.cbs.dtu.dk/services/NetNGlyc), respectively.

Similarly, the drought-induced 19 (Di19) family of transcription factors in rice genome were analyzed by Wang et al. (2014e) using the RGAP and Rice Annotation Project (RAP-DB; http://rapdb.dna.affrc.go.jp/index.html) databases. The presence of zf-Di19 and Di19_C domains was determined using InterProScan and Pfam servers, the physicochemical parameters of each gene were predicted using ExPASy, and the phylogenetic analyses were performed via MEGA software.

Moreover, de Oliveira et al. (2014) identified the family of wall-associated kinases (WAKs) in rice genomes using RGAP and Gramene (http://www.gramene.org) databases. The protein domains were analyzed via Pfam and SMART, and splicing predictions were performed using GENSCAN (http://genes.mit.edu/GENSCAN.html). The sequence alignment, similarity clustering, and phylogenetic analyses were performed via MUSCLE, Circoletto (Darzentas 2010), and MEGA software, respectively.

Zhang et al. (2015a) identified three members of rice CCT [CONSTANS (CO), CO-LIKE, and TIMING OF CAB1 (TOC1)] family, taking role in flowering, through the Gramene, RGAP, NCBI, and DRTF (http://drtf.cbi.pku.edu.cn/index.php) databases. Nucleotide diversity analysis and candidate gene-based association mapping were performed via RiceVarMap (http://ricevarmap.ncpgr.cn).

Nawaz et al. (2014) analyzed the cyclic nucleotide-gated ion channel (*CNGC*) gene family and their expression patterns in response to plant hormones and biotic

and abiotic stresses in rice. The *Arabidopsis CNGC* gene sequences were obtained from the TAIR database to find out the rice homologs in RGAP and RAP-DB databases. The non-redundant sequences were collected from Phytozome database, and the domain analysis was performed using the Pfam, SMART, CDD (http://www.ncbi.nlm.nih.gov/cdd), PROSITE (http://prosite.expasy.org), SUPERFAMILY (http://www.supfam.org/SUPERFAMILY), and Gene3D (http://gene3d.biochem.ucl.ac.uk). The chromosomal location and duplication status of the genes were determined using the Rice TOGO Browser (http://agri-trait.dna.affrc.go.jp) and the Plant Genome Duplication Database (PGDD; http://chibba.agtec.uga.edu/duplication), respectively. Exon–intron distribution and *cis*-acting regulatory elements were analyzed using Gene Structure Display Server (GSDS; http://gsds.cbi.pku.edu.cn) and the Plant *Cis*-acting Regulatory DNA Elements (PLACE; http://www.dna.affrc.go.jp/PLACE) database, respectively. The protein sequences were analyzed via ExPASy, their cellular localizations were determined using PSORT (http://psort.hgc.jp), and the posttranslational modifications were predicted via PROSITE tool. The amino acid sequences were aligned using MUSCLE and MAFFT (http://mafft.cbrc.jp/alignment/software), the phylogenetic analysis was performed via MEGA, and the conserved motifs were determined using MAST (http://meme.nbcr.net/meme/cgi-bin/mast.cgi) search tool.

Additionally, Saha et al. (2015) identified rice ATP-binding cassette (ABC) transporter family, taking role in salt stress, using RGAP and GenBank (http://www.ncbi.nlm.nih.gov/genbank) databases. The protein domains and conserved motifs were predicted via SMART and MEME tools, respectively. The phylogenetic analyses were performed using MEGA software, the genes were mapped to the rice chromosomes via Massively Parallel Signature Sequencing database (MPSS; http://mpss.udel.edu/in9311/mpss_index.php), and the expression patterns of the genes during salinity stress were investigated using Genevestigator (https://genevestigator.com/gv) database.

Maize

The gene families in maize have also been analyzed using both genome- and transcriptome-wide methodologies. The protein phosphatase family in maize was identified by Wei and Pan (2014) using Maize Genome Sequence Project (http://ftp.maizesequence.org/release-5b/filtered-set). The hidden Markov model (HMM) profiles and the catalytic domains of the proteins were predicted via HMMER (http://hmmer.janelia.org), Pfam, and SMART. The subcellular localization and pI of the proteins were determined using PSORT and ExPAsy, respectively. The phylogenetic analysis was performed using MEGA. The gene structure, chromosomal location, gene duplication, synteny, and sequence polymorphism analyses were performed via GSDS, MapDraw (Liu and Meng 2003), SyMAP (http://www.symapdb.org), and DnaSP (http://www.ub.edu/dnasp), respectively. The putative *cis*-acting elements in the promoters, and the microRNAs (miRNAs) were identified using PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html) and miRBase (http://www.mirbase.org), respectively.

The expansin gene family in maize genome and their expression in maize endosperm were investigated by Zhang et al. (2014a) using the MaizeSequence (http://ftp.maizesequence.org/ current) database. The protein domains and motifs were determined using Pfam, SMART, and MEME. The pIs, signal peptide cleavage sites, and the gene ontology (GO) terms were predicted using ExPASy, SignalP, and ProtFun (http://www.cbs.dtu.dk/services/ProtFun) servers, respectively. The phylogenetic, gene structure, and chromosomal location analyses were performed via MEGA, GSDS, and MapInspect (http://mapinspect.software.informer.com), respectively. The hormone-responsive elements in promoter regions were predicted using PlantCARE and PLACE.

Shiriga et al. (2014) investigated the genome-wide distribution of NAC family of transcription factors, and their expression pattern under drought stress in maize using the Plant Transcription Factor Database (PlnTFDB; http://plntfdb.bio.uni-potsdam.de). The protein domains were determined using Pfam, SMART, and InterProScan. The phylogenetic analysis was performed using MEGA as well as the membrane-bound proteins and the motifs were predicted via TMHMM (http://www.cbs.dtu.dk/services/TMHMM), and MEME servers, respectively.

Similarly, Fan et al. (2014) identified the NAC transcription factors in maize using the Phytozome database. The Pfam and SMART were used to predict protein domains. The gene and protein structures were analyzed via GSDS, ProtParam (http://web.expasy.org/protparam), and SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html), respectively. The chromosomal locations and conserved motifs were determined via MapInspect and MEME, respectively. The orthologous sequences were analyzed using OrthoMCL (http://www.orthomcl.org/orthomcl) and the data for NAC gene expression in various maize tissues and developmental stages were obtained from Genevestigator.

Chettoor et al. (2014) analyzed the RNA-seq of different reproductive organs of maize to find out the gametophyte functions, and small signaling proteins and various transcription factor gene families were identified. The data from different samples were determined using the Venny (http://bioinfogp.cnb.csic.es/tools/venny) and the Venn Diagrams were obtained using the BioInfoRx (http://apps.bioinforx.com/bxaf6/tools/app_overlap.php) tools. The GO terms, small peptide genes, and the transcription factors were identified using AgriGO (http://bioinfo.cau.edu.cn/agriGO) toolkit, MaizeSequence (updated link: http://ensembl.gramene.org/Zea_mays/Info/Index) database, and Grassius (http://grassius.org/grasstfdb.html) server. The sequences were aligned using MegAlign (DNASTAR; http://www.dnastar.com), and the phylogenetic analysis was performed using MrBayes (http://mrbayes.sourceforge.net).

Distribution of auxin-responsive *GH3* family genes in maize genome and their expression under abiotic stresses were analyzed by Feng et al. (2014b) using MaizeGDB (http://www.maizegdb.org). The protein domains were predicted using Pfam and InterProScan. The phylogenetic, gene structure, and synteny analyses were performed via MEGA, GSDS, and SyMAP, respectively.

Wheat

The distribution of nucleotide binding site–leucine-rich repeats (*NBS-LRR*) disease resistance genes in wheat (*Triticum aestivum*) genome was identified by Bouktila et al. (2015) using the NCBI database for wheat (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=CALP01) and analyzed via European Molecular Biology Open Software Suite (EMBOSS; http://www.ebi.ac.uk/Tools/emboss). The sequences were aligned using MUSCLE, and the HMM profiles were predicted via HMMER. The NBS sequences were obtained from EMBOSS, merged via DNA Baser sequence assembler (http://www.dnabaser.com), and the redundant contigs were checked using CD-HIT Suite (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-est). The candidate wheat NBS-encoding R genes were identified using FGENESH (http://www.softberry.com) and Pfam. The structural domain and orthologous cluster analyses were performed via Geneious (http://www.geneious.com) and InterProScan, and OrthoMCL database, respectively.

Thomas et al. (2014) investigated the Methyltransferase 1 (*MET1*) gene family in hexaploid wheat using the International Wheat Genome Sequencing Consortium (IWGSC; http://www.wheatgenome.org) and URGI (https://urgi.versailles.inra.fr) databases. The transposable elements were analyzed using TREP (http://wheat.pw.usda.gov/ITMI/Repeats) database. Gene structures and protein domains were identified using SIM4 (http://pbil.univ-lyon1.fr) and FancyGene (http://bio.ieo.eu/fancygene) tools, and Pfam, respectively. The sequences were aligned using MUSCLE, phylogenetic trees were constructed via Interactive Tree of Life (iTOL; http://itol.embl.de), and the exonic sequences were identified using FGENESH. For RNA-seq analysis, the cDNAs were translated using Transeq and Sixpack from the EMBOSS package. The sequence reads were aligned using TopHat (http://ccb.jhu.edu/software/tophat/index.shtml) and Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), and analyzed using Cufflinks (http://cole-trapnell-lab.github.io/cufflinks).

Ariyarathna et al. (2014) characterized the multigene family *TaHKT 2;1* in bread wheat using IWGSS database. The gene structures were predicted using Spidey (http://www.ncbi.nlm.nih.gov/spidey), GeneSeqer (http://www.plantgdb.org/cgi-bin/GeneSeqer/index.cg), and SIM4 software. Sequence and phylogenetic analyses were conducted using Geneious and MEGA, respectively. The structural and functional analyses of the proteins were performed via TMHMM and Membrane Protein Explorer (MPEx; http://blanco.biomol.uci.edu/mpex) tools. 3D structures of the proteins were predicted using PHYRE 2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). The *cis*-acting elements were analyzed via PlantCARE and PLACE databases.

Okay et al. (2014) identified the superfamily of WRKY transcription factors in bread wheat using Plant Transcription Factor Database (PlantTFDB; http://planttfdb.cbi.edu.cn) and NCBI databases. The HMM profiles were predicted using Pfam, the phylogenetic analysis was performed via MEGA, and the conserved motifs were determined using MEME Suite tool. To identify the drought-responsive WRKY members, the RNA-seq data were retrieved from SRA. The proteins were detected using PlantTFDB and analyzed via ExPASy portal.

Barley

Genome-wide characterization of the basic leucine zipper (bZIP) family of transcription factors in barley (*Hordeum vulgare*) and their expression patterns were investigated by Pourabed et al. (2015) using the PlantTFDB and the International Barley Sequencing Consortium (IBSC) IPK BLAST server (http://webblast.ipk-gatersleben.de/barley). The HMM profiles and conserved domains were analyzed using HMMER, Pfam, and SMART. The sequences were aligned and phylogenetic trees were constructed using ClustalX and MEGA, respectively. The bZIP motifs were predicted via MEME, and the gene structures were identified using GSDS. The UTR sequences were retrieved from Ensembl Plants database (http://plants.ensembl.org/index.html), the internal ribosome entry sites (IRES), and sRNA targets were predicted via UTRdb (http://utrdb.ba.itb.cnr.it) and psRNATarget (http://plantgrn.noble.org/psRNATarget) servers, respectively. The *cis*-regulatory elements were identified using Plant Promoter Analysis Navigator (PlantPAN; http://plantpan.mbc.nctu.edu.tw) database.

Pandey et al. (2014) identified the heat shock protein 20 (*HSP20*) gene family in wheat and barley as well as their expression pattern under heat stress using NCBI and Uniprot (http://www.uniprot.org) databases. The HMM profiles were analyzed using Pfam and HMMER. Open reading frames were predicted via ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The conserved HSP20 motifs and domains were analyzed using MEME and InterProScan, respectively. The physicochemical properties and the subcellular localization of the proteins were predicted using ProtParam, and ESLpred (http://www.imtech.res.in/raghava/eslpred/submit.html) and ProtComp (http://linux1.softberry.com/cgi-bin/programs/proloc/protcomppl.pl) servers, respectively. Simultaneous Alignment and Tree Construction using Hidden Markov Models - Jump Start (SATCHMO-JS (http://phylogenomics.berkeley.edu/q/satchmo) and PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php) servers were used for sequence alignment, and the phylogenetic tree was constructed using MEGA.

Brachypodium distachyon

The IQ67 domain (*IQD*) and growth-regulating factor (*GRF*) gene families found in *Brachypodium distachyon* genome were investigated by Filiz et al. (2013) and (2014), respectively using NCBI and Phytozome databases. The conserved domains were identified using Pfam and SMART servers. The sequences were aligned via ClustalW, motifs were predicted using MEME, and the phylogenetic tree was constructed using MEGA. The gene structures, ORFs, and physicochemical characteristics of the proteins were identified using GSDS, ORF finder, and ProtParam, respectively. The GO terms were annotated using Gramene resource.

Wen et al. (2014) identified the WRKY family of transcription factors in *B. distachyon* genome utilizing PlantTFDB, GramineaeTFDB (http://gramineaetfdb.psc.riken.jp), *B. distachyon* genome database (http://www.brachypodium.org),

Phytozome, UniProt, and NCBI databases. The reliability of the results was verified using UniProt, SMART, and the Brachy WRKY Database (http://www.igece.org/WRKY/BrachyWRKY/BrachyWRKYIndex.html). The sequences were aligned using ClustalW and the phylogenetic tree was constructed via MEGA. The conserved motifs and *cis*-acting elements were analyzed utilizing MEME and PlantCARE, respectively.

Wei et al. (2014a) analyzed the MADS-box gene family in *B. distachyon* genome using *B. distachyon* genome database. The sequences were aligned via ClustalX, and the phylogenetic tree was constructed using MEGA. The Mw and pI of the proteins were predicted using Editseq (DNASTAR), and the gene structures were analyzed utilizing GSDS. The conserved motifs were determined and annotated via MEME and SMART, respectively. The Ka/Ks values were calculated using PAL2NAL (http://www.bork.embl.de/pal2nal) program.

Zhu et al. (2014a) identified the family of protein disulfide isomerase (*PDI*) genes in *B. distachyon* genome utilizing NCBI and Phytozome databases. The protein sequences were analyzed using Pfam, CDD, Interpro (http://www.ebi.ac.uk/interpro), and ExPASy databases. The chromosomal locations were mapped using MapInspect software, and the syntenic relationships were investigated using PGDD. The transcription directions and the structures of the genes were analyzed utilizing Gramene and GSDS, respectively. The protein pI/Mw was calculated via ExPASy, and the signal peptides were predicted using SignalP. The transmembrane regions were predicted via TMHMM and SMART databases, phosphorylation and N-/O-glycosylation sites were determined utilizing NetPhos (http://www.cbs. dtu.dk/services/NetPhos) NetNGlyc (http://www.cbs.dtu.dk/services/NetNGlyc), and NetOGlyc (http://www.cbs.dtu.dk/services/NetOGlyc) servers, respectively. The sequences were aligned using ClustalW, and the phylogenetic tree was generated via MEGA software.

Foxtail Millet

The AP2/ERF transcription factor family in foxtail millet (*Setaria italica*) genome and their expression pattern were investigated by Lata et al. (2014). The HMM profiles were obtained from Pfam and searched against Phytozome for foxtail millet. The conserved domains of the proteins were predicted using HMMER. Chromosomal location, segmental duplication, and gene structure analyses were performed using MapChart (https://www.wageningenur.nl/en/ show/Mapchart.htm), Multiple Collinearity Scan (MCScan; http://chibba.agtec.uga.edu/duplication/mcscan) and Circos (http://circos.ca), and GSDS, respectively. Phylogenetic analysis, GO annotation, *cis*-regulatory element, and miRNA identification were performed using MEGA, Blast2GO (https://www.blast2go.com), PLACE, psRNATarget tools, respectively. For transcriptome analysis, the RNA-seq data were obtained from European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), mapped via CLC Genomics Workbench (http://www.clcbio.com/genomics), and the heat map profile was constructed using TIGR Multiexperiment Viewer (MeV; http://www.tm4.org/mev.html).

Likewise, the superfamily of MYB transcription factors in foxtail millet was investigated by Muthamilarasan et al. (2014) using Pfam and Phytozome. The protein domains were also analyzed using CDD and HMMER. The physical map of chromosomal location was generated using MapChart, and the segmental duplications were identified via MCScanX (http://chibba.pgml.uga.edu/mcscan2). The sequence alignment was performed using BioEdit, and the physicochemical properties of the proteins were determined via ExPASy. The phylogenetic tree was constructed using MEGA, and the GO annotation was performed using Blast2GO. The *cis*-acting elements and the miRNAs were analyzed using PLACE, Foxtail millet Transcription Factor Database (FmTFDb; http://59.163.192.91/FmTFDb), and psRNATarget, respectively. The RNA-seq data were retrieved from ENA and filtered using the NGS QC Toolkit (http://59.163.192.90:8080/ngsqctoolkit). The heat map for expression profile was generated via TIGR MeV. The orthologous relationships of grass MYBs were visualized and analyzed using Circos and PAL2NAL.

Zhu et al. (2014b) analyzed the *NBS-LRR* disease resistance genes in foxtail millet using the BioEdit software (http://www.mbio.ncsu.edu/bioedit/bioedit.html) to construct a local platform based on the genome sequence in ftp://ftp.jgi-psf.org/pub/JGI_data/Setaria_italica. The protein structures were analyzed using Pfam and COILS (http://ch.embnet.org/software/COILS_form.html) software. The sequences were aligned using ClustalX (http://www.clustal.org/clustal2) and the phylogenetic trees were constructed via MEGA.

The PHT1 family of phosphate transporters in foxtail millet was identified by Ceasar et al. (2014) using Phytozome. The phylogenetic analysis was conducted using MEGA, and the *cis*-acting elements were identified via PLACE. The conserved regions in the putative promoter regions were predicted using the Evolutionary Analysis of Regulatory Sequences (EARS; http://wsbc.warwick.ac.uk/wsbcTools-Webpage/user_case_form.php).

Sorghum

The *NBS-LRR* genes in sorghum (*Sorghum bicolor*) genome were identified by Mace et al. (2014) using Phytozome database. The conserved motifs and coiled coils were detected using Pfam and COILS, respectively. The sequences were aligned via ClustalW and MEGA. The phylogenetic trees were constructed and displayed using TreeBest (http://treesoft.sourceforge.net/treebest.shtml) and Dissimilarity Analysis and Representation for Windows (Darwin; http://darwin.cirad.fr) software, respectively. The synteny analysis was performed utilizing Circos software.

Filiz and Tombuloğlu (2015) investigated the distribution of superoxide dismutase (*SOD*) gene family in sorghum genome using Phytozome and NCBI databases. The protein domains were detected via Pfam and the physicochemical characteristics were determined using ProtParam tool. The gene structures were identified using GSDS, and transcript levels of *SbSOD* genes were determined via the NCBI expressed sequence tag (EST) database (http://www.ncbi.nlm.nih.gov/

dbEST). The conserved motifs were analyzed using MEME Suite, the sequences were aligned via ClustalW, and the phylogenetic analysis was conducted using MEGA software. The GO terms were annotated using AmiGO (http://amigo.geneontology.org/amigo), 3D structures of the proteins were predicted utilizing 3DLigandSite server (http://www.sbg.bio.ic.ac.uk/~3dligandsite), and the structural and stereochemical analyses were assessed via RAMPAGE Ramachandran plot analysis (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php).

Panahi et al. (2014) investigated the genes for alternative splicing in sorghum genome using NCBI and Phytozome databases. The transcripts were mapped to the chromosomes using SIM4, and the GTF file was investigated via Alternative Splicing transcriptional landscape visualization tool (Astalavista; http://genome.crg.es/astalavista). For GO analysis, the data were retrieved from Ensemble using BioMart tool (http://www.ensembl.org/info/data/biomart.html) and analyzed via AgriGO.

3.2 Identification of Gene Families in Dicots

3.2.1 Brassicaceae

Arabidopsis thaliana

The first plant genome sequence to be completed belongs to the model organism *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative 2000). Therefore, the genome and transcriptome studies on *A. thaliana* are numerous, and its genome sequence is used for mapping in many plant genome studies. Here are some of the recent studies on investigation of gene families in *A. thaliana* utilizing genome and/or transcriptome data:

Jali et al. (2014) investigated the HUA2-LIKE (*HULK*) gene family in *A. thaliana* genome using TAIR and Phytozome databases. The sequences were aligned using MUSCLE and the conserved blocks were obtained via Gblocks (http://molevol.cmima.csic.es/castresana/Gblocks.html). The phylogenetic tree was generated and visualized using TREE-PUZZLE (http://www.tree-puzzle.de) and Dendroscope (http://ab.inf.uni-tuebingen.de/software/dendroscope) software, respectively. For transcriptome analysis, the data were retrieved from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo). The RNA-seq data were aligned and mapped to *Arabidopsis* genome using TopHat and Bowtie2, and analyzed using Cufflinks.

Ballester et al. (2015) characterized the NGATHA (NGA) clade of transcription factors in *A. thaliana* using genome sequence database. NGA orthologs from different plant were searched via BLAST, and the pairwise alignments were performed using VISTA (http://genome.lbl.gov/vista/index.shtml). The conserved sequences were aligned using ClustalW tool in MacVector software (http://macvector.com).

Siriwardana et al. (2014) analyzed the NF-YA members of NUCLEAR FACTOR-Y (NF-Y) transcription factor families in *A. thaliana* genome, and their role in abscisic acid responses during seed germination. The sequence data were retrieved from TAIR and the phylogenetic analyses were performed using MEGA.

Kenzior and Folk (2015) identified a novel family of plant-specific PWWP/RRM (RNA recognition motif) domain proteins in *A. thaliana* using NCBI and TAIR databases. The sequences were aligned via MultAlin (http://multalin.toulouse.inra.fr/multalin). The Jpred (http://www.compbio.dundee.ac.uk/www-jpred) server was utilized for predicting secondary structures of RRM domain. The phylogenetic analysis was conducted using ClustalW and PHYLogeny Inference Package (PHYLIP; http://evolution.genetics.washington.edu/phylip.html) tools.

Brassica spp.

The MADS-box gene family in *Brassica rapa* (Chinese cabbage) genome was identified by Duan et al. (2014) using the *Brassica* database (BRAD; http://brassicadb.org/brad). The conserved domains were analyzed using Pfam, SMART, and NCBI databases. The MADS proteins from different plants were retrieved and analyzed using TAIR, Pfam, Phytozome, and PGDD databases. The sequences were aligned using ClustalW and the phylogenetic analysis was conducted using MEGA software. The conserved motifs were analyzed via MEME and SMART, and the gene structures were determined using GSDS. The ortholog groups were identified using OrthoMCL, the genes were linked to chromosomes via Circos, and the network relations were built using Cytoscape (http://www.cytoscape.org) software. The gene duplications (Ka/Ks values) were determined using KaKs_calculator (http://evolution.genomics.org.cn/software.htm).

Duan et al. (2015) investigated the ascorbic acid-related genes in *B. rapa* genome using BRAD. The sequences were analyzed using FGENESH and verified via NCBI database. The core eukaryotic genes and random genes were retrieved from CEGMA (http://korflab.ucdavis.edu/Datasets/cegma) and the synteny analysis was performed using MCScanX. The homologs were retrieved from Phytozome and Amborella Genome Database (http://www.amborella.org). The sequences were aligned using ClustalW, the phylogenetic trees were generated using MEGA, and the Ka/Ks values were calculated via KaKs_calculator. The conserved motifs were identified using MEME and the GO terms were annotated via InterProScan.

R2R3-MYB transcription factors in *B. rapa* genome were analyzed by Wang et al. (2015b) using BRAD and PlantTFDB. The conserved protein domains were analyzed using Pfam, SMART, and ExPASy. Mw and pI of the proteins were calculated using the Pepstats (http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats). The conserved motifs were identified via MEME. Multiple alignments of the sequences were performed using ClustalW and WebLogo (http://weblogo.berkeley.edu/logo.cgi), and the phylogenetic analysis was conducted via MEGA. The orthologous and paralogous genes were analyzed via OrthoMCL and plotted using Circos. The gene duplication analysis was performed using MCScanX.

Wang et al. (2014a) identified the *GRF* family of genes in *B. rapa* genome using BRAD. The sequences were aligned using DNAMAN (http://www.lynnon.com) and the phylogenetic analysis was performed utilizing MEGA software. The gene structures were identified via GSDS, the GC content was calculated via DNASTAR, and the physicochemical properties of the proteins were predicted using ProtParam. The Simple Sequence Repeat Identification Tool (SSRIT; http://archive.gramene.org/db/markers/ssrtool) was used to detect the SSR markers. The GO terms were annotated using Gramene, and the conserved motifs were predicted via MEME Suite tool.

The superfamily of WRKY transcription factors in *B. rapa* genome was investigated by Kayum et al. (2015) utilizing BRAD. The conserved domains and the properties of the proteins were analyzed using SMART and ExPASy, respectively. The chromosomal locations were identified via MapChart.

Ma et al. (2014b) analyzed the NAC transcription factor family in *B. rapa* genome using BRAD. The conserved motifs were predicted via MEME, the sequences were aligned utilizing ClustalW, and the phylogenetic tree was generated using MEGA software.

Arya et al. (2014a) investigated the family of heterotrimeric G-protein subunit genes in *B. rapa* genome using Phytozome database. The alignment of sequences and phylogenetic analysis were conducted using ClustalW and MEGA, respectively. The Ka/Ks values were calculated using DnaSP.

The family of *NBS-LRR* resistance genes in *B. oleracea* was analyzed by Kim et al. (2015). The RNA-seq reads were de novo assembled via Velvet (https://www.ebi.ac.uk/~zerbino/velvet) and Oases (https://www.ebi.ac.uk/~zerbino/oases) programs, validated using Phytozome, and mapped to the assembled unigenes using Bowtie. The number of mapped reads was normalized utilizing DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html) software. Functional enrichment and annotation analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov). The RNA-seq data were deposited in the National Agricultural Biotechnology Information Center (NABIC http://nabic.rda.go.kr) database. NBS-encoding genes in the other plant genomes were retrieved from BRAD and PlantTFDB, and the conserved domains were searched in the *Brassica oleracea* Genome Database (Bolbase; http://www.ocri-genomics.org/bolbase). The protein domains were analyzed using SMART, EMBOSS, and myHits (http://myhits.isb-sib.ch/cgi-bin/motif_scan). The conserved motifs were predicted via MEME, and the genes were linked to chromosomes using MapChart.

Yao et al. (2015) identified the genome-wide distribution of *WRKY* gene family in *B. oleracea* var. *capitata* using Bolbase database. The HMM profiles were analyzed via Pfam and HMMER, the sequences were aligned using MUSCLE, and the protein domains were identified via SMART. For RNA-seq analyses, the data were retrieved from SRA, the sequence reads were mapped using Bowtie and TopHat, and the transcript reconstruction was performed using Cufflinks. The conserved motifs were predicted via MEME, the sequences were aligned using MUSCLE and BioEdit, and the phylogenetic tree was generated using MEGA software.

3.2.2 Fabaceae (Leguminosae)

Chickpea

Distribution of Ethylene Responsive Factor (*ERF*) gene family in chickpea (*Cicer arietinum*) genome was identified by Deokar et al. (2015) utilizing PlnTFDB, PlantTFDB, and International Chickpea Genetics and Genomics Consortium (ICGGC; http://www.icrisat.org/gt-bt/ICGGC/GenomeManuscript.htm) databases. The protein domains were analyzed using CDD, SMART, and Pfam. The phylogenetic tree was constructed using MEGA, and the gene duplication was predicted via MCScanX. The sequence homology was analyzed using BLAST, and the protein sequence analyses were performed via ExPASy and Membrane protein IdeNtificatioN withOUt explicit use of hydropathy profiles and alignments (MINNOU; http://minnou.cchmc.org) tools. Secondary structures were predicted using YASPIN (http://www.ibi.vu.nl/programs/yaspinwww) and Advanced Protein Secondary Structure Prediction Server (APSSP; http://imtech.res.in/raghava/apssp). Subcellular localization of the proteins was analyzed using PredictProtein (https://www.predictprotein.org) and WoLF PSORT (http://wolfpsort.org) servers. The *cis*-acting elements were predicted via PLACE and PlantCARE. The DNA and amino acid analysis were performed using BioEdit and DNASTAR.

Sharma et al. (2014b) investigated the uridine diphosphate glycosyltransferase (*UGT*) family of genes in chickpea genome utilizing chickpea browser of Legume Information System (LIS; http://cicar.comparative-legumes.org). The family of UGTs and the conserved motifs were identified using SUPERFAMILY and MEME tools, respectively. The HMM profiles were analyzed using Pfam and HMMER. The dendrogram for evolutionary analysis was drawn via PhyML (http://atgc.lirmm.fr/phyml). To determine the functional specificity of UGTs, characterized proteins with known substrate specificity from different plants were retrieved from Swiss-Prot (http://www.uniprot.org) database. The molecular modeling of the proteins was performed using the Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do). The stereochemical properties of the 3D models were analyzed via PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK), Verify3D (http://services.mbi.ucla.edu/Verify_3D), ProSA (https://prosa.services.came.sbg.ac.at/prosa.php), and ERRAT (http://services.mbi.ucla.edu/ERRAT) tools. The UGT orthologs were predicted using Blast2GO, and the gene structures were identified via GSDS.

Sharma and Suresh (2015) analyzed the proteases and protease inhibitors in chickpea genome using LIS database. The HMM profiles were identified using Pfam and HMMER. The sequences were aligned, bootstrapped, and analyzed via ClustalX, PHYLIP, and ProtDist (http://evolution.genetics.washington.edu/phylip/doc/protdist.html), respectively, and the dendrograms were obtained utilizing the FigTree (http://tree.bio.ed.ac.uk/software/figtree). The conserved domains were analyzed using SMART and DomainGraph (http://domaingraph.bioinf.mpi-inf.mpg.de), and the signal peptides were predicted via SignalP. The orthologs were identified using Blast2GO, and the codon composition was analyzed via MEGA.

The gene structure was identified using GSDS. For gene expression analyses, the RNA-seq data were retrieved from SRA, mapped to the chickpea genome using TopHat, and the abundance of reads were estimated via Cufflinks.

Ha et al. (2014) investigated the NAC family of transcription factors in chickpea genome utilizing the Chickpea Transcriptome Database (CTDB; http://www.nipgr.res.in/ctdb.html) PlantTFDB and iTAK (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi) program. Additionally, the genome sequence of chickpea cultivar "desi" (CGAP; http://nipgr.res.in/CGAP/home.php) was used, and the HMM profiles were identified via TMHHM. The phylogenetic analyses were conducted using MEGA, and the sequence alignment was visualized via GeneDoc (http://www.nrbsc.org/gfx/genedoc).

Jain et al. (2014) performed the genome-wide identification of the miRNAs in chickpea. The RNA-seq data were pre-processed using miRTools (http://centre.bioinformatics.zj.cn/mirtools) software, and annotated for small nucleolar RNAs using Plant SnoRNAbase (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home), for tRNAs using Genomic tRNA Database (http://gtrnadb.ucsc.edu/download.html), and for rRNAs using RFAM (http://rfam.xfam.org) database. The rest of the sequence was screened via RepBase (http://www.girinst.org/server/RepBase). The chickpea miRNAs were identified in miRBase and the filtered reads were mapped to these miRNAs using Bowtie. The secondary structures of the genomic sequences were determined using RNAfold (http://www.tbi.univie.ac.at/RNA/RNAfold.html) software, and processed utilizing miRDeep-P core algorithm (http://faculty.virginia.edu/lilab/miRDP). The mature miRNA candidates were clustered into families via CD-HIT server, and the putative targets of the miRNAs were predicted using psRNATarget server. The conserved domains of the targets were identified via Pfam and HMMER program. The GO terms were annotated using BİNGO (http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html) software, and the eukaryotic orthologous groups (KOGs) were identified utilizing NCBI KOG server (ftp://ftp.ncbi.nih.gov/pub/COG/KOG).

Soybean

The cupin gene family in soybean (*Glycine max*) genome was investigated by Wang et al. (2014f) utilizing Phytozome. The conserved domains were identified via InterProScan. Sequence alignment and phylogenetic analysis were conducted using ClustalX and MEGA, respectively. The logos for amino acid residues in conserved domains were generated using WebLogo. The conserved motifs were analyzed via MEME, and annotated using SMART, Pfam, and NCBI database. The exon/intron organizations were identified via GSDS, and the chromosomal locations were mapped using Chromosome Visualization Tool (CViT) at the LIS database (http://cvit.comparative-legumes.org). The Ka/Ks values were calculated using DnaSp. For gene expression analysis, the data were obtained from the SoyBase database (http://soybase.org), analyzed using Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm), and the heat map was visualized via Java Treeview

(http://jtreeview.sourceforge.net). For evolutionary analysis, the SNP data were retrieved from the Soybean Knowledge Base (SoyKB; http://soykb.org).

Mainali et al (2014) identified the cyclophilin (*CYP*) gene family in soybean genome utilizing Phytozome database. The Mw and subcellular localization of the proteins were predicted using ProtParam program, and TargetP (http://www.cbs.dtu.dk/services/TargetP) and WoLF PSORT servers, respectively. The transcribed *CYP*s were analyzed using soybean gene index (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Glycine_max). The sequences were aligned and analyzed using ClustalX, MEGA, and iTOL. For transcriptome analysis, the data were retrieved from GEO, and the heat map was generated using the gplots CRAN library (http://cran.r-project.org/web/packages/gplots/index.html).

Li et al. (2014a) investigated the family of heat shock transcription factors (Hsfs) in soybean genome utilizing Phytozome and SoyBase. The conserved domains were analyzed using SMART, Pfam, Predict Nuclear Localization Signals (PredictNLS; https://rostlab.org/owiki/index.php/PredictNLS), and Nuclear Export Signals (NetNES; http://www.cbs.dtu.dk/services/NetNES) servers, and the genes were mapped to the chromosomes via MapDraw. The exon-intron substructures and the *cis*-acting elements were analyzed using GSDS and PLACE tools, respectively. The phylogenetic tree was constructed using MEGA software.

The family of R2R3-MYB transcription factors in soybean was identified by Aoyagi et al. (2014) utilizing PlantTFDB, SoyDB, and Phytozome. The conserved domains and motifs were predicted using Pfam and SMART, and MEME, respectively. For transcriptome analysis, RNA-seq data were retrieved from LGE Soybean Genome Project (http://bioinfo03.ibi.unicamp.br/soja) and Genevestigator databases. The C-terminal amino acid sequences of the proteins were analyzed via MEME, and the phylogenetic tree was generated using MEGA.

Belamkar et al. (2014) analyzed the distribution of homeodomain leucine zipper (HD-Zip) transcription factor family in soybean using Phytozome database. Following BLAST search, the sequences were aligned using MUSCLE and processed via SeaView (http://doua.prabi.fr/software/seaview). The phylogenetic trees were generated using CLUSTAL, PhyML, and FigTree. The HMM profiles were analyzed using HMMER. For RNA-seq analysis, the data were obtained from SoyBase and SoyKB. The GO terms were annotated using Blast2GO and SoyDB (http://casp.rnet.missouri.edu/soydb), and the heat map was obtained via the gplots CRAN library.

Bencke-Malato et al. (2014) investigated the family of WRKY transcription factors in soybean genome utilizing Phytozome, PlantTFDB, PLAZA (http://bioinformatics.psb.ugent.be/plaza), and SoyBase. The protein domains were identified using SMART. The coding sequences were analyzed via GENSCAN and FGENESH, and the conserved motifs were predicted using MEME. The functional analysis of the proteins was performed via FancyGene. The sequences were aligned using MUSCLE option in MEGA, and the phylogenetic analyses were conducted via the Bayesian Evolutionary Analysis Sampling Trees (BEAST; http://beast.bio.ed.ac.uk) software. The best-fit model of amino acid replacement was analyzed using ProtTest (https://code.google.com/p/prottest3), and the phylogenetic trees were visualized via FigTree.

Feng et al. (2014a) investigated the family of *IQD* genes in soybean using Phytozome. The HMM profiles were determined utilizing Pfam and SMART. The Mw and pI of the proteins were predicted using ExPASy, and the subcellular localizations were determined via TargetP and WoLF PSORT. The sequence alignment and phylogenetic tree construction were performed using ClustalX and MEGA, respectively. The gene structures were identified using GSDS, and the conserved motifs were determined via MEME. The putative calmodulin-binding sites were predicted using the Calmodulin Target Database (http://calcium.uhnres.utoronto.ca/ctdb/ctdb/home.html). The chromosomal locations of the genes were mapped via MapInspect, and the segmental duplications were analyzed using SoyBase. The Smith-Waterman algorithm (http://www.ebi.ac.uk/Tools/psa) was utilized to calculate the local alignment of two protein sequences. The amino acid sequences were aligned via ClustalX, the codon alignments were generated using PAL2NAL, and Ka/Ks values were calculated using CODEML program of Phylogenetic Analysis by Maximum Likelihood (PAML; http://abacus.gene.ucl.ac.uk/software/paml.html) package. The microsynteny analysis was conducted using PGDD. For transcriptome analysis, the RNA-seq data were retrieved from SoyBase.

The soybean glutamate decarboxylase (*GAD*) gene family was analyzed by Hyun et al. (2014) utilizing the Phylogeny.fr server (http://phylogeny.lirmm.fr/phylo_cgi/index.cgi) and PGDD. The sequence alignment was performed via PAL2NAL, and the Ka/Ks calculation was conducted using PAML. The start codon for *GAD* genes was obtained from Phytozome, and the *cis*-acting elements were analyzed using PLACE.

Medicago truncatula

The family of auxin/indoleacetic acid (*Aux/IAA*) genes in *Medicago truncatula* genome was investigated by Shen et al. (2014) using Phytozome. The conserved domains and the synteny blocks were analyzed via InterProScan and SyMAP, respectively. Multiple sequence alignment was performed via ClustalW, and the phylogenetic tree was constructed using MEGA and visualized via TreeView. The chromosomal locations of the genes were mapped using Circos, and the motif analysis was conducted via MEME.

The distribution of Gretchen Hagen 3 (*GH3*) gene family in *M. truncatula* genome was identified by Yang et al. (2015) using *M. truncatula* Genome Database (MtGDB; http://www.plantgdb.org/MtGDB). The conserved domains were analyzed via InterProScan. The sequence alignment was performed via ClustalW and visualized using GeneDoc. The phylogenetic tree was generated using MEGA software. The synteny blocks were analyzed using SyMAP, and the conserved motifs were predicted via MEME. The *cis*-regulating elements were determined using PLACE.

The LEED..PEED (*LP*) gene family, unique to the *Medicago* lineage, was identified by Trujillo et al. (2014) in *M. truncatula* genome using NCBI, Phytozome, LIS, J. Craig Venter Institute (http://www.jcvi.org/medicago), Kazusa DNA Research

Institute (http://www.kazusa.or.jp/lotus), and Dana Farber Cancer Institute—Gene Indices (http://compbio.dfci.harvard.edu/tgi) databases. The small peptides were analyzed using SPADA program (Zhou et al. 2013), and the conserved domains were identified via InterProScan. The synteny analysis was conducted using MUMmer (http://mummer.sourceforge.net). The gene homology patterns were analyzed via Genome Evolution Analysis (GEvo; https://genomevolution.org/CoGe/GEvo.pl) and visualized using Multi-Genome Synteny Viewer (mGSV; http://cas-bioinfo.cas.unt.edu/mgsv) tools. The sequences were aligned using ClustalW, the phylogenetic analysis was conducted using MEGA and MrBayes, and the tree was visualized via FigTree. Gene duplications were analyzed and displayed using DILTAG (http://www-lbit.iro.umontreal.ca/DILTAG) program.

3.2.3 Solanaceae

Tomato

The family of CLAVATA3/EMBRYO-SURROUNDING REGION-RELATED (*CLV3/ESR*, *CLE*) genes in tomato (*Solanum lycopersicum*) genome was identified by Zhang et al. (2014c) using Phytozome and the tomato resource in SOL Genomics Network (SGN; http://solgenomics.net/organism/Solanum_lycopersicum/genome). The sequence alignment and phylogenetic analysis were conducted using ClustalX and MEGA, respectively. The conserved motifs and *cis*-acting elements were analyzed via MEME and PLACE, respectively.

Zhang et al. (2014d) identified the family of HD-Zip transcription factors in tomato using SGN. The protein domains were analyzed via Pfam and SMART. Mw and pI of the proteins were determined using ProtParam, and the subcellular localizations were predicted via CELLO (http://cello.life.nctu.edu.tw). The multiple alignment of the sequences was performed using ClustalX, and the phylogenetic tree was constructed via MEGA. The gene structures were identified using GSDS. The conserved motifs were analyzed via MEME. The *cis*-acting elements were predicted using PLACE.

Cao and Li (2014) analyzed the family of late embryogenesis abundant (*LEA*) genes in tomato using Phytozome. The physicochemical properties of the proteins were investigated via ProtParam, and intrinsically disordered proteins were analyzed using IUPred (http://iupred.enzim.hu) server. The subcellular localizations were predicted via CELLO server and PSORT. Multiple sequence alignment and phylogenetic analysis were performed via MUSCLE and MEGA, respectively. The gene duplication/lost analysis was conducted using NOTUNG (http://www.cs.cmu.edu/~durand/Notung) software. The K-Estimator program (https://bioweb.biology.uiowa.edu/labs/comeron/software) was used for Ka/Ks calculation. The *cis*-elements were analyzed via PLACE, and the recombination events were predicted using the Recombination Detection Program (RDP; http://web.cbio.uct.ac.za/~darren/rdp.html). Additionally, the site-specific positive selection and purifying selection was analyzed using the Selecton Server (http://selecton.tau.ac.il).

For transcriptome-wide expression analysis, the data were retrieved from GEO and processed via Genesis program (http://genome.tugraz.at).

Wu et al. (2014) investigated the mitogen-activated protein kinase (MAPK) kinase (MAPKK) and MAPKKK family in tomato genome using SGN and the Kazusa Full-length Tomato cDNA Database (KafTom; http://www.pgb.kazusa.or.jp/kaftom/blast.html). The protein domains were analyzed using Pfam and SMART. The Mw and pI of the proteins were determined using ExPASy, and the subcellular localizations were predicted via CELLO. The sequences were aligned via ClustalX, and the phylogenetic analysis was performed using MEGA. The Plant Phosphorylation (PlantsP; http://plantsp.genomics.purdue.edu/index.html) was used for motif and domain analysis, and the *cis*-elements were analyzed using PLACE. The chromosomal locations were identified via SGN, and the synteny analysis was conducted using PGDD.

Chen et al. (2014) identified the mildew resistance locus o (*MLO*) gene family in tomato genome utilizing SGN and the Plant Genome and Systems Biology (PGSB; http://pgsb.helmholtz-muenchen.de/plant/tomato/searchjsp/index.jsp) database. The conserved domains were identified via Pfam, the sequences were aligned using ClustalX, and the phylogenetic tree was generated using MEGA. The gene structure was determined via GSDS, and the genes were mapped to the chromosomes using MapDraw. The MEME tool was utilized for prediction of conserved motifs.

The basic helix-loop-helix (bHLH) family of transcription factors in tomato genome was investigated by Sun et al. (2015) via SGN. The conserved domains were analyzed using Pfam, SMART, and HMMER. The motif analysis was performed using MEME. The sequences were aligned using MultAlin and Clustal Omega (http://www.clustal.org/omega) and visualized via WebLogo. The phylogenetic analysis was performed using MEGA and FigTree. The gene duplications were analyzed via MUMmer and mapped to the chromosomes using MapChart. The *cis*-acting elements were identified using PLACE.

Potato

The family of ERF transcription factors in potato (*Solanum tuberosum*) genome was investigated by Charfeddine et al. (2014) using Phytozome. The conserved domains were searched using Pfam, and the gene structures were corrected using FGENESH. The sequence alignment was conducted via ClustalW and the phylogenetic analysis was performed using MEGA. The genes were mapped to the chromosomes using MapChart, and the Pamilo–Bianchi–Li substitution model in MEGA was utilized for a codon-based Z-test for each block. The pI of the proteins was calculated using ProtParam, and the FoldIndex (http://bioportal.weizmann.ac.il/fld-bin/findex) program was used for prediction of protein folding. The subcellular localization of the protein was analyzed via TargetP. For RNA-seq analysis, the data were retrieved from SRA and clustered using MeV. The conserved motifs were predicted via MEME, and the gene structures were analyzed using GSDS.

Charfeddine et al. (2015) investigated the *LEA* gene family in potato genome using Phytozome. The conserved domains were analyzed via Pfam and FGENESH. The sequences were aligned using ClustalW and the phylogenetic tree was generated via MEGA. The signal peptides, transmembrane regions, and subcellular localization of the proteins were predicted using SignalP, TMHHM, and TargetP, respectively. The gene structures were predicted via GSDS, and the conserved motifs were analyzed using MEME. The chromosomal locations of the genes were mapped using MapChart, and a codon-based Z-test was applied using MEGA. The physicochemical and folding properties of the proteins were predicted via ProtParam and FoldIndex, respectively.

Yang et al. (2014) identified the miR159 family and MYB transcription factors as their targets in potato using mirBase, NCBI, PlantTFDB, the Potato Genome Sequencing Consortium (PGSC; http://potatogenome.net/index.php/Main_Page) database, and the Unified Nucleic Acid Folding and hybridization package (UNAFold; http://www.bioinfo.rpi.edu/applications/mfold). The miRNA targets were predicted via psRNATarget, the sequences were aligned using ClustalX, and the phylogenetic tree was generated using MEGA. The conserved domains of the MYBs were analyzed using Pfam, the physicochemical properties were determined via ExPASy, and the gene structures were identified using Splign (http://www.ncbi.nlm.nih.gov/sutils/splign).

Sharma et al. (2014a) investigated the BEL1-like (BELL) family of transcription factors in potato using PGSC server. The sequences were aligned via ClustalW algorithm in BioEdit, and the phylogenetic tree was generated using MEGA. The ORFs were predicted using FGENESH.

3.2.4 Trees

Apple

The *NBS-LRR* gene family in apple (*Malus x domestica*) genome was investigated by Arya et al. (2014b) using Phytozome. The conserved domains were identified via Pfam and HMMER. The sequence alignment was performed via ClustalW, and the coil-coiled motif in proteins was identified using COILS program. The conserved NBS–leucine-rich repeat (LRR) motifs were analyzed via MEME. The genes were mapped to the chromosomes using MapInspect, and the duplication events were analyzed using MCScanX. The phylogenetic analysis was conducted via the Randomized Axelerated Maximum Likelihood tool (RAxML; http://sco.h-its.org/exelixis/web/software/raxml/index.html).

The cystatin gene family in apple genome was identified by Tan et al. (2014) using NCBI and *M. x domestica* genome data in Genome Database for Rosaceae (GDR; http://www.rosaceae.org/species/malus/malus_x_domestica). The protein domains were analyzed using Pfam and SMART. The MapDraw was utilized to map the genes to the chromosomes. The signal peptides were predicted via SignalP, and the Mw and pI of the proteins were calculated using ExPASY. The multiple

alignment of the sequences was performed using CLC Combined Workbench, and the phylogenetic tree was generated using MEGA software. The gene structures were predicted via GSDS, and the cis-acting elements were analyzed utilizing PLACE and PlantCARE databases.

Wei et al. (2014b) investigated the sugar transporter (*SUT*) gene family in apple genome using GDR and the apple genome database in the Istituto Agrario San Michele all'Adige (IASMA; http://genomics.research.iasma.it). The sequences were aligned via DNAMAN and MUSCLE, and the phylogenetic tree was obtained using MEGA. The subcellular localizations were predicted via TargetP and WoLF PSORT. The chromosomal locations of the genes were mapped using MapDraw.

The distribution of teosinte branched1/cycloidea/proliferating cell factor1 (TCP) family of transcription factors in apple genome was identified by Xu et al. (2014) using the Apple Gene Function and Gene Family DataBase (AppleGFDB; http://www.applegene.org), NCBI, and GDR databases. The conserved domains were analyzed using Pfam and SMART. The Mw and pI of the proteins were calculated via ExPASy. The sequence alignment was conducted using ClustalX and MUSCLE, and the phylogenetic tree was generated via MEGA. The genes were mapped to the chromosomes using MapDraw, and the gene structures were analyzed using GSDS.

Tian et al. (2015) analyzed the MADS-box gene family in apple genome using the IASMA database. The conserved domains were determined using Pfam and CDD. The sequence logo was generated via WebLogo. The protein structure homology models were predicted using SWISS-MODEL (http://swissmodel.expasy.org). The 3D structure models were presented using RasTop (http://www.geneinfinity.org/rastop). The sequence alignment was conducted using CLC Combined Workbench, and the phylogenetic tree was generated via MEGA. The conserved motifs were analyzed using MEME. Chromosomal location of the genes was mapped via MapInspect. The intron–exon structures were identified and visualized via PLAZA and SigmaPlot (http://www.sigmaplot.com), respectively.

Shao et al. (2014) investigated the sucrose non-fermenting-1-related protein kinase 2 (*SnRK2*) gene family in *M. prunifolia* (Chinese apple) genome utilizing GDR and IASMA databases. The conserved domains were analyzed using CDD, Pfam, SMART, and PROSITE. The sequences were aligned via ClustalW, MUSCLE, and DNAMAN, and the synteny analysis was performed using PGDD. The gene structure and conserved motif analyses were conducted via GSDS and MEME, respectively. The phylogenetic tree was constructed using MEGA software.

Pessina et al. (2014) identified *MLO* gene family in apple using GDR. The conserved protein motifs were analyzed using HMMER. The membrane spanning helices were identified using InterPro. The sequence alignment was conducted via CLC Sequence Viewer. The orthology and synteny analyses were performed using GBrowse-Syn tool at GDR (http://www.rosaceae.org/gb/gbrowse_syn/peach_apple_strawberry) and Mercator (https://www.biostat.wisc.edu/~cdewey/mercator), respectively.

Citrus spp.

The MYB transcription factor family in sweet orange (*Citrus sinensis*) genome was identified by Hou et al. (2014) using the Orange Genome Annotation Project (OGAP; http://citrus.hzau.edu.cn/orange/index.php) database. The conserved domains were analyzed using Pfam, and the sequences were aligned via ClustalW. The gene structures were analyzed using GSDS. The physicochemical properties and the subcellular localizations of the proteins were analyzed using ProtParam and the Protein Localization Server (PLOC; http://www.genome.jp/SIT/plocdir). The conserved motifs were predicted using MEME, and the phylogenetic tree was constructed via MEGA. The GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations were conducted using Blast2GO. The chromosomal locations of the genes were mapped via MapChart, and the Ka/Ks values were determined using CODEML module of the PAML. The repetitive elements were analyzed using Tandem Repeats Finder (TRF; http://tandem.bu.edu/trf/trf.html) and Inverted Repeats Finder (IRF; http://tandem.bu.edu/irf/irf.download.html). The low copy repeats (LCRs) and transposable elements (TEs) were identified using RepeatMasker (http://www.repeatmasker.org) and the simple sequence repeats (SSRs) were determined via the Simple Sequence Repeat Identification Tool (SSRIT; http://archive.gramene.org/db/markers/ssrtool).

Ito et al. (2014) identified the AP2/ERF superfamily of transcription factors in sweet orange genome using the Citrus Genome Database (http://www.citrus-genomedb.org/species/sinensis). The ORFs were detected using ORF finder, the sequences were aligned, and the phylogenetic tree was generated using ClustalW and MEGA, respectively. The conserved motifs of the proteins were predicted via MEME, and the gene structures were identified using GSDS tool.

Xie et al. (2014) investigated the *R2R3-MYB* gene family in the genomes of sweet orange and clementine (*C. clementina*) utilizing Phytozome and PlantTFDB. The conserved domains of the proteins were predicted via PROSITE and SMART. Multiple sequence alignment was conducted using ClustalX and adjusted via BioEdit. The gene structures were analyzed using GSDS, and the phylogenetic analysis was conducted via MEGA.

Lin et al. (2015) analyzed the heat shock transcription factors in Ponkan (*C. reticulata* Blanco cv. Ponkan) using the Citrus Genome Database. The CAP3 sequence assembly program was used to eliminate redundant sequences and alignment was performed via ClustalX. The phylogenetic tree was generated using TreeView. The conserved motifs were identified using MEME Suite tool.

Poplar

The WRKY transcription factor family in poplar (*Populus trichocarpa*) genome was investigated by Jiang et al. (2014) using Phytozome and PlantTFDB. The sequence alignment was performed using ClustalX, and the phylogenetic tree was constructed via MEGA. The *cis*-regulatory elements were analyzed using

PlantCARE. For gene expression analysis, the data were retrieved from Phytozome and the sequence reads were mapped to the poplar genome using Short Oligonucleotide Analysis Package (SOAP; http://soap.genomics.org.cn).

Ma et al. (2014a) identified the *IQD* gene family in poplar genome using Phytozome. The conserved domains were analyzed using Pfam and SMART. The physicochemical properties of the proteins were determined via ExPASy, and the subcellular localization of the proteins was identified using WoLF PSORT. The calmodulin-binding sites were predicted via the Calmodulin Target Database. The gene structures and the conserved motifs were identified using GSDS and MEME, respectively. The genes were mapped to the chromosomes using MapInspect, and the synteny analysis was performed using the Vista Synteny browser (http://pipeline.lbl.gov/cgi-bin/gateway2). The sequences were aligned using ClustalX, and the phylogenetic analysis was conducted via MEGA. The Ka/Ks values were calculated using CODEML program in PAML after multiple alignment via PAL2NAL. For gene expression analysis, the data were obtained from Gene Indices (http://compbio.dfci.harvard.edu/tgi) and GEO. The heat map was visualized via Heatmapper Plus (http://bar.utoronto.ca/ntools/cgi-bin/ntools heatmapper plus.cgi) and Cluster tools.

Li and Lu (2014) analyzed the SQUAMOSA PROMOTER BINDING PROTEIN LIKE (*SPL*) gene family in poplar using Phytozome. The conserved domains and motifs were identified using Pfam and CDD, and MEME, respectively. Sequence logos were generated via WebLogo. The paralogs were identified via PGDD, and the Ka/Ks calculation was performed using DnaSP. The Mw and pI of the proteins were predicted using ExPASy, and the gene structure was analyzed via GSDS. The sequence alignment was constructed using ClustalW, and the phylogenetic tree was generated via MEGA software.

Chai et al. (2014) identified the *R2R3-MYB* gene family in poplar utilizing Phytozome. The conserved domains were analyzed using Pfam, the sequences were aligned via ClustalX, and phylogenetic analysis was conducted using MEGA. The gene structures and the conserved motifs were identified using GSDS and MEME, respectively. Detected motifs were searched in databases using MAST. The gene expression was analyzed using the data retrieved from GEO, and normalization was conducted via Genesis.

Eucalyptus grandis

The R2R3-MYB transcription factor family in *Eucalyptus grandis* was identified by Soler et al. (2014) using Phytozome. The protein domains were analyzed via InterProScan, the sequences were aligned using MAFFT, and the phylogenetic tree was generated using MEGA. The gene structures were retrieved from Phytozome and represented via FancyGene. The physical positions of the genes on the corresponding chromosomes were mapped using MapChart. The gene duplication was analyzed using a Z-test in MEGA. The conserved motifs were predicted using MEME. For transcriptome analysis, the RNA-seq data were obtained from

EucGenIE (http://www.eucgenie.org) and normalized via the EXpression Analyzer and DisplayER (EXPANDER; http://acgt.cs.tau.ac.il/expander).

The *Aux/IAA* gene family in *E. grandis* genome was investigated by Yu et al. (2015) utilizing Phytozome. The conserved domains were identified via Pfam and CDD. The gene models were processed using FGENESH and mapped to the chromosomes via MapChart. The physicochemical features of the proteins were predicted using ProtParam, and the conserved motifs were analyzed using MEME. The exon–intron structures were retrieved from Phytozome and visualized using FancyGene. The sequences were aligned via ClustalX and the phylogenetic tree was constructed using MEGA.

Yu et al. (2014) analyzed the family of AUXIN RESPONSE FACTOR (*ARF*) genes in *E. grandis* genome using Phytozome. The protein domains were analyzed using CDD and Pfam, and the gene models were processed via FGENESH and mapped to the related chromosomes using MapChart. The sequences were aligned via ClustalX, and the phylogenetic analysis was conducted using MEGA. The gene structures were represented using FancyGene. The small RNA target sites were predicted via psRNATarget. The stem-loop structures of the RNAs were analyzed and visualized via RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) and RNAstructure (http://rna.urmc.rochester.edu/RNAstructure.html) servers, respectively.

Hussey et al. (2014) identified the NAC family of transcription factors in *E. grandis* genome using Phytozome, PlantTFDB, Eucspresso (http://eucspresso.bi.up.ac.za), and EucGenIE. The protein domains were analyzed using Pfam and HMMER, and the transmembrane helix structures were predicted via TMHMM. The gene models were processed using FGENESH, and the gene structures were predicted via GSDS. The multiple sequence alignment was generated via MUSCLE and trimmed using Gblocks. The phylogenetic tree was constructed and visualized using PhyML and MEGA, respectively. The conserved motifs were predicted via MEME; overrepresented motifs were annotated using Pfam-A and Pfam-B, and schematically represented via DomainDraw (http://domaindraw.imb.uq.edu.au). The genes were mapped to the chromosomes using MapChart. The sequences were aligned via MUSCLE, and the phylogenetic analysis was conducted using MEGA. For transcriptome analysis, the data were retrieved from EucGenIE and analyzed using TopHat and Cufflinks. The expression values were clustered using the QT clustering tool in the MeV.

3.2.5 Other Plants

Cotton

The family of WRKY transcription factors in the genomes of *Gossypium raimondii* and *G. arboreum* was investigated by Ding et al. (2015) using Phytozome and Cotton Genome Project database (CGP; http://cgp.genomics.org.cn/page/species/index.jsp), respectively. The FGENESH was utilized for gene and protein prediction. The conserved domains were analyzed using HMMER, Pfam, and SMART,

and revised using PlantTFDB. Tandem duplications were detected using MCScanX. Ka/Ks calculation was performed via Ka_Ks Calculator. The gene conversion events were identified via GENECONV (http://www.math.wustl.edu/~sawyer/geneconv). The sequences were aligned and the phylogenetic tree was generated using MUSCLE and MEGA, respectively. The gene structures were evaluated via the Plant Intron Exon Comparison and Evolution (PIECE; http://wheat.pw.usda.gov/piece) database. The functional divergence of the subgroups was identified using DIVERGE (http://xungulab.com/software/diverge2/diverge2.html). The selective pressures on codons were analyzed using CODEML package of PAML.

The *SPL* gene family in *G. hirsutum* genome was investigated by Zhang et al. (2015b) using the CGP and CottonGen (http://www.cottongen.org). The genes were mapped to the corresponding chromosomes using MapChart. The phylogenetic analysis was conducted via MEGA. The conserved motifs were predicted using MEME. The miRNAs targeting the *GhSPL*s were identified using miRBase and psRNATarget.

Zhang et al. (2014b) investigated the MAPK family in *G. raimondii* genome using Phytozome and NCBI EST database. The conserved domains were analyzed via HMMER, Pfam, InterProScan, SMART, PlantsP, and MOTIF (http://www.genome.jp/tools/motif) tools. The chromosomal locations of the genes were mapped using MapInspect. The subcellular localization of the proteins was detected using CELLO. The sequences were aligned via ClustalX and the phylogenetic tree was obtained using MEGA.

Yurchenko et al. (2014) identified the omega-3 fatty acid desaturase (*FAD*) gene family in *G. hirsutum* genome using the databases NCBI, Phytozome, and Cotton Genome Database (CottonDB; http://www.cottondb.org/wwwroot/cdbhome.php). The intron/exon structures were determined using Softberry package (http://www.softberry.com). The sequences were aligned using T-Coffee (http://tcoffee.crg.cat) and cleaned via Gblocks. The phylogenetic tree was generated using PhyML and visualized via FigTree.

Wang et al. (2014c) analyzed the family of heat shock transcription factors in *G. hirsutum* genome using the NCBI EST database. The conserved domains and motifs were analyzed via SMART and MEME, and visualized using ProSite. The sequences were aligned using DNAMAN and ClustalX. The phylogenetic analyses including gene duplication were performed using MEGA. Physicochemical features of the proteins were analyzed via ExPASy.

The aldehyde dehydrogenase (*ALDH*) gene superfamily in *G. raimondii* genome was investigated by He et al. (2014) utilizing NCBI and Phytozome. The protein domains were determined using Pfam. Multiple sequences were aligned and edited using ClustalW and BioEdit, respectively. The phylogenetic analysis was performed via MEGA. The intron–exon structures were identified using FancyGene, and the sequence repeats were determined using RepeatMasker. The synteny analysis was conducted using MCScanX. For gene expression analysis, the microarray and RNA-seq data were retrieved from the GEO and Plant Expression Database (PLEXdb; http://www.plexdb.org), and SRA, respectively. RNA-seq reads were mapped to the gene models via TopHat, and differentially regulated genes at the transcriptional or

post-transcriptional level were estimated using Cuffdiff (http://cole-trapnell-lab.github.io/cufflinks/cuffdiff).

Grapevine

The MADS-box transcription factors in grapevine (*Vitis vinifera*) were identified by Wang et al. (2014d) using the GENOSCOPE database (http://www.genoscope.cns.fr/spip/Vitis-vinifera.html). The conserved domains were identified via Pfam, SMART, and Domain Graph software (DOG; http://dog.biocuckoo.org). Gene duplication events were analyzed using PGDD. Multiple sequence alignment was conducted using ClustalW, and the phylogenetic analysis was performed via MEGA. The gene structures were identified using GSDS.

The MAPKKK family in grapevine was analyzed by Wang et al. (2014b) using Grape Genome Database (http://genomes.cribi.unipd.it/grape), Vitis-URGI (http://urgi.versailles.inra.fr/Species/Vitis), and NCBI. The conserved domains were identified via Pfam, HMMER, and SMART. The sequence alignment was performed using ClustalX and GeneDoc, and the phylogenetic tree was generated using Phylogeny.fr. The gene structures were analyzed using GSDS.

The subtilase gene family in grapevine was investigated by Cao et al. (2014). The protease-associated subtilisin-like domain (PA_subtilisin_like domain) was searched using CDD. The primary structural analyses were performed using TargetP and PredoTar (https://urgi.versailles.inra.fr/Tools/Predotar). The sequences were aligned via MUSCLE and ClustalW, and the phylogenetic analyses were performed using PhyML and PHYLIP. The protein substitution model and rate heterogeneity were evaluated using ModelGenerator (http://bioinf.nuim.ie/modelgenerator). The chromosomal locations were identified using GENOSCOPE, and the gene structure information was gathered from NCBI and Phytozome. Ka/Ks analysis was performed using K-Estimator, and the conserved motifs were identified via MEME. The functional divergence was analyzed using DIVERGE. Site-specific selection analyses were conducted via SLAC, REL, and FEL methods in Datamonkey web interface (http://www.datamonkey.org/dataupload.php).

Matus et al. (2014) identified the BURP superfamily in grapevine using Grape Genome Database and GENOSCOPE database. The sequences were aligned via MUSCLE and the phylogenetic trees were obtained using MEGA and FigTree. The conserved motifs were identified using MEME.

4 Conclusion and Future Perspective

Identification of gene families in the genome/transcriptome of a plant gives important clues about the organism's phylogenetic position, genome mobility, stress tolerance, gene expression profiles, and so on. Hence, the number of articles on this subject is getting higher day by day as more plant genome sequences are released.

Currently, various bioinformatic tools are present to analyze a vast amount of data provided by the next-generation sequencing technologies, and new tools are still produced depending on the need of analysts. In this chapter, mainly the methodologies used in the recent studies analyzing the plant gene families were mentioned. The majority of these bioinformatic tools are accessible on the internet platforms; therefore, their website addresses were provided for ease of use. However, it should be noted that sometimes the links for these addresses might be broken or they might be moved to another address.

As the number of sequenced plant genomes/transcriptomes increased, more studies will be performed on the analysis of gene families. New analysis methods will give rise to the discovery of novel bioinformatic tools. Hereby, our understanding on the roles of these gene families will broaden, especially having an impact on the molecular breeding of stress-tolerant cultivars. Additionally, the knowledge obtained via in silico analyses will be used for the functional gene expression and/or gene silencing studies in plants.

References

Aoyagi LN, Lopes-Caitar VS, de Carvalho MCCG, Darben LM, Polizel-Podanosqui A, Kuwahara MK, Nepomuceno AL, Abdelnoor RV, Marcelino-Guimarães FC (2014) Genomic and transcriptomic characterization of the transcription factor family R2R3-MYB in soybean and its involvement in the resistance responses to *Phakopsora pachyrhizi*. Plant Sci 229:32–42

Ariyarathna HA, Ul-Haq T, Colmer TD, Francki MG (2014) Characterization of the multigene family *TaHKT 2;1* in bread wheat and the role of gene members in plant Na^+ and K^+ status. BMC Plant Biol 14:159

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40:W597–W603

Arya GC, Kumar R, Bisht NC (2014a) Evolution, expression differentiation and interaction specificity of heterotrimeric G-protein subunit gene family in the mesohexaploid *Brassica rapa*. PLoS One 9(9):e105771

Arya P, Kumar G, Acharya V, Singh AK (2014b) Genome-wide identification and expression analysis of NBS-encoding genes in *Malus x domestica* and expansion of NBS genes family in Rosaceae. PLoS One 9(9):e107987

Bailey TL, Bodén M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208

Ballester P, Navarrete-Gómez M, Carbonero P, Oñate-Sánchez L, Ferrándiz C (2015) Leaf expansion in *Arabidopsis* is controlled by a TCP-NGA regulatory module likely conserved in distantly related species. Physiol Plant doi. doi:10.1111/ppl.12327

Belamkar V, Weeks NT, Bharti AK, Farmer AD, Graham MA, Cannon SB (2014) Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (*Glycine max*) during dehydration and salt stress. BMC Genomics 15:950

Bencke-Malato M, Cabreira C, Wiebke-Strohm B, Bücker-Neto L, Mancini E, Osorio MB, Homrich MS, Turchetto-Zolet AC, De Carvalho MC, Stolf R, Weber RL, Westergaard G, Castagnaro AP, Abdelnoor RV, Marcelino-Guimarães FC, Margis-Pinheiro M, Bodanese-Zanettini MH (2014) Genome-wide annotation of the soybean WRKY family and functional

characterization of genes involved in response to *Phakopsora pachyrhizi* infection. BMC Plant Biol 14:236
Bouktila D, Khalfallah Y, Habachi-Houimli Y, Mezghani-Khemakhem M, Makni M, Makni H (2015) Full-genome identification and characterization of NBS-encoding disease resistance genes in wheat. Mol Genet Genomics 290(1):257–271
Cao J, Li X (2014) Identification and phylogenetic analysis of late embryogenesis abundant proteins family in tomato (*Solanum lycopersicum*). Planta. doi:10.1007/s00425-014-2215-y
Cao J, Han X, Zhang T, Yang Y, Huang J, Hu X (2014) Genome-wide and molecular evolution analysis of the subtilase gene family in *Vitis vinifera*. BMC Genomics 15:1116
Ceasar SA, Hodge A, Baker A, Baldwin SA (2014) Phosphate concentration and arbuscular mycorrhizal colonisation influence the growth, yield and expression of twelve *PHT1* family phosphate transporters in foxtail millet (*Setaria italica*). PLoS One 9(9):e108459
Chai G, Wang Z, Tang X, Yu L, Qi G, Wang D, Yan X, Kong Y, Zhou G (2014) *R2R3-MYB* gene pairs in *Populus*: evolution and contribution to secondary wall formation and flowering time. J Exp Bot 65(15):4255–4269
Chanroj S, Wang G, Venema K, Zhang MW, Delwiche CF, Sze H (2012) Conserved and diversified gene families of monovalent cation/H^+ antiporters from algae to flowering plants. Front Plant Sci 3:25
Charfeddine M, Saïdi MN, Charfeddine S, Hammami A, Gargouri Bouzid R (2014) Genome-wide analysis and expression profiling of the ERF transcription factor family in potato (*Solanum tuberosum* L.). Mol Biotechnol doi:10.1007/s12033-014-9828-z
Charfeddine S, Saïdi MN, Charfeddine M, Gargouri-Bouzid R (2015) Genome-wide identification and expression profiling of the late embryogenesis abundant genes in potato with emphasis on dehydrins. Mol Biol Rep. doi:10.1007/s11033-015-3853-2
Chen Y, Wang Y, Zhang H (2014) Genome-wide analysis of the mildew resistance locus o (*MLO*) gene family in tomato (*Solanum lycopersicum* L.). Plant Omics J 7(2):87–93
Chettoor AM, Givan SA, Cole RA, Coker CT, Unger-Wallace E, Vejlupkova Z, Vollbrecht E, Fowler JE, Evans MM (2014) Discovery of novel transcripts and gametophytic functions via RNA-seq analysis of maize gametophytic transcriptomes. Genome Biol 15(7):414
Darzentas N (2010) Circoletto: visualizing sequence similarity with Circos. Bioinformatics 26(20):2620–2621
de Oliveira LF, Christoff AP, de Lima JC, de Ross BC, Sachetto-Martins G, Margis-Pinheiro M, Margis R (2014) The Wall-associated Kinase gene family in rice genomes. Plant Sci 229:181–192
Deokar AA, Kondawar V, Kohli D, Aslam M, Jain PK, Karuppayil SM, Varshney RK, Srinivasan R (2015) Funct Integr Genomics 15(1):27–46
Ding M, Chen J, Jiang Y, Lin L, Cao Y, Wang M, Zhang Y, Rong J, Ye W (2015) Genome-wide investigation and transcriptome analysis of the WRKY gene family in *Gossypium*. Mol Genet Genomics 290(1):151–171
Duan W, Song X, Liu T, Huang Z, Ren J, Hou X, Du J, Li Y (2014) Patterns of evolutionary conservation of ascorbic acid-related genes following whole-genome triplication in *Brassica rapa*. Genome Biol Evol 7(1):299–313
Duan W, Song X, Liu T, Huang Z, Ren J, Hou X, Li Y (2015) Genome-wide analysis of the MADS-box gene family in *Brassica rapa* (Chinese cabbage). Mol Genet Genomics 290(1):239–255
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797
Fan K, Wang M, Miao Y, Ni M, Bibi N, Yuan S, Li F, Wang X (2014) Molecular evolution and expansion analysis of the NAC transcription factor in *Zea mays*. PLoS One 9(11):e111837
Feng L, Chen Z, Ma H, Chen X, Li Y, Wang Y, Xiang Y (2014a) The IQD gene family in soybean: structure, phylogeny, evolution and expression. PLoS One 9(10):e110896

Feng S, Yue R, Tao S, Yang Y, Zhang L, Xu M, Wang H, Shen C (2014b) Genome-wide identification, expression analysis of auxin-responsive *GH3* family genes in maize (*Zea mays* L.) under abiotic stresses. J Integr Plant Biol doi:10.1111/jipb.12327

Filiz E, Tombuloğlu H (2015) Genome-wide distribution of superoxide dismutase (SOD) gene families in *Sorghum bicolor*. Turk J Biol 39:49–59

Filiz E, Tombuloglu H, Ozyigit II (2013) Genome-wide analysis of IQ67 domain (*IQD*) gene families in *Brachypodium distachyon*. Plant Omics J 6(6):425–432

Filiz E, Koç İ, Tombuloğlu H (2014) Genome-wide identification and analysis of growth regulating factor genes in *Brachypodium distachyon*: in silico approaches. Turk J Biol 38:296–306

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Frech C, Chen N (2010) Genome-wide comparative gene family classification. PLoS One 5(10):e13409

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652

Ha CV, Esfahani MN, Watanabe Y, Tran UT, Sulieman S, Mochida K, Nguyen DV, Tran LS (2014) Genome-wide identification and expression analysis of the CaNAC family members in chickpea during development, dehydration and ABA treatments. PLoS One 9(12):e114107

He D, Lei Z, Xing H, Tang B (2014) Genome-wide identification and analysis of the *aldehyde dehydrogenase* (*ALDH*) gene superfamily of *Gossypium raimondii*. Gene 549(1):123–133

Hofberger JA, Zhou B, Tang H, Jones JD, Schranz ME (2014) A novel approach for multi-domain and multi-gene family identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants. BMC Genomics 15:966

Hou XJ, Li SB, Liu SR, Hu CG, Zhang JZ (2014) Genome-wide classification and evolutionary and expression analyses of citrus MYB transcription factor families in sweet orange. PLoS One 9(11):e112375

Hussey SG, Saïdi MN, Hefer CA, Myburg AA, Grima-Pettenati J (2014) Structural, evolutionary and functional analysis of the NAC domain protein family in *Eucalyptus*. New Phytol. doi:10.1111/nph.13139

Hyun TK, Eom SH, Han X, Kim JS (2014) Evolution and expression analysis of the soybean glutamate decarboxylase gene family. J Biosci 39(5):899–907

Ito TM, Polido PB, Rampim MC, Kaschuk G, Souza SG (2014) Genome-wide identification and phylogenetic analysis of the AP2/ERF gene superfamily in sweet orange (Citrus sinensis). Genet Mol Res 13(3):7839–7851

Jain M, Chevala VV, Garg R (2014) Genome-wide discovery and differential regulation of conserved and novel microRNAs in chickpea via deep sequencing. J Exp Bot 65(20):5945–5958

Jali SS, Rosloski SM, Janakirama P, Steffen JG, Zhurov V, Berleth T, Clark RM, Grbic V (2014) A plant-specific *HUA2-LIKE* (*HULK*) gene family in *Arabidopsis thaliana* is essential for development. Plant J 80(2):242–254

Jiang Y, Duan Y, Yin J, Ye S, Zhu J, Zhang F, Lu W, Fan D, Luo K (2014) Genome-wide identification and characterization of the *Populus* WRKY transcription factor family and analysis of their expression in response to biotic and abiotic stresses. J Exp Bot 65(22):6629–6644

Kayum MA, Jung HJ, Park JI, Ahmed NU, Saha G, Yang TJ, Nou IS (2015) Identification and expression analysis of *WRKY* family genes under biotic and abiotic stresses in *Brassica rapa*. Mol Genet Genomics 290(1):79–95

Kenzior A, Folk WR (2015) *Arabidopsis thaliana* MSI4/FVE associates with members of a novel family of plant specific PWWP/RRM domain proteins. Plant Mol Biol doi. doi:10.1007/s11103-014-0280-z

Kim Y-W, Jung H-J, Park J-I, Hur Y, Nou I-S (2015) Response of NBS encoding resistance genes linked to both heat and fungal stress in *Brassica oleracea*. Plant Physiol Biochem 86:130–136

Lata C, Mishra AK, Muthamilarasan M, Bonthala VS, Khan Y, Prasad M (2014) Genome-wide investigation and expression profiling of AP2/ERF transcription factor superfamily in foxtail millet (*Setaria italica* L.). PLoS One 9(11):e113092

Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. Nucleic Acids Res 43:D257–D260

Li C, Lu S (2014) Molecular characterization of the SPL gene family in *Populus trichocarpa*. BMC Plant Biol 14:131

Li PS, Yu TF, He GH, Chen M, Zhou YB, Chai SC, Xu ZS, Ma YZ (2014a) Genome-wide analysis of the Hsf family in soybean and functional identification of *GmHsf-34* involvement in drought and heat stresses. BMC Genomics 15:1009

Li Q, Zhang N, Zhang L, Ma H (2014b) Differential evolution of members of the rhomboid gene family with conservative and divergent patterns. New Phytol. doi:10.1111/nph.13174

Lin Q, Jiang Q, Lin J, Wang D, Li S, Liu C, Sun C, Chen K (2015) Heat shock transcription factors expression during fruit development and under hot air stress in Ponkan (Citrus reticulata Blanco cv. Ponkan) fruit. Gene pii: S0378-1119(15)00040-2

Liu RH, Meng JL (2003) MapDraw: a microsoft excel macro for drawing genetic linkage maps based on given genetic linkage data. Yi Chuan 25(3):317–321

Ma H, Feng L, Chen Z, Chen X, Zhao H, Xiang Y (2014a) Genome-wide identification and expression analysis of the *IQD* gene family in *Populus trichocarpa*. Plant Sci 229:96–110

Ma J, Wang F, Li M-Y, Jiang Q, Tan G-F, Xiong A-S (2014b) Genome wide analysis of the NAC transcription factor family in Chinese cabbage to elucidate responses to temperature stress. Scientia Horticulturae 165:82–90

Ma T, Ma H, Zhao H, Qi H, Zhao J (2014c) Identification, characterization, and transcription analysis of xylogen-like arabinogalactan proteins in rice (*Oryza sativa* L.). BMC Plant Biol 14:299

Mace E, Tai S, Innes D, Godwin I, Hu W, Campbell B, Gilding E, Cruickshank A, Prentis P, Wang J, Jordan D (2014) The plasticity of NBS resistance genes in sorghum is driven by multiple evolutionary processes. BMC Plant Biol 14:253

Mainali HR, Chapman P, Dhaubhadel S (2014) Genome-wide analysis of *Cyclophilin* gene family in soybean (*Glycine max*). BMC Plant Biol 14(1):282

Martinez M (2013) From plant genomes to protein families: computational tools. Comput Struct Biotechnol J 8:e201307001

Matus JT, Aquea F, Espinoza C, Vega A, Cavallini E, Dal Santo S, Cañón P, Rodríguez-Hoces de la Guardia A, Serrano J, Tornielli GB, Arce-Johnson P (2014) Inspection of the grapevine BURP superfamily highlights an expansion of *RD22* genes with distinctive expression features in berry development and ABA-mediated stress responses. PLoS One 9(10):e110372

Mochida K, Shinozaki K (2011) Advances in omics and bioinformatics tools for systems analyses of plant functions. Plant Cell Physiol 52(12):2017–2038

Muthamilarasan M, Khandelwal R, Yadav CB, Bonthala VS, Khan Y, Prasad M (2014) Identification and molecular characterization of MYB transcription factor superfamily in C_4 model plant foxtail millet (*Setaria italica* L.). PLoS One 9(10):e109920

Nawaz Z, Kakar KU, Saand MA, Shu QY (2014) Cyclic nucleotide-gated ion channel gene family in rice, identification, characterization and experimental analysis of expression response to plant hormones, biotic and abiotic stresses. BMC Genomics 15:853

Nguyen QN, Lee YS, Cho LH, Jeong HJ, An G, Jung KH Genome-wide identification and analysis of *Catharanthus roseus* RLK1-like kinases in rice. Planta doi:10.1007/s00425-014-2203-2

Okay S, Derelli E, Unver T (2014) Transcriptome-wide identification of bread wheat WRKY transcription factors in response to drought stress. Mol Genet Genomics 289(5):765–781

Pan X, Peng FY, Weselake R (2015) Genome-wide analysis of *PHOSPHOLIPID:DIACYLGLYCEROL ACYLTRANSFERASE* genes in plants reveals the eudicot-wide *PDAT* gene expansion and altered selective pressures acting on the core eudicot *PDAT* paralogs. Plant Physiol pii: pp.114.253658

Panahi B, Abbaszadeh B, Taghizadeghan M, Ebrahimie E (2014) Genome-wide survey of alternative splicing in *Sorghum bicolor*. Physiol Mol Biol Plants 20(3):323–329

Pandey B, Kaur A, Gupta OP, Sharma I, Sharma P (2014) Identification of *HSP20* gene family in wheat and barley and their differential expression profiling under heat stress. Appl Biochem Biotechnol. doi:10.1007/s12010-014-1420-2

Pessina S, Pavan S, Catalano D, Gallotta A, Visser RG, Bai Y, Malnoy M, Schouten HJ (2014) Characterization of the *MLO* gene family in Rosaceae and gene expression analysis in *Malus domestica*. BMC Genomics 15:618

Pourabed E, Ghane Golmohamadi F, Soleymani Monfared P, Razavi SM, Shobbar ZS (2015) Basic leucine zipper family in barley: genome-wide characterization of members and expression analysis. Mol Biotechnol 57(1):12–26

Rawal HC, Singh NK, Sharma TR (2013) Conservation, divergence, and genome-wide distribution of *PAL* and *POX A* gene families in plants. Int J Genomics 2013:678969

Saha J, Sengupta A, Gupta K, Gupta B (2015) Molecular phylogenetic study and expression analysis of ATP-binding cassette transporter gene family in *Oryza sativa* in response to salt stress. Comput Biol Chem 54:18–32

Saito F, Suyama A, Oka T, Yoko-O T, Matsuoka K, Jigami Y, Shimma YI (2014) Identification of novel peptidyl serine α-galactosyltransferase gene family in plants. J Biol Chem 289:20405–20420

Shao Y, Qin Y, Zou Y, Ma F (2014) Genome-wide identification and expression profiling of the SnRK2 gene family in *Malus prunifolia*. Gene 552(1):87–97

Sharma R, Suresh CG (2015) Genome-wide identification and structure-function studies of proteases and protease inhibitors in *Cicer arietinum* (chickpea). Comput Biol Med 56:67–81

Sharma P, Lin T, Grandellis C, Yu M, Hannapel DJ (2014a) The BEL1-like family of transcription factors in potato. J Exp Bot 65(2):709–723

Sharma R, Rawat V, Suresh CG (2014b) Genome-wide identification and tissue-specific expression analysis of UDP-glycosyltransferases genes confirm their abundance in *Cicer arietinum* (chickpea) genome. PLoS One 9(10):e109715

Shen C, Yue R, Yang Y, Zhang L, Sun T, Xu L, Tie S, Wang H (2014) Genome-wide identification and expression profiling analysis of the *Aux/IAA* gene family in *Medicago truncatula* during the early phase of *Sinorhizobium meliloti* infection. PLoS One 9(9):e107495

Shiriga K, Sharma R, Kumar K, Yadav SK, Hossain F, Thirunavukkarasu N (2014) Genome-wide identification and expression pattern of drought-responsive members of the NAC family in maize. Meta Gene 2:407–417

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Siriwardana CL, Kumimoto RW, Jones DS, Holt BF III (2014) Gene family analysis of the *Arabidopsis NF-YA* transcription factors reveals opposing abscisic acid responses during seed germination. Plant Mol Biol Rep 32(5):971–986

Soler M, Camargo EL, Carocha V, Cassan-Wang H, San Clemente H, Savelli B, Hefer CA, Paiva JA, Myburg AA, Grima-Pettenati J (2014) The *Eucalyptus grandis* R2R3-MYB transcription factor family: evidence for woody growth-related evolution and function. New Phytol. doi:10.1111/nph.13039

Sun H, Fan H-J, Ling H-Q (2015) Genome-wide identification and characterization of the bHLH gene family in tomato. BMC Genomics 16:9

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

Tan Y, Wang S, Liang D, Li M, Ma F (2014) Genome-wide identification and expression profiling of the cystatin gene family in apple (*Malus* × *domestica* Borkh.). Plant Physiol Biochem 79:88–97

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815

Thomas M, Pingault L, Poulet A, Duarte J, Throude M, Faure S, Pichon JP, Paux E, Probst AV, Tatout C (2014) Evolutionary history of Methyltransferase 1 genes in hexaploid wheat. BMC Genomics 15:922

Tian Y, Dong Q, Ji Z, Chi F, Cong P, Zhou Z (2015) Genome-wide identification and analysis of the MADS-box gene family in apple. Gene 555(2):277–290

Trujillo DI, Silverstein KA, Young ND (2014) Genomic characterization of the LEED.PEEDs, a gene family unique to the Medicago lineage. G3 (Bethesda) 4(10):2003–2012

Wang F, Qiu N, Ding Q, Li J, Zhang Y, Li H, Gao J (2014a) Genome-wide identification and analysis of the growth-regulating factor family in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). BMC Genomics 15:807

Wang G, Lovato A, Polverari A, Wang M, Liang YH, Ma YC, Cheng ZM (2014b) Genome-wide identification and analysis of mitogen activated protein kinase kinase kinase gene family in grapevine (*Vitis vinifera*). BMC Plant Biol 14:219

Wang J, Sun N, Deng T, Zhang L, Zuo K (2014c) Genome-wide cloning, identification, classification and functional analysis of cotton heat shock transcription factors in cotton (*Gossypium hirsutum*). BMC Genomics 15:961

Wang L, Yin X, Cheng C, Wang H, Guo R, Xu X, Zhao J, Zheng Y, Wang X (2014d) Evolutionary and expression analysis of a MADS-box gene superfamily involved in ovule development of seeded and seedless grapevines. Mol Genet Genomics. doi:10.1007/s00438-014-0961-y

Wang L, Yu C, Chen C, He C, Zhu Y, Huang W (2014e) Identification of rice Di19 family reveals *OsDi19-4* involved in drought resistance. Plant Cell Rep 33(12):2047–2062

Wang X, Zhang H, Gao Y, Sun G, Zhang W, Qiu L (2014f) A comprehensive analysis of the Cupin gene family in soybean (*Glycine max*). PLoS One 9(10):e110092

Wang J, Chu S, Zhu Y, Cheng H, Yu D (2015a) Positive selection drives neofunctionalization of the UbiA prenyltransferase gene family. Plant Mol Biol. doi:10.1007/s11103-015-0285-2

Wang Z, Tang J, Hu R, Wu P, Hou XL, Song XM, Xiong AS (2015b) Genome-wide analysis of the R2R3-MYB transcription factor genes in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*) reveals their stress and hormone responsive patterns. BMC Genomics 16(1):17

Wei K, Pan S (2014) Maize protein phosphatase gene family: identification and molecular characterization. BMC Genomics 15:773

Wei B, Zhang RZ, Guo JJ, Liu DM, Li AL, Fan RC, Mao L, Zhang XQ (2014a) Genome-wide analysis of the MADS-box gene family in *Brachypodium distachyon*. PLoS One 9(1):e84781

Wei X, Liu F, Chen C, Ma F, Li M (2014b) The *Malus domestica* sugar transporter gene family: identifications based on genome and expression profiling related to the accumulation of fruit sugars. Front Plant Sci 5:569

Wen F, Zhu H, Li P, Jiang M, Mao W, Ong C, Chu Z (2014) Genome-wide evolutionary characterization and expression analyses of WRKY family genes in *Brachypodium distachyon*. DNA Res 21(3):327–339

Wu J, Wang J, Pan C, Guan X, Wang Y, Liu S, He Y, Chen J, Chen L, Lu G (2014) Genome-wide identification of MAPKK and MAPKKK gene families in tomato and transcriptional profiling analysis during development and stress response. PLoS One 9(7):e103032

Xie R, Li Y, He S, Zheng Y, Yi S, Lv Q, Deng L (2014) Genome-wide analysis of citrus *R2R3MYB* genes and their spatiotemporal expression under stresses and hormone treatments. PLoS One 9(12):e113971

Xu R, Sun P, Jia F, Lu L, Li Y, Zhang S, Huang J (2014) Genomewide analysis of *TCP* transcription factor gene family in *Malus domestica*. J Genet 93(3):733–746

Yang J, Zhang N, Mi X, Wu L, Ma R, Zhu X, Yao L, Jin X, Si H, Wang D (2014) Identification of miR159s and their target genes and expression analysis under drought stress in potato. Comput Biol Chem 53PB:204–213

Yang Y, Yue R, Sun T, Zhang L, Chen W, Zeng H, Wang H, Shen C (2015) Genome-wide identification, expression analysis of *GH3* family genes in *Medicago truncatula* under stress-related hormones and *Sinorhizobium meliloti* infection. Appl Microbiol Biotechnol 99(2):841–854

Yao QY, Xia EH, Liu FH, Gao LZ (2015) Genome-wide identification and comparative expression analysis reveal a rapid expansion and functional divergence of duplicated genes in the WRKY gene family of cabbage. Brassica oleracea var. capitata. Gene 557(1):35–42

Yu H, Soler M, Mila I, San Clemente H, Savelli B, Dunand C, Paiva JA, Myburg AA, Bouzayen M, Grima-Pettenati J, Cassan-Wang H (2014) Genome-wide characterization and expression profiling of the *AUXIN RESPONSE FACTOR* (*ARF*) gene family in *Eucalyptus grandis*. PLoS One 9(9):e108906

Yu H, Soler M, San Clemente H, Mila I, Paiva JA, Myburg AA, Bouzayen M, Grima-Pettenati J, Cassan-Wang H (2015) Comprehensive genome-wide analysis of the *Aux/IAA* gene family in *Eucalyptus*: evidence for the role of *EgrIAA4* in wood formation. Plant Cell Physiol pii:pcu215

Yurchenko OP, Park S, Ilut DC, Inmon JJ, Millhollon JC, Liechty Z, Page JT, Jenks MA, Chapman KD, Udall JA, Gore MA, Dyer JM (2014) Genome-wide analysis of the omega-3 fatty acid desaturase gene family in *Gossypium*. BMC Plant Biol 14:312

Zhang W, Yan H, Chen W, Liu J, Jiang C, Jiang H, Zhu S, Cheng B (2014a) Genome-wide identification and characterization of maize expansin genes expressed in endosperm. Mol Genet Genomics 289(6):1061–1074

Zhang X, Wang L, Xu X, Cai C, Guo W (2014b) Genome-wide identification of mitogen-activated protein kinase gene family in *Gossypium raimondii* and the function of their corresponding orthologs in tetraploid cultivated cotton. BMC Plant Biol 14(1):345

Zhang Y, Yang S, Song Y, Wang J (2014c) Genome-wide characterization, expression and functional analysis of *CLV3/ESR* gene family in tomato. BMC Genomics 15:827

Zhang Z, Chen X, Guan X, Liu Y, Chen H, Wang T, Mouekouba LD, Li J, Wang A (2014d) A genome-wide survey of homeodomain-leucine zipper genes and analysis of cold-responsive HD-Zip I members' expression in tomato. Biosci Biotechnol Biochem 78(8):1337–1349

Zhang L, Li Q, Dong H, He Q, Liang L, Tan C, Han Z, Yao W, Li G, Zhao H, Xie W, Xing Y (2015a) Three CCT domain-containing genes were identified to regulate heading date by candidate gene-based association mapping and transformation in rice. Sci Rep 5:7663

Zhang X, Dou L, Pang C, Song M, Wei H, Fan S, Wang C, Yu S (2015b) Genomic organization, differential expression, and functional analysis of the *SPL* gene family in *Gossypium hirsutum*. Mol Genet Genomics 290(1):115–126

Zhou P, Silverstein KAT, Gao L, Walton JD, Nallu S, Guhlin J, Young ND (2013) Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). BMC Bioinformatics 14:335

Zhu C, Luo N, He M, Chen G, Zhu J, Yin G, Li X, Hu Y, Li J, Yan Y (2014a) Molecular characterization and expression profiling of the protein disulfide isomerase gene family in *Brachypodium distachyon* L. PLoS One 9(4):e94704

Zhu YB, Xie XQ, Li ZY, Bai H, Dong L, Dong ZP, Dong JG (2014b) Bioinformatic analysis of the nucleotide binding site-encoding disease-resistance genes in foxtail millet (*Setaria italica* (L.) Beauv.). Genet Mol Res 13(3):6602–6609

Epigenetics and Applications in Plants

Çağatay Tarhan and Neslihan Turgut-Kara

Contents

Abstract As seen in other eukaryotic cells, DNA is coiled tightly around the histone proteins in plant cells. Pathways that end with cytosine DNA methylation, posttranslational histone modifications, and RNA interference (RNAi) contribute importantly to the regulation of chromatin structure and hence affect many cellular events. High-throughput sequencing analysis on a genome-wide scale brings new understanding about plant genomes and the functions of epigenetic pathways. Although epigenetics has become an important research field in the post-genomic era and even though we can use many model organisms whose epigenomes have been sequenced for many years, we still far from having full knowledge about the regulation of gene expression. This chapter mainly focuses on the characteristics of the field of epigenetics and its applications in plants.

Keywords Epigenetics • Plant • Next-generation sequencing • Methylation

Ç. Tarhan • N. Turgut-Kara (✉)
Department of Molecular Biology and Genetics, Faculty of Science,
Istanbul University, Istanbul, Turkey
e-mail: neslihantk@istanbul.edu.tr

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_10

1 Introduction

Epigenetics involve the changes in gene expression that do not stem from the changes in DNA sequences but which are hereditary at the same time. Thus, it is a regulating process at a level above that of classical genetic mechanisms. Although the field acquired its name about 50 years ago, it has become a rapidly developing discipline only in recent years. As a result, it has caused major changes in the classical view of inheritance.

Epigenetics has become one of the most attractive research topics in molecular biology because it appears encouraging for the decoding of the control mechanisms of gene expression. Many cellular processes that could be categorized as "classical genetics themes," such as transcription, replication, DNA repair, gene transposition, and cell differentiation, are generally controlled by the elements such as promoters, enhancers, and inducer or repressor proteins. In addition, many molecular processes that take place in a cell can also be regulated by epigenetic mechanisms. Epigenetic regulation is mediated by DNA methylation, histone modifications, histone variants, chromatin remodeling, small RNAs, and transposable elements (TEs), which are involved in the transcriptional and posttranscriptional control of gene expression (Fig. 1). In addition to these control mechanisms, epigenetic modifications function in the regulation of noncoding DNA sequences, thereby assuring the safety and the stability of genomes. For instance, deactivating certain DNA regions such as centromeres and telomeres enables the microtubules to attach correctly to the components of the cellular skeleton. Thus, reducing unwanted recombination events and intercepting TE transpositions prohibits insertional mutagenesis (Dupont et al. 2009), showing that epigenetic modifications affect genome expression by changing the chromatin structure. Among these modifications, histone variants, histone posttranslational modifications, and DNA methylation could be considered (Chinnusamy and Zhu 2009).

As seen in other eukaryotes, epigenetic mechanisms have been implicated in the regulation of plant gene expression. Methylation is one of the epigenetic mechanisms that involves methylation of cytosine residues present in specific parts of the DNA molecule (Bestor 2000; Bird 2002). The enzymes that carry out the methylation reaction have been well characterized (Okano et al. 1999), as is the mechanism by which the configuration of methylated positions is propagated through DNA replication (Groth et al. 2007). The best known result of a methylation event in a genomic region is the repression of the genes that are located in the affected region (Bird 2002).

Another kind of epigenetic modification occurs at the level of chromatin. Chromatins consist of DNA and histone octamers. Wrapping the DNA molecule around the histone proteins forms the chromatin structure, and this structure may prevent the DNA transcription events from taking place because parts of the DNA are inaccessible. Through some chemical reactions such as acetylation, methylation, sumoylation, and ubiquitylation, however, modified histone proteins can make the DNA accessible and enable DNA transcription (Kouzarides 2007).

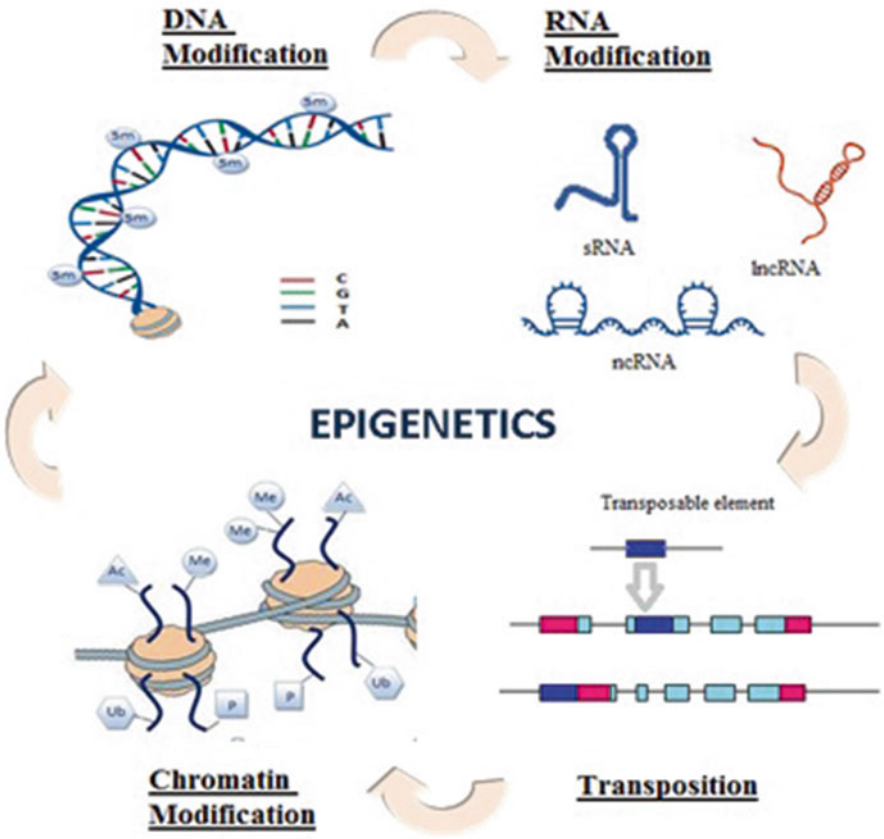


Fig. 1 Summary of the basic epigenetic mechanisms

One of the mechanisms of epigenetic regulation is noncoding RNAs (ncRNAs). These molecules can be modified and function in RNA interference pathways. As a result of this process small RNA molecules are formed, and these molecules can interact with an immature RNA or DNA molecule. These interactions can cause inhibition of the gene expression process (Zaratiegui et al 2007; Mattick et al 2009). Also, some other ncRNA molecules interfere in silencing of large chromosome segments, and consequently these segments become transcriptionally inactive (Clark 2007; Yang and Kuroda 2007).

Studies of interactions between DNA methylation and histone modification in plants have confirmed their significant role in jointly silencing gene expression (Fuks 2005). In addition, histone modifications are more easily modified than epigenetic mark-DNA methylation that is more stable and resistant to change. Plants can protect their genomes from detrimental mobile elements, viruses, and viroids using a small RNA silencing mechanism. In this process, DNA methylation and histone modification are crucial chemical modifications. With these mechanisms,

plants are also able to regulate the expression of certain genes in response to environmental changes (Chen et al. 2010).

To understand the epigenetic mechanisms in both animals and plants in more detail, multidisciplinary approaches have been included. These approaches cover especially a combination of genetic, genomic, biochemical, and computational sciences. Strahl and Allis (2000) suggested the "histone code hypothesis" that histone modifications occur at selected residues of histone protein tails, and some of the patterns shown have been closely linked to a biological response. Since then, a great many histone-modifying proteins have been discovered, but there are many more histone proteins to be characterized, and correspondingly many more histone modification mechanisms remain to be described. Although we understand the dynamics of many of the histone modifiers yet we do not know how these modifications are maintained during continuous cell division. The role of small RNAs in gene silencing has also been documented, but we do not know enough about maintaining small RNA activity during the cell division process. In animals, the embryo is the structure where the progressional program is designed but in plants, this design is made after the embryonic stage through an unclear mechanism. In a similar manner, it is found that some epigenetic regulators (histone modifications and chromatin remodeling) organize the RNA splicing process, but there are underlying molecular mechanisms to be explained. Taken together, these findings suggest that there is much more to be discovered regarding unknown epigenetic mechanisms (Ahmad et al. 2010). Packaging DNA and histone proteins and thereby forming chromatin is not only an efficient way of packaging hereditary material but also serves as a basic structure for a variety of regulating mechanisms.

The importance of histone modifications especially comes from their contribution to the regulation of gene expression, coordination of development, and evolution of genome functionality. Currently, most of the studies focus on the mechanism of this contribution. To that end, high-throughput and high-resolution technologies such as sequencing and microarray have been used as very efficient tools. When the *Arabidopsis* genome was first sequenced, researchers started to obtain an overall understanding of a plant. Besides, the genomic data obtained provided a fruitful substructure for determining gene functions. Today we have more than 100 sequenced plant genomes: some of these belong to the group of model organisms, and others are crop species. These sequenced genomes are especially important for annotations such as the determination of gene loci and mapping of the quantitative trait loci (QTL). Advances in this field have led to initiation of a plant ENCODE (pENCODE) project. Some of the findings from the human ENCODE project are of direct interest to plant scientists. For example, the epigenomic maps of different developmental stages contribute to understanding of the genotype–phenotype relationship. In this respect the integration of transcription factor-binding sites, RNA expression states, DNase I hypersensitivity sites, and chromatin modification maps are of special interest to a plant molecular biologist. Another relevant finding in the human ENCODE project was the identification of large numbers of trait-associated sequence variants localized to regulatory DNA elements (Maurano et al. 2012; Lane et al. 2014). Combined with the generated genome-wide maps of sequence

variation, RNA types, chromatin modifications, protein–DNA interactions, and inter- or intrachromosomal interactions, these ENCODE projects have also developed the protocols required to generate these data, the software required to analyze them, and the genome browsers required to visualize them (Lane et al. 2014).

One of the key features of pENCODE is that plants provide an ideal organism to study how the environment interacts with the genome to coordinate phenotypic changes. Plant species do not contain a nervous system but instead take advantage of a complex transcriptional regulatory code to execute many of the same responses that animals experience; this is partly exemplified by the massive expansion of transcription factor (TF) families present in plant genomes. Plant genomes also offer an excellent system to understand how genomes manage newly duplicated sequences, such as genes, chromosomes, or genomes (Lane et al. 2014).

Plants are very convenient organisms for epigenomics research. In contrast to certain model organisms such as yeast, nematodes and fruit flies, plants use DNA methylation frequently, as seen in humans. Also, plants use histone modifications and gene silencing as defense mechanisms against mobile elements and viruses. In *Arabidopsis*, small RNAs such as microRNAs (miRNAs) and small interfering RNAs (siRNAs) are synthesized through genetically determined pathways (Chapman and Carrington 2007). Null mutations in many chromatin regulators are lethal in animals. But as an important feature, plants can cope with these kinds of mutations despite undergoing similarly complex developmental transitions as animals. Genome sequences of many plants such as *Arabidopsis*, rice, and maize are available; thus, they are potent sources for studying genome-wide analyses of DNA methylation, histone modifications, and their relationships to coding as well as noncoding RNAs. These features make plants very useful model organisms in the fields of epigenetics and epigenomics (Epigenomics of Plants International Consortium (EPIC) 2015). This chapter focused on the epigenetic studies and applications that have been successfully performed on plants.

2 DNA Methylation Analysis

Epigenetic studies in plants can be grounded on the discovery of epigenetic regulation of genetic control in plant development and stress tolerance. As one of these epigenetic regulation mechanisms, DNA methylation has a pivotal role in regulating gene expression regulation and directing other epigenetic mechanisms to function at the right time and place (Chen et al. 2010).

There are plenty of methods for determining methylated cytosine, and these methods can be divided into two groups. The first is the bisulfate treatment method. Unmethylated cytosine can be converted to uracil while methylcytosine remains unmodified. In a polymerase chain reaction (PCR), uracil is a template-like thymidine and all original cytosine residues are converted to thymidine. Ultimately, the PCR products are sequenced and mapped onto genome data to locate unchanged cytosine, which is the methylcytosine. The advantage of this method is that DNA

integrity can be maintained; on the other hand, the rate of conversion must be controlled to minimize bias. The second class is a biological process that requires methylation-sensitive enzymes such as HpaII and NotI. The DNA digestion process is often performed with isoschizomers as well. Therefore, one of the two enzymes is methylation loci sensitive and the other recognizes the same DNA sequence but is independent of methylcytosine. DNA methylation loci can be determined through analysis of the different fragment patterns. This method is easy to perform and robust. However, failure to properly control the activity of the isoschizomers on their targets could lead to incomplete DNA digestion. The development of new sequencing technologies has renewed interest in these methods (Chen et al. 2010).

DNA methylation in mammals can be seen only in symmetrical CG sequences (Bird 2002), whereas plant DNA methylation is found at CG sites, CHG sites (H indicating A, C, or T), and CHH sites (an asymmetrical site). Genome-wide analysis of DNA methylation in plants by McrBC digestion, methylcytosine immunoprecipitation (MeDIP), or sequencing of bisulfate-treated DNA indicates that transposons are heavily methylated at both CG and non-CG sites, whereas the density of methylcytosines in genes is much lower and limited to CG sites (Zhang et al. 2006; Zilberman and Henikoff 2007; Cokus et al. 2008; Lister et al. 2008). Methyltransferase 1 (MET1), the homologue of mammalian DNMT1 (mammalian DNA methyltransferase 1), maintains DNA methylation at CG sites (Finnegan and Dennis 1993; Kankel et al. 2003). Plant-specific DNA methyltransferase, chromomethylase 3 (CMT3), can maintain DNA methylation at CHG sites (Jackson et al. 2002). Domains rearranged methylase 1 and 2 (DRM1/2), together with siRNAs, maintain DNA methylation at CHH sites (Zhang 2008). Analysis of rice gene methylation also showed that the rice genome contains more methylated promoters than does the *Arabidopsis* genome (Li et al. 2008), raising an interesting question about the comparability of dicotyledon and monocotyledon DNA methylation patterns. High-throughput DNA methylation sequencing in *Arabidopsis thaliana*, *Oryza sativa*, and maize genomes also revealed that gene coding regions may be candidates of moderate methylation (Zhang et al. 2006; Zilberman and Henikoff 2007; Cokus et al. 2008; Li et al. 2008; Lister et al. 2008; Wang et al. 2009). Although DNA methylation is associated with silencing genes, it has been shown that methylated housekeeping genes in coding regions have a higher level of expression (Zhang et al. 2006). A remarkable finding is that protein-coding gene methylation is polymorphic in different *Arabidopsis* ecotypes (Vaughn et al. 2007), suggesting there are some methylation patterns, such as DNA methylation in promoters, that can influence gene expression levels and are pivotal for plant development. It is now well established that DNA methylation can induce chromatin remodeling (Woo and Richards 2008).

Recent research revealed that the *ibm1* mutation results in hypermethylation in thousands of genes (Miura et al. 2009). Gene methylation is restricted to CG sites in wild-type *Arabidopsis*, whereas many CHG sites are methylated in the *ibm1* mutation. These data suggest that one or more critical factors are pivotal in directing

DNA methylation to correct gene sites and preventing other methyltransferases from modifying gene regions such as the CHG or CHH sites. In conclusion, there is strong evidence that disparate factors work in an orderly manner to control gene expression by adding methyl groups to the correct targets (Chen et al. 2010).

3 Histone Modification Analysis

The expression of genes can be regulated by histone modifications (Vaillant and Paszkowski 2007). Moreover, it has been shown in a wide range of plant species that these modifications are crucial in plant development and plant defense processes (Sokol et al. 2007; Kim et al. 2008). Interestingly, studies on these modifications have shown that histone modifications can interact with each other or DNA methylation (Chen et al. 2010). When examining plant development, studies have moved beyond classical genetic approaches and investigated especially the epigenetic modification of chromatin. Modifications in chromatin structure can robustly influence the organ development and physiological response to environmental changes and stress conditions such as salinity and temperature. These modifications can be inherited by daughter cells, and in this way maintenance of transcriptional states and proper responses to environmental changes is ensured. Chromatin remodeling is a well-conserved mechanism in a wide variety of eukaryotic organisms. On the other hand, because of the postembryonic development of plants and their efficient adaptive response to environmental changes, the flexibility of the chromatin modification in plants probably greater than in animals. The abundance of complex constituents of chromatin modification process could confer this flexibility to plants. Thus, loss-of-function mutants can be viable. The increasing number of sequenced genomes, applicability of high-throughput investigations, and the feasibility of many reverse genetic tools make possible distinguishing the chromatin modifiers and to study the interactions between chromatin complexes, thereby enabling understanding their effect on the regulation of transcriptional activity. Results from these studies can shed more light on the nature of chromatin complexes, formation and maintenance of chromatin modifications and epigenetic marks, and the effect of the different combinatorial chromatin complexes on the developmental plasticity of plants (Jarillo et al. 2009).

For the determination of histone modifications, antibody techniques such as Western blot and immunocytochemistry were used for a long time. However, because of the diversity of modification patterns, designing a specific sort of antibodies that can be functional in genome-wide analysis is not so easy (Villar-Garea and Imhof 2006). On the other hand, the emergence of certain methods that use analytical techniques enables the utilization of mass spectrometry for the detection of chromatin modification locations. For example, Villar-Garea et al. (2008) reported using high-resolution mass spectrometry as a tool for studying histone posttranslational modifications. This technique is also used for histone variant analysis.

4 Noncoding RNAs

The noncoding RNAs (ncRNAs) are functional molecules that do not encode a protein. These RNAs are highly abundant and consist of a heterogeneous group of RNA molecules. Their location, length, and biological functions can be used as classifying criteria (Costa 2005; Ben Amor et al. 2009). Ranging from small to long noncoding RNAs, these molecules function in the regulation of gene expression, genome stability, and defense against foreign genetic elements. For example, small RNAs can modify the structural properties of chromatin, thereby silencing transcription. To do this, they guide argonaute-containing complexes to target sites in partially complementary RNA scaffolds and then mediate the recruitment of histone and DNA methyltransferases. In addition, recent studies suggest that, independent from small RNAs, chromatin-modifying complexes can also be recruited by long noncoding RNAs. Thanks to these silencing mechanisms, powerful RNA regulation systems can detect and silence abnormal transcription events and provide a memory of these events via self-reinforcing epigenetic loops (Holoch and Moazed 2015).

RNA-directed DNA methylation (RdDM) was first observed in tobacco plants (Wassenegger et al. 1994) and siRNAs involved in this RdDM event. It was reported that there is a correlation between the small RNA population and DNA methylation patterns in *Arabidopsis* floral tissues. The quantity of small RNAs to be methylated are 25 times greater than sequences without small RNA matches. This finding indicates that DNA methylation occurs genome wide, but only one third of total methylated loci are related to small RNAs (Lister et al. 2008). Therefore, there must be another mechanism responsible for the remaining two thirds of genomic cytosine methylation. Chromatin remodeling and certain histone modifications might be candidate mechanisms for DNA methyltransferases to be directed to the loci of interest (Zhu 2008).

5 Transposable Elements

Transposable elements and other repetitive elements, also known as jumping genes, are mobile pieces of DNA sequences that have the ability to move through the genome. They can change their position in the genome and transfer themselves into protein-coding or regulator regions. As a result, the gene expression pattern can be affected in an unfavorable manner and cause, for example, insertional inactivation of the gene of interest. To prevent this adverse effect, silencing and immobilizing of these mobile elements need to be ensured. RNA interference and epigenetic DNA methylation events are responsible for this preventive effectiveness (Suzuki and Bird 2008; Sekhon and Chopra 2009). It has been shown that the inverted repeats at the end of TEs can be methylated when these TEs are transformed in *Arabidopsis* (Chen et al. 2010). DNA methylation is not a completely stable modification, and the methylated regions can be actively demethylated (Zhu et al. 2007; Zhu 2008).

In addition to these mechanisms, it has been reported in *Arabidopsis* that there is a transposon-silencing mark which is responsible for the suppression of the transposon replication. In this way, the plant ensures its genome stability. This mechanism may also take place in actively dividing cells as a control process for genome stability (Feng et al. 2010; Jacob et al. 2010).

6 Next-Generation Sequencing Technologies

Genome sequencing projects have taken advantage of a variety of technologies, and later these technologies could be applied to various epigenetic investigations such as DNA methylation patterns, posttranslational modifications of histones, and nucleosome positioning on a genome-wide scale (Rival et al. 2008). In this regard, the pENCODE project enabled obtaining the genome-wide maps of DNA methylation in many plant species (Lane et al. 2014). The time needed for sequencing the human genome using the conventional Sanger Method was 3 to 4 years and the cost was about $300 million. On the other hand, next-generation sequencing (NGS) technologies are as much as 200 times faster and much less expensive than the Sanger method. Moreover, this technique produces more accurate and reliable results. These properties make this technique very attractive, especially for individual researchers. Next-generation sequencing enables comprehensive analysis of genomes, epigenomes, transcriptomes, and interactomes to become much more cheaper and widespread, and thus can make a major contribution to the area of biological and biomedical research (Shendure and Ji 2008; Metzker 2010). So far, multiple high-throughput sequencing studies have reported to generate epigenomic data to understand the regulation of gene expression and resulting morphological variations. The two main classes of epigenetic modification data types are methylation by covalent modification of cytosine-5′ in DNA and posttranslational modifications of histone proteins (Callinan and Feinberg 2006). Epigenetic changes in gene expression and regulation do not stem from the changes in DNA sequences. The other tools for the epigenetic control of the gene expression are small RNAs. These kinds of epigenetic changes are very important, especially in plant developmental processes. Thus, the next-generation sequencing technologies offer the potential to accelerate epigenomic research substantially (Rival et al. 2010).

Using bisulfite sequencing technology, two complementary studies revealed extensive DNA methylation throughout the *Arabidopsis* genome (Lister et al. 2008; Cokus et al. 2008), and this result was consistent with the results obtained in previous studies that mapped methylation in the same plant using microarrays (Zhang et al. 2006; Zilberman and Henikoff 2007). The methylation was found to be especially high in heterochromatic regions, dispersed in euchromatic regions, and widespread in integral parts of genes. Nowadays, whole-genome bisulfite sequencing analysis of methylation patterns of duplicated and single-copy genes of soybean and common bean have provided insights into the functional consequences of polyploidy and epigenetic regulation in legume plant genomes (Kim et al. 2015).

List of high-throughput-sequencing types to generate epigenomic data

Method name	Definition	Details	References
ChIP-seq	Chromatin immunoprecipitation followed by next-generation DNA sequencing used to analyze DNA–protein interactions.	It is used for the determination of how proteins interact with specific regions of the genome. It can be used to detect the DNA sequences for transcription factors to interact, as well as the positions of histones with specific modification of their N-terminal tails. Antibodies that selectively bind methylated DNA may also be used to determine the position of methylated cytosines.	Pellegrini and Ferrari (2012)
DNase-seq	DNase I digestion of chromatin is combined with next-generation sequencing to identify regulatory regions of the genome, including enhancers and promoters.	DNase I digests every 10 bp of DNA around nucleosomes. Low concentrations of DNase I liberate accessible chromatin characterized as DNase I hypersensitive sites (DHSs). Chromatin conformation of active genes can be digested with DNase I. Bound regulatory factors within DHSs can inhibit DNase I cleavage and generate footprints [digital genomic footprinting (DGF) or DNase I footprinting]; this allows detecting TFs at nucleotide resolution in a qualitative and quantitative manner.	Tsompana and Buck (2014)
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements followed by sequencing to determine regulatory regions of the genome.	In this technique, chromatin is crosslinked with formaldehyde in vivo, sheared by sonication, and phenol-chloroform extracted. By much higher crosslinking efficiency of histones, nucleosome-depleted chromatins are released to the aqueous phase of the solution, compared to other regulatory factors. DNA in the aqueous phase is fluorescently labeled and used in DNA microarray.	Tsompana and Buck (2014)

ATAC-seq	Assay for transposase-accessible chromatin using sequencing that combines next-generation sequencing with in vitro transposition of sequencing adapters into native chromatin.	The technique is based on the ability of hyperactive Tn5 transposase to fragment DNA and integrate into active regulatory regions in vivo. The ATAC-seq protocol can identify accessible locations and nucleosome positioning simultaneously. However, its ability to map nucleosomes genome-wide is limited to regions in close proximity to accessible sites: 5×10^4 cells are sufficient to perform the technique, and the whole protocol lasts 3 h in total.	Tsompana and Buck (2014)
MNase-seq	Micrococcal nuclease (MNase) digestion of chromatin is followed by next-generation sequencing to identify loci of high nucleosome occupancy.	In MNase-seq experiment, mononucleosomes are extracted by MNase treatment of chromatin that has been crosslinked with formaldehyde. The nucleosomal population is subsequently submitted to single-end (identifies one end of template) or paired-end (identifies both ends of template). MNase-seq thus probes chromatin accessibility *indirectly* by unveiling the areas of the genome occupied by nucleosomes and other regulatory factors. MNase-seq is a superior method to localize many chromatin-bound proteins, and assessing TF occupancy in a range of cell types. However, it requires a large number of cells and careful enzymatic titrations for accurate and reproducible evaluation of differential substrates.	Tsompana and Buck (2014)

(continued)

(continued)

Method name	Definition	Details	References
ChIA-PET-seq	Chromatin interaction analysis by paired-end tag sequencing. A method that combines chromatin immunoprecipitation-based enrichment and chromatin proximity ligation with paired-end next-generation sequencing to determine genome-wide chromatin interactions.	ChIA-PET is better at its higher resolution associated with a protein of interest for functional study, and lays a solid foundation for studying long-range chromatin interactions in a three-dimensional (3D) manner, as well as provides a more reliable way to determine transcription factor (TF) binding sites and identify chromatin interactions.	Li et al. (2014)
Hi-C-seq	It is an extension of chromosome conformation capture that uses next-generation sequencing to observe long-range interaction frequencies between different regions of the genome.	Hi-C adapts the above approach to enable purification of ligation products followed by massively parallel sequencing. Hi-C allows unbiased identification of chromatin interactions across an entire genome. Hi-C can also be used to construct comprehensive, genome-wide interaction maps at finer scales by increasing the number of reads. This should enable the mapping of specific long-range interactions between enhancers, silencers, and insulators.	Lieberman-Aiden et al. (2009)
MethylC-seq	Involves shotgun sequencing of DNA treated with bisulfite, a chemical which converts unmethylated cytosines but not methylated cytosines to uracil.	This revealed extensive, previously undetected, DNA methylation, enabled both the context and level of methylation at each site to be assessed, and identified effects of the local sequence composition upon DNA methylation state.	Lister et al. (2008)

7 Conclusion

Epigenomics is the field that identifies variations at the level of RNA, protein–DNA interactions, chromatin accessibility, and modifications. In this way, it facilitates explaining phenotypic variations and expands our interpretation abilities. The epigenome is a very dynamic structure that can be shaped by various factors such as environmental perturbations and developmental signals. For this reason, epigenomic investigations must be individualized and epigenomes must be sequenced even for a single organism, implying epigenome sequencing is much more exhaustive than genome sequencing.

Studies on *Arabidopsis thaliana* have made important contributions to the explanation of how DNA methylation functions in plant genomes. In general, the plant genomes methylate cytosines in DNAs, but this is not the only path for epigenetic regulation. As more DNA methylomes are explored, additional mechanisms and novel pathways will be discovered. Thus, it is important to have genome-wide, high-resolution maps to understand the epigenetic regulation. Next-generation sequencing technologies allow us to determine the genome-wide epigenetic alterations such as methylated DNA and histone modifications. The data obtained from these technologies will shed more light on the mechanisms of action, epigenome characteristics, and the evolution of gene regulation. Recently, a method was developed for the methylome sequencing at the level of single embryonic mouse stem cells (Smallwood et al. 2014), which indicates that individualized epigenomic studies at the single-cell level are feasible. If the difficulties in DNA extraction from a single cell can be overcome, progression in plant genome studies will be easier. This improvement will provide the identification of regions and genes in the single-cell genome that are silenced by epigenetic mechanisms.

References

Ahmad A, Zhang Y, Cao XF (2010) Decoding the epigenetic language of plant development. Mol Plant 3(4):719–728

Ben Amor B, Wirth S, Merchan F, Laporte P, Aubenton-Carafa Y, Hirsch J, Maziel A, Malloy A, Lucas A, Deragon JM, Vaucheret H, Thermes C, Crespi M (2009) Novel long non-protein coding RNAs involved in *Arabidopsis* differentiation and stress responses. Genome Res 19(1): 57–69

Bestor TH (2000) The DNA methyltransferases of mammals. Hum Mol Genet 9(16):2395–2402

Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16(1):6–21

Callinan PA, Feinberg AP (2006) The emerging science of epigenomics. Hum Mol Genet 15(Spec No 1):R95–R101

Chapman EJ, Carrington JC (2007) Specialization and evolution of endogenous small RNA pathways. Nat Rev Genet 8:884–896

Chen M, Lv S, Meng Y (2010) Epigenetic performers in plants. Dev Growth Differ 52(6): 555–566

Chinnusamy V, Zhu JK (2009) Epigenetic regulation of stress responses in plants. Curr Opin Plant Biol 12(2):133–139

Clark SJ (2007) Action at a distance: epigenetic silencing of large chromosomal regions in carcinogenesis. Hum Mol Genet 16(Spec No 1): R88–R95

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature (Lond) 452(7184):215–219

Costa FF (2005) Non-coding RNAs: new players in eukaryotic biology. Gene (Amst) 357(2): 83–94

Dupont C, Armant DR, Brenner CA (2009) Epigenetics: definition, mechanisms and clinical perspective. Semin Reprod Med 27(5):351–357

EPIC, Epigenomics of Plants International Consortium (https://www.plant-epigenome.org/plants-leading-systems-epigenetics-and-epigenomics-research)

Feng F, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. Science 330(6004):622–627

Finnegan EJ, Dennis ES (1993) Isolation and identification by sequence homology of a putative cytosine methyltransferase from *Arabidopsis thaliana*. Nucleic Acids Res 21(10):2383–2388

Fuks F (2005) DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev 15(5):490–495

Groth A, Rocha W, Verreault A, Almouzni G (2007) Chromatin challenges during DNA replication and repair. Cell 128(4):721–733

Holoch D, Moazed D (2015) RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet 16(2):71–84

Jackson JP, Lindroth AM, Cao X, Jacobsen SE (2002) Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature (Lond) 416(6880):556–560

Jacob Y, Stroud H, LeBlanc C, Feng S, Zhou L, Caro E, Hassel C, Gutierrez C, Michaels SD, Jacobsen SE (2010) Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature (Lond) 466(7309):987–991

Jarillo JA, Pineiro M, Cubas P, Martinez-Zapater JM (2009) Chromatin remodelling in plant development. Int J Dev Biol 53(8-10):1581–1596

Kankel MW, Ramsey DE, Stokes TL, Flowers SK, Haag JR, Jeddeloh JA, Riddle NC, Verbsky ML, Richards EJ (2003) *Arabidopsis* MET1 cytosine methyltransferase mutants. Genetics 163(3):1109–1122

Kim JM, To TK, Ishida J, Morosawa T, Kawashima M, Matsui A, Toyoda T, Kimura H, Shinozaki K, Seki M (2008) Alterations of lysine modifications on the histone H3 N-tail under drought stress conditions in *Arabidopsis thaliana*. Plant Cell Physiol 49(10):1580–1588

Kim KD, El Baidouri M, Abernathy B, Iwata-Otsubo A, Chavarro C, Gonzales M, Libault M, Grimwood J, Jackson SA (2015) A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. Plant Physiol 168(4):1433–1447

Kouzarides T (2007) SnapShot: histone-modifying enzymes. Cell 131(4):822

Lane AK, Niederhuth CE, Ji L, Schmitz RJ (2014) pENCODE: a plant encyclopedia of DNA elements. Annu Rev Genet 48:49–70

Li X, Wang X, He K, Ma Y, Su N, He H, Stolc V, Tonqprasit W, Jin W, Jiang J, Terzaghi W, Li S, Denq XW (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell 20(2):259–276

Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, Kolchanov NA, Ruan Y (2014) Chromatin interaction analysis with paired-end tag (ChlA-PET) sequencing technology and application. BMC Genomics 12:S11

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Raqoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Snadstorm R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326:289–293

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(3): 523–536

Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF (2009) RNA regulation of epigenetic processes. Bioessays 31(1):51–59

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E et al (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1190–1195

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11(1):31–46

Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, Jacobsen SE, Ashikari M (2009) A metastable DWARF1 epigenetic mutant affecting plant stature in rice. Proc Natl Acad Sci USA 106(27):11218–11223

Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99(3):247–257

Pellegrini M, Ferrari R (2012) Epigenetic analysis: ChIP-chip and ChIP-seq. Methods Mol Biol 802:377–387

Rival A, Jaligot E, Beule T, Finnegan EJ (2008) Isolation and expression analysis of genes encoding MET, CMT, and DRM methyltransferases in oil palm (*Elaeis guineensis* Jacq.) in relation to the 'mantled' somaclonal variation. J Exp Bot 59(12):3271–3281

Rival A, Beule T, Bertossi FA, Tregear J, Jaligot E (2010) Plant epigenetics: from genomes to epigenomes. Not Bot Hort Agrobot Cluj 38(2):Special Issue 09–15

Sekhon RS, Chopra S (2009) Progressive loss of DNA methylation releases epigenetic gene silencing from a tandemly repeated maize Myb gene. Genetics 181(1):81–91

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 11(8):817–820

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 (Berl) 227(1):245–254

Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature (Lond) 403(6765):41–45

Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9(6):465–476

Tsompana M, Buck MJ (2014) Chromatin accessibility: a window into the genome. Epigenetics Chromatin 7(1):33

Vaillant I, Paszkowski J (2007) Role of histone and DNA methylation in gene regulation. Curr Opin Plant Biol 10(5):528–533

Vaughn MW, Tanurdzic M, Lippman Z, Jiang H, Carrasguillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, Colot V, Doerge RW, Martienssen RA (2007) Epigenetic natural variation in *Arabidopsis thaliana*. PLoS Biol 5(7), e174

Villar-Garea A, Imhof A (2006) The analysis of histone modifications. Biochim Biophys Acta 1764(12):1932–1939

Villar-Garea A, Israel L, Imhof A (2008) Analysis of histone modifications by mass spectrometry. Curr Protoc Protein Sci Chapter 14, Unit 14 10

Wang X, Elling AA, Li X, Li N, Peng Z, He G, Sun H, Qi Y, Liu XS, Deng XW (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell 21(4):1053–1069

Wassenegger M, Heimes S, Riedel L, Snager HL (1994) RNA-directed de novo methylation of genomic sequences in plants. Cell 76(3):567–576

Woo HR, Richards EJ (2008) Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. BMC Plant Biol 8:92

Yang PK, Kuroda MI (2007) Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell 128(4):777–786

Zaratiegui M, Irvine DV, Martienssen RA (2007) Noncoding RNAs and gene silencing. Cell 128(4):763–776

Zhang X (2008) The epigenetic landscape of plants. Science 320(5875):489–492

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126(6):1189–1201

Zhu JK (2008) Epigenome sequencing comes of age. Cell 133(3):395–397

Zhu J, Kapoor A, Sridhar VV, Agius F, Zhu JK (2007) The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in *Arabidopsis*. Curr Biol 17(1):54–59

Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. Development (Camb) 134(22):3959–3965

Next-Generation Sequencing Technologies and Plant Improvement

Fakiha Afzal, Alvina Gul, and Abdul Mujeeb Kazi

Contents

Abstract Hidden information lying underneath the genetic material is yet to be explored. Deciphering genetic information is the basic and primary step in bioscience research. For many years, capillary electrophoresis (CE)-based Sanger's method prevailed in the scientific world for elucidation of genetic information.

F. Afzal • A. Gul (✉)
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

A.M. Kazi
National Agricultural Research Centre (NARC), Islamabad, Pakistan

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_11

Due to lack in resolution, throughput, scalability, speed, and efficiency, it has now been replaced by spectacular next-generation sequencing (NGS) technologies since 5 years. Applications of NGS technologies in the field of plant biology is genome-wide scan for variants, rapid parallel sequencing, marker discovery, epigenetics, transcriptomics, de novo sequencing, resequencing, and high-resolution mapping in less time and money. This chapter briefly describes NGS technologies and utilization of these technologies in studying plant genome for its improvement and better development.

Keywords Capillary electrophoresis (CE) • Next-generation sequencing (NGS) • De novo sequencing • Epigenetics • Transcriptomics

1 Introduction

Deciphering DNA sequence is primary and the most important step in all fields of biological research. Scientific world has met charismatic next-generation sequencing (NGS) technologies for about 5–6 years which has led to the beginning of remarkable information regarding genetics, epigenetics, and transcriptomics which was previously not known (Austin et al. 2014). Thus, NGS opens a new door of research in the advancement of health, agriculture, and environment which is not only economical but also less time consuming and comparatively effortless. The amazing pace at which genome sequences are becoming available is largely due to the improvement in sequencing technologies both in terms of cost and speed (Bhavisha and Vrinda 2014). Modern sequencing technologies allow the sequencing of multiple cultivars of smaller crop genomes at a reasonable cost (Bolger et al. 2014).

DNA sequence determination methods were first established by two scientists, Fred Sanger and Alan R. Coulson, in 1977 (Sanger et al. 1977). These methods dramatically revolutionized biology in terms of determining sequences of genes and genomes. Then later on two other scientists, Maxam and Gilbert, proposed another method for DNA sequencing in the same year (Maxam and Gilbert 1977). Until 2008, these methods were successfully used to identify the sequence of bases in genome (Schuster 2008a, b). Later on automated capillary electrophoresis was developed for better interpretation of genes and genomes into DNA sequence (Margulies et al. 2006). Though Sanger's sequencing method is widely adapted all over the world for a longer period of time it has few and considerable limitations. This method lacks scalability and speed with low throughput and resolution.

Sequencing centers containing thousands of DNA sequencing instruments in order to automate and parallelize the process then started to develop by different enterprises in major countries of the world functioned by cohorts of workers. These never-ending efforts of scientists led to the successful completion of human genome project (HGP) (Schuster 2008a, b). Nevertheless the hunger for more economical

and fast-throughput sequencing was not ended which results to the innovation of "NGS" technologies revolutionizing the scientific world.

In plant genetics and breeding, NGS technologies open the new door of improving the quality of research. According to a prediction major parts of the world will face a terrible shortage of food. To meet the upcoming challenges regarding food security, NGS would firmly play a very positive role (Visendi et al. 2014). From the discovery of genes to the marker discovery, NGS technologies have proven to assist development of new better crop varieties. With great precision and accuracy NGS technologies led to the identification of novel genes and genetic variants for marker development (Chikara et al. 2014).

Major advantages of NGS over conventional sequencing technologies are the following:

1. The major advantage of NGS is that it runs in tremendously parallel manner in contrary to old sequencing methods which were limited to a single DNA strand or very few DNA strands. Being highly scalable, parallel sequencing, i.e., multiplexing in NGS, has led to the advancement of speedy sequencing (Bolger et al. 2014).
2. NGS involves highly sophisticated instruments which are capable of dealing with large genomic DNA fabricating hundreds of gigabase or even terabase of data in just single sequencing run (Schnable 2013).
3. Thus NGS is comparatively cheaper, less time consuming, and less laborious than conventional sequencing methods and gives more reproducible information (Bhavisha and Vrinda 2014).
4. NGS can be used for de novo sequencing in which there is no reference genome and base pair reads are either directly compared to the reads of other unknown sequences or assembled to reconstruct the sequence (Llaca 2012).
5. NGS is also used for resequencing in which reads are aligned to the reference genome majorly used in polymorphism discovery and transcription profiling (Hui 2012).
6. NGS technologies can deal with not only large and complicated genomes but also smaller genomes such as of viruses and bacteria. Not only this, it can be done for the genomes with no reference genome available (Matsuba et al. 2013).
7. NGS is comparatively flexible of each designed experiment.
8. Transforming present biology, NGS has become a universal biological tool due to its flexible and powerful nature.
9. NGS gives very high resolution giving precision up to a single nucleotide base (Mardis 2008).
10. NGS has widened the scope of metagenomics enabling the sequencing of antique and environmentally derived DNA samples (Mardis 2008).
11. NGS can make the dream of personal medicine come true due to reduced cost and precise knowledge of genetic information.
12. NGS has remarkable applications in plant breeding and genetics.
13. NGS has led to the discovery of high number of genomic variants and markers in crops lacking markers (Chen et al. 2014).

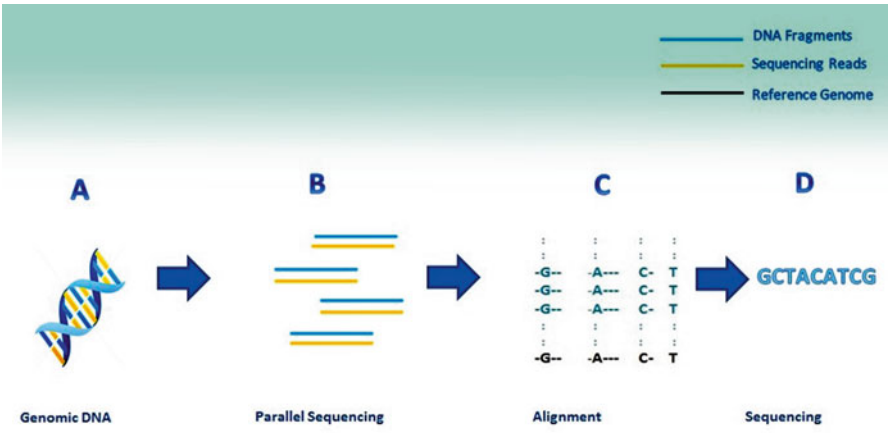


Fig. 1 Overview of next-generation sequencing technologies. (**a**) Genomic DNA is extracted. (**b**) Parallel sequencing of fragmentized genomic DNA. (**c**) Alignment of individual fragment to reference genome. (**d**) Whole-genome sequence is derived

14. Investigations in chromatin alterations and transcription factor (TF)-binding sites and DNA sequencing of target DNA sites bound to the DBPs (DNA-binding proteins) through their antibodies (Mikkelsen et al. 2007) can be done. Thus, NGS opens the new doors of research in the fields of transcriptomics and proteomics (Pettersson et al. 2009). The brief concept behind NGS is summarized in Fig. 1.

2 Principle of NGS with Respect to Their Respective Platforms

These novel techniques are highly throughput with great speed and accuracy as compared to Sanger's sequencing which takes not only several years but also numerous coinages. NGS has several platforms which work on their respective specific principle (Pettersson et al. 2009). The major advantage of these platforms is that there is no need of cloning and sequence is detected directly from amplified DNA fragment. As huge amount of data is generated in terms of several gigabase per run highly sophisticated software are installed to control this huge data (Ansorge 2009). Following are present sequencing generations with their respective platform.

2.1 454 Genome Sequencer FLX

This sequencer was introduced in 2005 by 454 Life Sciences and it works on the principle of parallelized pyrosequencing. It is also known as "highly parallel miniaturized pyrosequencing." Pyrosequencing is a process in which a molecule of

pyrophosphate is released by incorporation of each nucleotide into the growing strand of DNA by DNA polymerase. It causes initiation of reactions producing light signals by the production of light-emitting enzyme called luciferase which is detected by the detector. Amount of light production is directly proportional to the number of nucleotides being incorporated (Mardis 2008).

It was the first next-generation system introduced in the market. In recent days, the causative agent for epidemic of honey-bee disease was identified by this system which is its remarkable application (http://www.454.com; Schuster et al. 2008). Following are the major steps involved in 454 Genome Sequencer by parallelized pyrosequencing.

1. First of all population library is prepared and each fragment of the library is then linked to the specific adapters. These fragments with attached adapters are then mixed with agarose beads. These beads carry oligonucleotides attached to their surfaces which are specific to the adapters linked to each fragment. Thus, a single bead gets attached to the single fragment of the population via adapters (Mardis 2008).
2. Water-in-oil emulsion PCR is performed. It generated many (up to thousands) of amplified bead-fragment complexes (Lipshutz et al. 1995).
3. These single-clone amplified beads are put into the wells of special plate called "pico titer plate" (PTP). It is a thousand wells plate carrying a single bead in each well. Individual pyrosequencing reaction then takes place in each well providing fixed and immovable location for each reaction which can also be easily monitored (Patil et al. 2001).
4. Pyrosequencing then takes place by the enzyme attached to the beads in PTP which acts as a flow cell. All the four bases are added sequentially in a cyclic fashion (Mikkelsen et al. 2014).
5. After addition of each known nucleotide, light is generated by an enzyme cascade which is then detected via CCD camera.
6. The linked software then calibrates the emitted light at the incorporation of each base.

Total 500 Mbp of data is read through a PTP in single run. Figure 2 explains the major steps of this system. The major drawbacks of this system are;

(a) High cost
(b) Poor read accuracy in the stretches of DNA with identical bases (homopolar)

454 FLX Titanium is an upgraded version of this system with new PTP increasing the efficiency (Mikkelsen et al. 2014).

2.2 Applied Biosystems SOLiD™ Sequencer

Principle behind ABI SOLiD™ sequencer is ligation-based sequencing and is founded upon Polonator technology. It includes immobilization of DNA library to a solid support by utilizing emulsion PCR followed by cyclic ligation-based

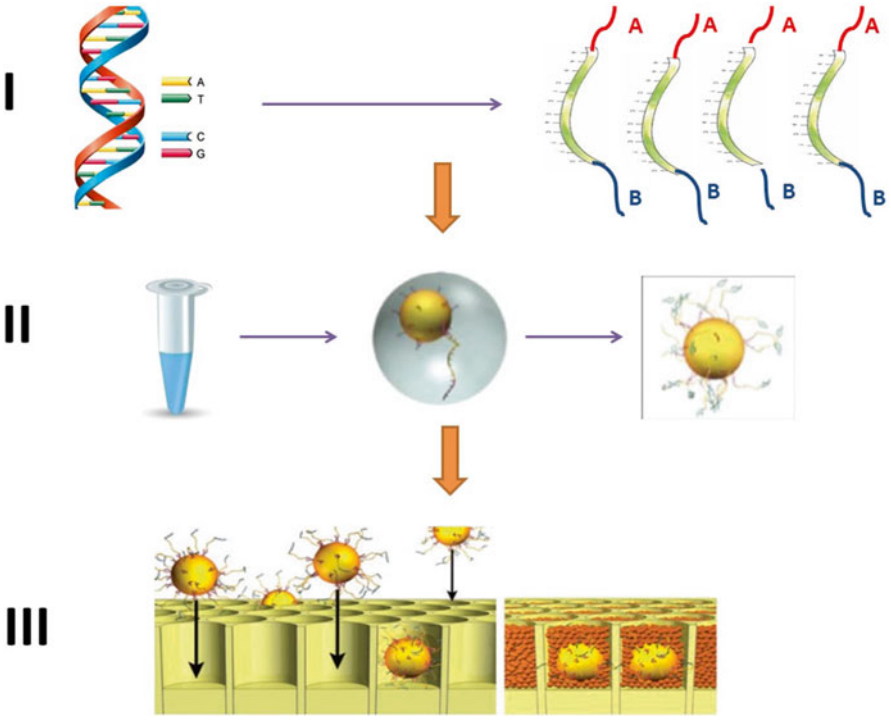


Fig. 2 Steps in 454 genome sequencer: (I) preparation of population library and linking of each fragment of the library to specific adapters; (II) a single bead gets attached to the single fragment of the population via adapters; (III) single-clone-amplified beads are put into the wells of special plate called "pico titer plate" (PTP) where pyrosequencing takes place

sequencing (Myllykangas et al. 2012). It was introduced to world in fall 2007. This system was updated in 2008 with improvements increasing the output from 3 to up to 10 Gb/run. This upgradation decreases the time span of total run to 1 week (Wang et al. 2014).

2.2.1 Overview of SOLiD™ System

Preparation of Library

Formulate one of the two SOLiD™ System sequencing-fragment libraries (Fig. 3). The selection of library is dependent upon the application one is performing and the information required from the experiments (Mardis 2009).

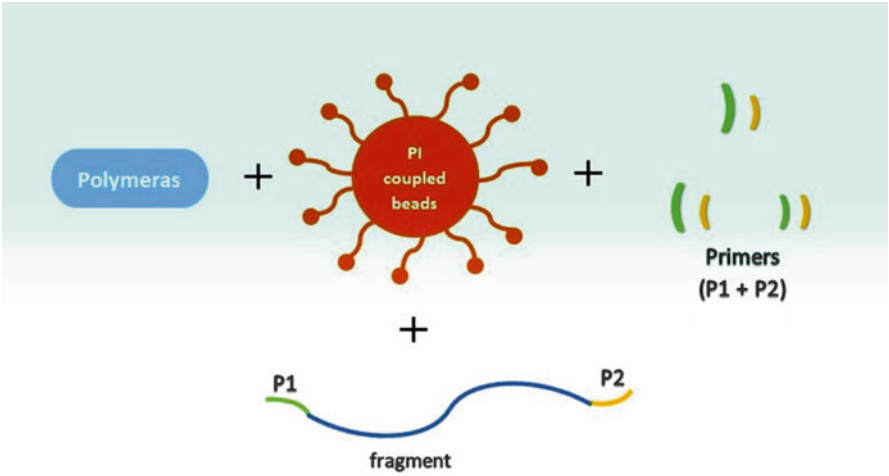


Fig. 3 Complex in 454 genome sequencer technology containing polymerase, P1-coupled beads, primer P1 and P2 attached which later get attached to the DNA fragment

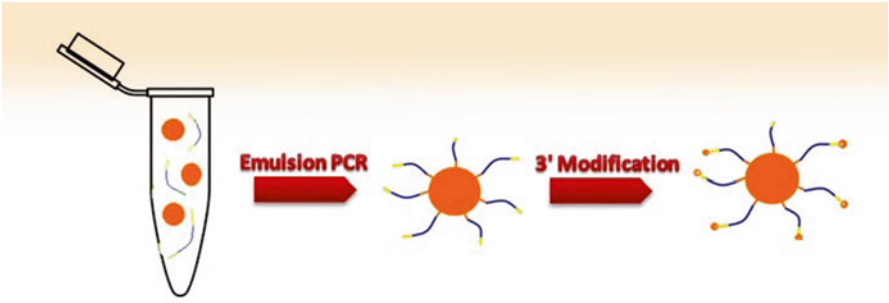


Fig. 4 Process of bead enrichment in which after emulsion PCR 3′ ends of enriched beads are modified so that they may form a covalent bond with the slide

Emulsion PCR

This step is also known as "bead enrichment." It is similar to other sequencing platforms using a beaded adapter to link with the template band and after emulsion PCR 3′ ends of enriched beads are modified so that they may form a covalent bond with the slide (Fig. 4) (Schuster et al. 2008).

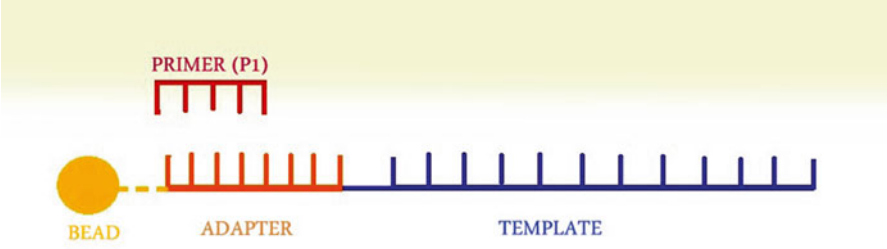


Fig. 5 Illustration of primer attachment to the fragment via adapter which is linked to the bead on the other side

Bead Positioning

Position 3′-modified beads onto a glass slide. The deposition chambers allow a slide to be segmented into one, four, or eight sections during the process of bead loading. A major benefit of the system is its ability to cope up with the increasing densities of beads per slide. Consequently, a higher level of throughput is achieved from the same system (Mardis 2008).

Sequencing by Ligation

Respective primers to the adapters ligate to amplify the bands with the four sets of fluorescent nucleotides. After every completed cycle, product is removed and cycle is repeated with the new complex (www3.appliedbiosystems.com/index.html).

Primer Reset

The primer pair is reset five times for each sequence tag. Through this reset process, nearly every base is cross-examined in two independent ligation reactions by two dissimilar primers. Figure 5 shows the primer template attached with bead.

Precise Call Chemistry

As sequencing is done with an additional primer using a multi-base encoding scheme, the Exact Call Chemistry Module achieves an accuracy of approximately equal to 99.99 %.

Multiplexing in SOLiDTM sequencer enables scientists to cope with variable number of samples, thus reducing the cost. Multiplexing also helps in data validation, comparisons, expression analysis, mutation studies, and epigenetics to be studied.

Fig. 6 Barcode is added to the 3′ end of the template band via modified P2 adapter

2.3 SOLiD™ System Barcodes

SOLiD™ System has introduced a unique barcode system to enhance multiplexing and to increase accuracy. Sixteen unique barcodes are designed having same melting temperatures and sequences unique to color space. A modified adapter named "P2" is used which is attached to 3′ end of the template. Barcode attaches to these adapters. Thus a complex is formed containing a beaded adapter linked at its one side to the template band and to its side to the barcode system.

The combination of two, eight-segment sequencing slides and the proficiency of sixteen distinct barcodes allows the investigation of up to 256 samples in a solo round. Using its corresponding identifier, the sequence data can then be backtracked to a particular sample through data analysis (Kircher and Kelso 2010). Complex of barcode to the adapter and template fragment is shown in Fig. 6.

Table 1 describes the fundamental principles of this system making it highly robust and accurate. It has been claimed that accuracy rate of greater than 99.94 % is being achieved by this system. It provisions exciting new-fangled applications such as

(a) Digital gene expression
(b) Large-scale resequencing
(c) Methylation studies
(d) Hypothesis-free ChIP

2.4 Illumina Genome Analyzer (Solexa)

The headquarters for Illumina Genome Analyzer is based in San Diego, California, having more than 1800 workers worldwide. Solexa was acquired by Illumina in 2006, commercializing the technology in 2007 (Metzker 2010). Illumina sequencing technology works on the principle of sequencing by Synthesis (SBS). Like other technologies it also provisions parallel sequencing. It is known for the detection of single-nucleotide change.

In SBS, all four dNTPs are bound to a fluorescent reversible terminator. After addition of each dNTP, a fluorescent reversible terminator is imaged. A new base is incorporated after cleavage in each new reaction. Sequencer is fed with new dNTPs

Table 1 Three fundamental principles of SOLiD™ system and attributes achieved by their combination

Three fundamental principles of SOLiD	Production by these fundamental principles	Advantages achieved
1. High-fidelity ligase enzymology	Enables massively parallel sequencing of clonally amplified DNA fragments linked to magnetic beads	Preventing dephasing
2. Primer reset functionality	After seven cycles of ligation, the old primer is stripped off from the template and new primer hybridizes enabling interrogation at n1 position	Reducing systemic noise and allowing longer read lengths
3. Two base encoding	Utilization of two dyes for 16 possible two-base combination enabling working with more complex templates with single-nucleotide polymorphism (SNPs), copy number variations (CNVs), insertions and deletions (INDELs)	Discriminating measurement errors versus true polymorphism

bound to a fluorescent reversible terminator in each cycle; therefore a natural competition develops which reduces incorporation bias resulting in accurate base-by-base sequencing (Fig. 7) (Milne et al. 2009).

Illumina genome analyzer covers a lot of applications including

(a) Whole genome re-sequencing
(b) Targeted re-sequencing
(c) ChIP sequencing
(d) MicroRNA discovery
(e) Gene expression

Further detail and advantages of applications using Illumina genome sequencing approach have been summarized in Table 2. Overall procedure of Illumina's Genome Analyzer utilizes the random attachment of fragmented DNA to the flow cells. Sequence of about 75 base pairs at the end of each fragment is read. High-density distribution of DNA fragment into flow cells is done. Flow cell is comprised of eight lanes and each lane carries volume of 300 Mbp. Each lane further contains three columns made up of 100 tiles. Due to this, generation of 100 million data is possible in single run. This method is also known as paired-end method. When both of the fragment ends are read then generation of data is doubled. Each tile during the whole process is imaged four times per cycle, i.e., one image/base. Therefore, 345,600 images for a 36-cycle run are taken, hence providing high accuracy (El-Metwally et al. 2014). Therefore Illumina genome analyzer can be used for single read as well as paired-end analysis. Following are the major steps of the analyzer.

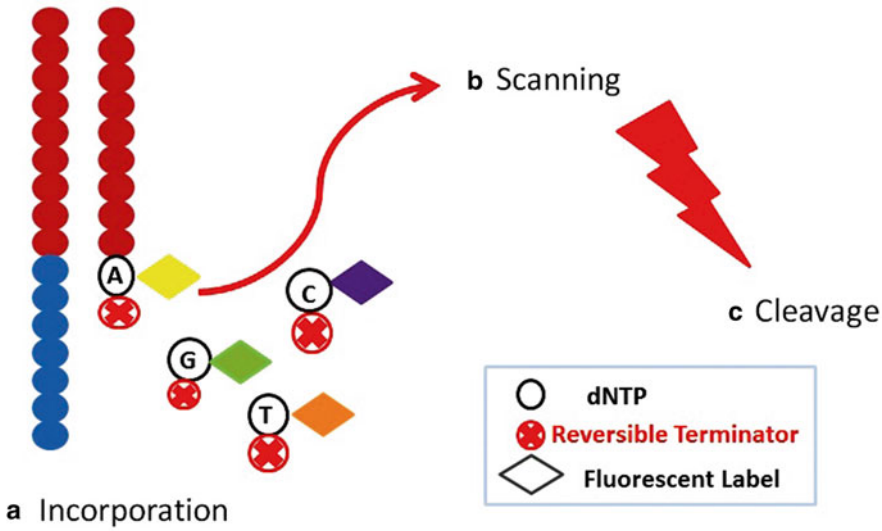


Fig. 7 All four dNTPs are bound to a fluorescent reversible terminator. (**a**) Incorporation of complementary fluorescently labelled nucleotide with reversible terminator. (**b**) A picture is taken every time a new base is added. (**c**) After addition of base the growing chain is cleaved and cycle is repeated

2.4.1 Library Preparation

Genomic DNA library is prepared. First DNA is randomly fragmentized to approximately 200 base pair size and ligated to adapters of two different types on both ends of fragments. Sample preparation normally takes 1–5 days.

2.4.2 Sample Addition to Flow Cell

DNA fragments ligated to adapters are attached to the solid surface, i.e., flow cells.

2.4.3 Solid-Phase Bridge Amplification

DNA fragments are amplified by bridge amplification producing more than 70 million of dense clusters of double-stranded DNA from each channel of the flow cell. This step usually takes 1–2 days.

Table 2 Powerful applications of Illumina sequencer (Solexa) and their advantages

Powerful applications of Illumina sequencer (Solexa)		Advantages
1	De novo sequencing	Paired reads: high library diversity by generating long scaffolds and highly accurate contigs with the help of multiple insert lengths Read lengths: allows de novo assembly by using paired-end reads in 2 × 100 bases Raw read precision: generated long error-free contigs Assembly tools: velvet, SOAPdenovo, forge
2	Resequencing	Single-nucleotide polymorphism (SNP): maximum accuracy with the help of direct base interrogation Insertions and deletions (INDELS): accurately detect both short insert paired end and long insert mate pairs Copy number variants (CNVs): provide comprehensive, uniform coverage by enabling genome characterization of CNVs Structural variation: enables to perform local de novo assembly
3	Transcript profiling	High throughput: profiling transcripts in single day Supreme sensitivity: quantification and identification of both rare and common transcript Protocol flexibility
4	Bisulfite sequencing	Precision of mapping: utilizing accurate 100 base paired-end reads to generate correct maps Library groundwork: easy preparation of highly diverse libraries necessary for comprehensive epigenome Universal protocol: utilization of standard reversible terminator technology Cost-effective coverage: evenly distributed long reads maximize the yield of usable data
5	ChIP sequencing	Low sample contribution: as less as 10 ng of sample can generate precise and accurate DNA-protein interaction maps Accurate mapping-rare binding events are identified Extraordinary sensitivity
6	Transcript discovery	Wide-range characterization: in a single experiment enables characterization of splice variants, coding SNPs, and comparative expression of alleles Low sample contribution: only 1–10 µg of sample can create paired-end library Unbiased coverage: reproducible and unbiased coverage Strand specificity: discovery and profiling of overlapping transcripts by differentiating between plus and minus strands

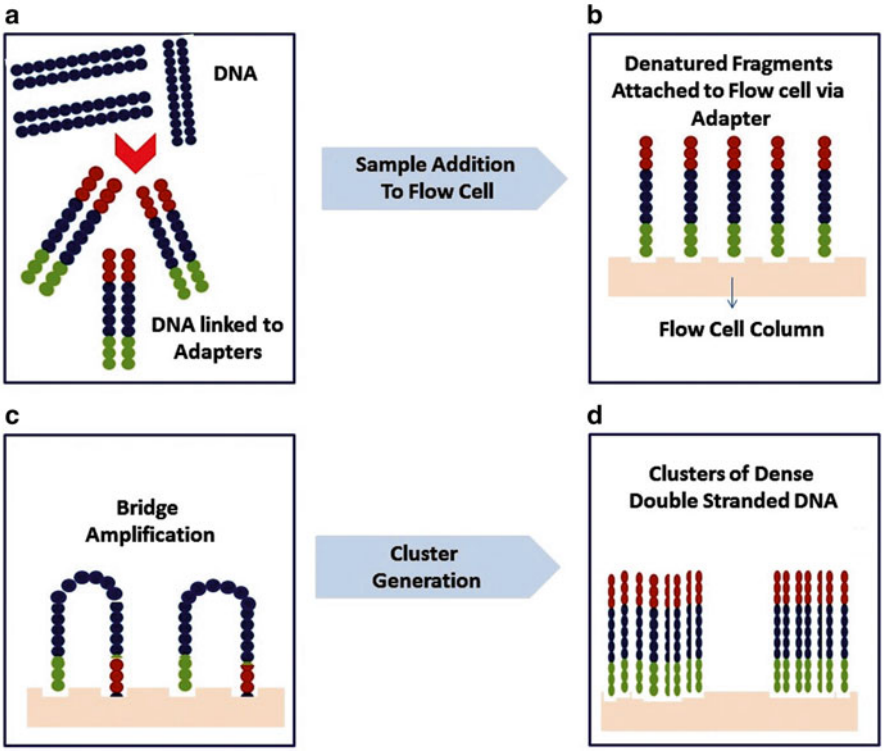


Fig. 8 The Illumina Solexa SBS approach. (**a**) Preparation of genomic DNA library by attaching double-stranded DNA molecules to adapters. (**b**) Incorporation of DNA fragments via adapters to flow cell where (**c**) bridge amplification is done by making them into single stranded. (**d**) Final amplification produces millions of dense clusters of double-stranded DNA fragments. Imaging is done in the lanes of columns inside the flow cells after addition of each single nucleotide. Once imaging is done the fluorescent molecule is cleaved. This cycle is repeated again and again by continuous addition of reagents to the sequencer

2.4.4 Sequencing, Data Processing, and Analysis

After addition of each fluorescently labelled reversible terminator, signal is detected and imaged. Sequencing step takes 1–2 days while it takes days to months to analyze and process the data generated. Figure 8 explains the major steps of Illumina technology (Robertson et al. 2007; Fields 2007; Johnson et al. 2007).

There exist some limitations of Solexa technology such as increase in error rate after 32 bp. Another limitation is the need of high coverage to detect polymorphism as missed SNPs are likely due to low coverage. Solexa can't sequence through repeats and de novo assembly is not yet possible with short reads. Recently analysis of gene expression patterns in floral bud emergence is done by construction and de novo characterization of transcriptome (Huang et al. 2014).

Table 3 Comparison between all three next-generation sequencing technologies

Comparison between next-generation sequencing technologies				
Technology name	Reads/run	Average read length (base pair)	Base pair per run	Data output
454 Genome sequencer FLX	400,000	250–310	70 Millions	20 GB
Applied Biosystems SOLiDTM sequencer	40 Millions	36	1 Billion	1.5 TB
Illumina genome analyzer (Solexa)	88–132 Millions	35	1 Billion	1.5–3.0 TB

Hence technology is improving day by day and scientists eager to find much more will never end. Table 3 compares the above three mentioned technologies in respect to reads/run, average read length, base pair/run, and data output.

3 Other Next-Generation Sequencing Platforms

There are various other NGS platforms available in the scientific market but they have limited uses in plant biology. The basic concept is the same but each of them uses different methods for template preparation, immobilization of template, and detection of the nucleotide sequence (Egan et al. 2012). Other NGS platforms are HeliScope (marketed by Helicos), Max-Seq Genome sequencer (marketed by Azco-Biotech), Polonator, and Pacific Bioscience PacBio (Llaca 2012).

4 Implications of Next-Generation Sequencing Technology for Crop Genetics and Breeding

In this modern era, genomics will soon turn into the integral unit of every life sciences branch. It is only possible due to easy and accessible genome sequencing. Regarding plant biology, NGS has tremendous and commendable applications leading scientists from improving crop quality to developing new better crop varieties by glancing deep into the genome of plants (Egan et al. 2012). Never-ending efforts are being done by scientists for the last 15–20 years in order to sequence genomes of different plants but in present years this dream would be fulfilled by NGS technologies (Feuillet et al. 2011). Reference genome should be available in order to resequence the genome or for those plants having complex genome with no reference genome de novo sequencing is done (Jackson et al. 2011). Due to the presence of large proportion of repetitive DNA sequences and polyploidization events, it is difficult to sequence plant genome (Schnable et al. 2009). One of the important crops, wheat, has genome which is six times more complex than *Homo sapiens* and still needs to be explored (Llaca et al. 2012). Arabidopsis and rice were the first

plants whose genomes were sequenced by shotgun sequencing based on Sanger's method. But it took several years and high cost to complete this project. It has been estimated that it took 70 million US$ to sequence *Arabidopsis* genome (Feuillet et al. 2011). Therefore need of high-density maps in less time and cost became the primary need of scientists. Here are some implications of NGS technologies on crop breeding and genetics.

4.1 Development of Molecular Markers for Construction of High-Resolution Genetic Maps

Variable regions in genome can act as marker for any desirable trait. NGS technologies have led to the discovery of thousands of high-coverage markers with fewer gaps between them. NGS technologies also allow to improve the existing maps both in polyploid and diploid organisms through resequencing studies, thus increasing marker density and integrating new markers such as SNPs into genomes (Oliver et al. 2011). SNPs are single-nucleotide polymorphism or point mutations naturally present in the genome. Due to their abundance in genome, they are the most favorable tool to be used as markers. Marker development has led to marker-assisted selection (MAS), i.e., linking of marker to the desired phenotypic trait (Egan et al. 2012). Previously, expressed sequence tags (ESTs) were used to identify SNPs in plant genome but due to low coverage and high cost this method is not useful now. SNP discovery by NGS is quiet challenging for the genome of plants with no reference genome available such as wheat (Azam et al. 2012). In a study by Azam et al. (2012), highly accurate and precise SNPs were predicted by using four tools for short read alignment of chick pea (*Cicer arietinum* L.) which has no reference genome. They named it "coverage-based consensus calling" (CbCC) for SNP discovery. In another study linkage map was generated by using SNP marker-associated ESTs for field pea (*Pisum sativum* L.) providing evenly dispersed genome-wide coverage (Leonforte et al. 2013).

SSRs (simple sequence repeats or microsatellite) are tandem repeats of DNA sequence usually of 1–6 base pairs. Before the advent of NGS, development of SSRs was very laborious. It involved complex library preparation, cloning, and then sequencing. A large number of SSRs have been developed by NGS using Illumina and 454 pyrosequencing (Zalapa et al. 2012). Further SSR identification can lead to the development of cultivar fingerprinting (Ahmed et al. 2013), quantitative trait loci (QTL) mapping (Jeennor and Volkaert 2014), marker-assisted selection (MAS) (Li et al. 2014), evolutionary studies (Tabbasam et al. 2014), development of linkage map (Cosson et al. 2014), and gene flow.

The release of a reference genome for *Arabidopsis thaliana* in 2000 has been a huge bonus for the study of plant genetics. Large-scale efforts to characterize natural genomic variation in *A. thaliana* have revealed remarkable intraspecific variation in this species, ranging from single-nucleotide differences to large structural rearrangements (Hollister 2014).

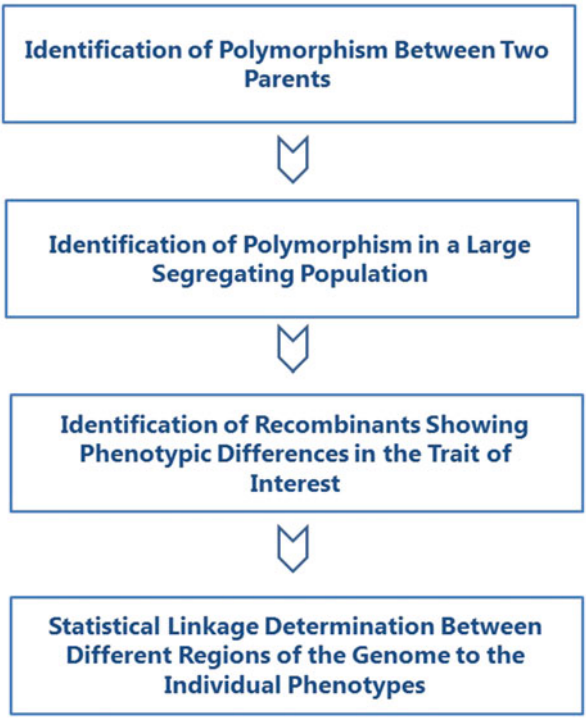


Fig. 9 Steps in linkage mapping for QTLs

4.2 Genome-Wide Association Studies and QTL Mapping

In plants, traits like fruit weight, yield of crop, sugar content, protein content, proline content, resistance to biotic stresses, and flowering time are quantitative and agronomically very important (Paran and Zamir 2003). For more than 100 years, scientists are studying QTLs and are now trying to map the underlying QTLs of these quantitative traits. Before 2005, many QTL maps were developed but major problem was low resolution of these maps (Salvi and Tuberosa 2005). Previously before NGS, strategies of linkage mapping were used to identify and clone most of the QTLs. Steps of linkage mapping are described in Fig. 9. Such linkage maps have low resolution due to less number of recombination events and number of generations (Liu et al. 2002). Later QTL mapping is done by genome-wide association studies (GWAS) after the advent of NGS technologies (Schneeberger and Weigel 2011). The major differences between GWAS and linkage mapping are as follows:

1. In GWAS, a large number of recombination events are generated within population by exploitation of natural diversity in contrast to linkage mapping (Nordborg and Weigel 2008).
2. As a result a high-resolution map is generated.

3. Whole-genome coverage is achieved in GWAS by millions of genetic markers unlike linkage mapping (Zhu et al. 2008).
4. Recombinant inbred lines (RILs) are generated in order to minimize genome complexity.

4.3 Linkage Disequilibrium (LD) Mapping or Association Mapping (AM)

Identification of inherited markers associated to the genetic factor affecting QTL is the main objective of mapping. For QTL mapping, F1 and F2 generations derived from two parents to get meiotic recombination are done (Holland 2007). Still sufficient number of meiosis is not achieved, thus hampering fine mapping. Therefore another better approach, association mapping, is introduced in unrelated genotypes in which association was created in distant past. Although it is quiet similar to QTL mapping but advantages of association mapping over QTL mapping are (1) high mapping resolution as compared to typical bi-parental cross approaches, (2) less time consuming, and (3) large number of alleles (Zhu et al. 2008). AM exploits linkage disequilibrium to discover trait marker relationship by taking a benefit of previous recombination event. Brief methodology of AM is described in Fig. 10.

In candidate gene association mapping, with the help of high-throughput sequencing technologies especially Solexa, it is now relatively easy and quick to identify large number of genomic variants, especially SNPs generated by combining PCR

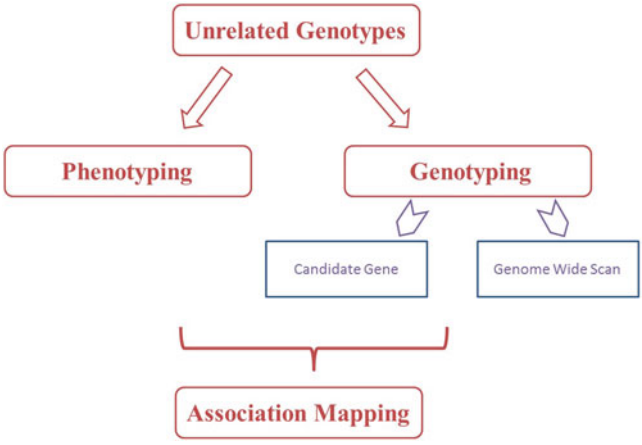


Fig. 10 Association mapping is done by relating genotypic data with phenotypic data of selected unrelated genotypes. Association mapping can be done by two different approaches, i.e., genome-wide association mapping and candidate-gene association mapping depending upon the designed experiment

amplicons of candidate gene from hundreds of natural population genotypes (Varshney et al. 2009). In whole genome-wide AM, with the help of NGS technologies large sets of genomic markers are identified by scanning whole genome which was impossible before (Nordborg and Weigel 2008). In barley, genome-wide association scans (GWAS) can be done to explore the genetic traits with respect to their alleles (Waugh et al. 2014). Abiotic and biotic stress tolerance genes in plants can be linked to the marker for improvement of crop quality and selection of tolerant crop varieties (Jenks and Hasegawa 2014).

In identification of candidate genes for complex quantitative traits in chickpea by association mapping, large-scale validation and high-throughput genotyping of numerous informative genic microsatellite markers are required. As the screening and genotyping of such informative markers in individual genotypes/whole-association panels for trait association mapping is very expensive alternative time-saving and cost-effective pool-based trait association mapping approach by combining pooled DNA analysis (with 616 genic microsatellite markers) and individual genotype (large structured association panel) genotyping is recently developed by NGS technologies (Kujur et al. 2014).

4.4 Resequencing Plant Genome

Resequencing is a powerful application of NGS for those plant species which have well-characterized genome. In this, by using NGS technologies genotypes are sequenced for marker discovery. Reduced representative genome or whole-genomic DNA of various genotypes is used as sequencing data. cDNA population can also be used. Afterwards alignment is done by using different bioinformatics tools of reference genome to the sequenced data. Comparison is done to the reference genome to generate large number of variants.

For example in a study on maize (*Zea mays*) Roche/454 sequencing generated 261 000 ESTs from shoot apical meristem, of which 30 % were novel; ~400 unique ESTs were also identified (Emrich et al. 2007).

4.5 De Novo Sequencing

Assembly and sequencing of plant genome in the absence of reference genome are known as de novo sequencing. De novo sequencing as compared to resequencing is very tedious because of absence of reference genome. Plant species that have complex genome containing large number of recombination events and polyploidy are hard to sequence. With the help of NGS technologies it is now possible to decipher DNA sequence of such species. De novo sequencing is done by generating sequence data of many different genotypes. Alignment of short reads generated by sequence

data is then done by sophisticated bioinformatics tools (Varshney et al. 2009). Alignment is done with the species of plant closely related to the sequence data providing sequence variants. These sequence variants such as SNP can lead to marker development in crops with no markers (Schuster 2008a, b).

4.6 Transcriptome Sequencing of Plants

Functional genomics is developing day by day. Though transcriptome de novo sequencing is very demanding and troublesome, scientists are trying their level best to explore plant transcriptome. Annotating plant transcriptome is highly useful in plant genetics and crop improvement providing the knowledge of genes and rare transcripts (Pérez-de-Castro et al. 2012). Recently with the help of NGS technologies, US Department of Energy Joint Genome Institute (JGI) has successfully conducted an experiment on de novo sequencing of maize by generating a large number of short reads of RNA-sequence through alternative splicing (Martin et al. 2014). In another study whole transcriptome of *Nicotiana benthamiana* (sugar beet) is sequenced by Illumina after an infection of *beet necrotic yellow vein virus* causing Rhizomania which is a major hindrance in crop production of sugar beet worldwide. With the help of this study scientists identified the gene involved in resistance to this disease (Fan et al. 2014). Many protocols to study plant transcriptome have been optimized and a new method of RNA extraction to study the expression of gene for the use in NGS technologies has recently been proposed by Yockteng et al. 2013.

4.7 NGS and Plant Virology

Plant virology is one of the emerging fields of science as a lot of crops are deteriorated due to infections of deadly viruses worldwide. Sequencing of double-stranded DNA, total RNA, double-stranded RNA, and small interfering (si) RNA of many viruses has been done with the help of NGS technologies (Barba et al. 2014). In a study using Roche 454 sequencing platform total RNA of *Australian grapevine viroid* was successfully done (Al Rwahnih et al. 2009). Similarly in another study by Solexa and Illumina sequencing platforms, a novel virus similar to luteovirus was discovered and named Citrus vein enation virus (Vives et al. 2013). Cassava mosaic disease, caused by cassava begomoviruses, is the most serious disease for cassava in Africa. The pathogenesis of this disease is poorly understood. With the help of high-throughput digital gene expression profiling based on the Illumina Solexa sequencing technology the global transcriptional response of cassava to African *cassava mosaic* virus infection was investigated and it was found that 3210 genes were differentially expressed in virus-infected cassava leaves (Liu et al. 2014).

4.8 Plant Epigenetics

Epigenetics is a subbranch of genetics in which alteration of gene expression without DNA sequence change is studied. Epigenetic involves DNA methylation and histone modifications. Today epigenetics has gained a lot of importance because discovery of many epigenetics mechanisms that affects activity of gene is made possible. Genome of an organism remains unchangeable throughout the life but epigenome is not static. It changes according to environmental or other factors (Sheth and Thaker 2014). Epigenetic changes are important in cell differentiation throughout the development of plant and also in maintaining gene expression profiles. Previously ChIP-chip was done which involves chromatin immunoprecipitation (ChIP) to determine regulation of gene expression followed by microarray hybridization (Mardis 2007). NGS ChIP sequencing technologies have replaced ChIP-chip as it is less laborious, fast, reliable, accurate, and comparatively cost effective. But very less research is done in plant epigenetics by NGS technologies. In a study comparison was done between wild-type *Arabidopsis* to the mutant one by sequencing cytosine methylome by using Solexa technology (Barski et al. 2007). This study also includes direct sequencing of mRNA-seq and smRNAseq and revealed the complete loss of CpG DNA in mutant type (Lister et al. 2008). NGS is being adopted with very high pace in epigenetics by scientists due to its remarkable applications and up till now many protocols have been optimized to study epigenome of plants by NGS technologies (Meaburn and Schulz 2012).

Other applications of NGS implicating plant breeding and genetics are genotyping by sequencing (GBS), gene mining, metagenomics, restriction site-associated DNA sequencing (RAD), evolutionary studies, and phylogenetic.

5 Conclusion and Future Prospects

After the advent of NGS technologies biological sciences have been revolutionized by revelation of fabulous information from genomic material which can be DNA, RNA, siRNA, dsRNA, smRNA, or transcriptome. In terms of plant genetics and breeding, NGS has played significant countless roles. Wheat which is considered to be one of the world's leading cereal grains in terms of the area harvested, production, and nutrition has been studied by NGS technologies for better grain yield. Due to its complex and large genome it was previously impossible to sequence whole genome but due to NGS technologies it would soon be sequenced by reducing the complexity of its genome. Barley genome sequencing is lagging behind the status achieved for many other crop genomes although barley is ranking worldwide as fifth most important crop species. By the introduction of NGS technology this situation has changed and fascinating new possibilities opened up for in-depth barley genome analysis and whole-genome sequencing (Visendi et al. 2014). One of the applications of NGS is marker discovery which would be helpful in assigning

markers to those species which are marker deficient; thus MAS would be possible for them. In future, NGS would replace microarray experiment as NGS is less time consuming, more accurate, and comparatively cheap. For crops like wheat and pigeonpea, NGS technologies are currently being used for de novo sequencing. It would not be an exaggeration to say that in future it would be a matter of few hundred dollars to sequence genomes of every plant species rather than model plants. Regarding resequencing of genomes high-resolution maps with great marker coverage are built. Phylogenetic advancements and evolutionary studies of plants by NGS technologies can be helpful to link predecessors and successors in terms of gene inheritance. It would help to identify ancestors of different crops. Improving crop quality to get higher yield is the primary goal of scientists. With the help of NGS technologies, it is easy to identify biotic and abiotic stress-tolerant genes for producing transgenic plants with high tolerable features. Conventional Sanger's method utilizes more than 3000$/Mbp for read length of 1000 bps generating 0.001 Mbp per run while 454 Roche utilizes 66$/Mbp for read length of 450 bps generating 450 Mbp per run. Similarly Illumina utilizes 0.07$/Mbp for read length of 100 bps generating 270,000 Mbp per run and SOLiD utilizes 0.07$/Mbp for read length of 50 bps generating 270,000 Mbp per run. Therefore, in comparison to Sanger's method, NGS has proven to be much better in terms of cost, time, and accuracy.

References

Ahmed MM, Guo H, Huang C, Zhang X, Lin Z (2013) Selection of core SSR markers for fingerprinting upland cotton cultivars and hybrids. Aust J Crop Sci 7(12):1912–1920

Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnol 25(4):195–203

Austin RS, Chatfield SP, Desveaux D, Guttman DS (2014) Next-generation mapping of genetic mutations using bulk population sequencing. Methods Mol Biol 1062:301–315

Azam S, Thakur V, Uperao PR, Shah T, Balaji J, Amindala B, Farmer AD et al (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results for the identification of SNPs in chickpea (Cicer arietinum; Fabaceae), a crop species without a reference genome. Am J Bot 99:186–192

Barba M, Czosnek H, Hadidi A (2014) Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6:106–136

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129(4):823–837

Bhavisha PS, Vrinda ST (2014) Plant systems biology: insights, advances and challenges. Planta. doi:10.1007/s00425-014-2059-5

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF (2014) Plant genome sequencing-applications for crop improvement. Curr Opin Biotech 26:31–37

Chen W, Yao J, Chu L, Li Y, Guo X, Zhang Y (2014) The development of specific SNP markers for chromosome 14 in cotton using next-generation sequencing. Plant Breeding 133(2): 256–261

Chikara SK, Pandey M, Pandey S, Vaidya K, Chaudhary S (2014) Next generation sequencing: a revolutionary tool for plant variety improvement. AJSIH 137–154

Cosson P, Decroocq V, Revers F (2014) Development and characterization of 96 microsatellite markers suitable for QTL mapping and accession control in an *Arabidopsis* core collection. Plant Methods 10:2

Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99(2):175–185
El-Metwally S, Ouda OM, Helmy M (2014) Next-generation sequence assemblers. In: El-Metwally S, Ouda OM, Helmy M (eds) Next generation sequencing technologies and challenges in sequence assembly, vol 7. Springer, New York, pp 103–116
Emrich SJ et al (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17:69–73
Fan H, Sun H, Wang Y, Zhang Y, Wang X, Li D, Yu J, Han C (2014) Deep sequencing-based transcriptome profiling reveals comprehensive insights into the responses of *Nicotiana benthamiana* to beet necrotic yellow vein virus infections containing or lacking RNA4. PLoS One 9(1):1–12
Feuillet C et al (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16(2): 77–88
Fields S (2007) Molecular biology. Site-seeing by sequencing. Science 316(5830):1441–1442
Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156–161
Hollister JD (2014) Genomic variation in *Arabidopsis*: tools and insights from next-generation sequencing. Chromosome Res. doi:10.1007/s10577-014-9420-1
Hui P (2012) Next generation sequencing: chemistry, technology and applications. Top Curr Chem. doi:10.1007/128_2012_329
Jackson SA et al (2011) Sequencing crop genomes: approaches and applications. New Phytol 191(4):915–925
Jeennor S, Volkaert H (2014) Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (Elaeis guineensis Jacq.). Tree Genet Genomes 10(1):1–14
Jenks MA, Hasegawa PM (2014) QTL and association mapping for plant abiotic stress tolerance trait characterization and introgression for crop improvement. In: Fleury D, Langridge P (eds) Plant abiotic stress. Wiley, New York
Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497–1502
Kircher M, Kelso J (2010) High-throughput DNA sequencing—concepts and limitations. Bio Essays 32(6):524–536
Kujur A, Bajaj D, Saxena MS, Tripathi S, Upadhyaya HD, Gowda CLL, Singh S, Tyagi AK, Jain M, Parida SK (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 Breeding. doi:10.1007/s11032-014-0033-3
Leonforte A, Sudheesh S, Cogan NO, Salisbury PA, Nicolas ME, Materne M, Forster JW, Kaur S (2013) SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (*Pisum sativum* L.). BMC Plant Biol 13:161
Li MY, Wang F, Jiang Q, Ma J, Xiong AS (2014) Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens L.*) by deep transcriptome sequencing. Horticulture Res doi:10.1038/hortres.2014.10
Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP (1995) Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 19:442–447
Lister R et al (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:1–14
Liu J et al (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc Natl Acad Sci U S A 99(20):13302–13306
Liu J, Yang J, Bi H, Zhang P (2014) Why mosaic? Gene expression profiling of African cassava mosaic virus-infected cassava reveals the effect of chlorophyll degradation on symptom development. J Integr Plant Biol 56(2):122–132
Llaca V (2012) Sequencing technologies and their use in plant biotechnology and breeding. In: Munshi A (ed) DNA sequencing—methods and applications doi: 10.5772/37918
Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4:613–614

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
Mardis ER (2009) New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med 1(4):40
Margulies M et al (2006) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380
Martin J, Gross S, Schnable J, Choi C, Wang M, Singh K, Lindquist E, Chen F, Wei C, Wang Z (2014) Deep sequencing of a plant transcriptome. U.S. Department of Energy Joint Genome Institute (JGI), Walnut Creek, CA
Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schafer P, Kudrna D, Wing RA, Bolger AM et al (2013) Evolution of a complex locus for terpene biosynthesis in solanum. Plant Cell 25:2022–2036
Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74(2):560–564
Meaburn E, Schulz R (2012) Next generation sequencing in epigenetics: Insights and challenges. Sem Cell Dev Biol 23(2):192–199
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46
Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, Donovan AO, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560
Mikkelsen M, Hansen RF, Hansen AJ, Morling N (2014) Massively parallel pyrosequencing 454 methodology of the mitochondrial genome in forensic genetics. Forensic Sci Int Genet. doi:10.1016/j.fsigen.2014.03.014
Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2009) Tablet—next generation sequence assembly visualization. Bioinformatics 26(3):401–402
Myllykangas S, Buenrostro J, Hanlee PJ (2012) Overview of sequencing technology platforms. In: Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM Springer (ed) Bioinformatics for high throughput sequencing, Springer NY, pp 11–25
Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature 456:720–723
Oliver RE, Lazo GR, Lutz JD, Rubenfield MJ, Tinker NA, Anderson JM, Morehead NHW, Adhikary D, Jellen EN, Maughan PJ, Guedira GLB, Chao S, Beattie AD, Carson ML, Rines HW, Obert DE, Bonman JM, Jackson EW (2011) Model SNP development for complex genomes based onhexaploid oat using high-throughput 454 sequencing technology. BMC Genomics 12:77–92
Paran I, Zamir D (2003) Quantitative traits in plants: beyond QTL. Trends Genet 19(6):303–306
Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723
Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual L, Blanca JM, Díez MJ, Prohens J, Picó B (2012) Application of genomic tools in plant breeding. Curr Genomics 13:179–195
Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–111
Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4(8):651–657
Rwahnih A, Daubert M, Golino S, Rowhani D (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387:395–401
Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci 10(6):297–304

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467

Schnable PS (2013) Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 64:71–88

Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115

Schneeberger K, Weigel D (2011) Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci 16(5):282–288

Schuster SC (2008a) Next-generation sequencing transforms today's biology. Nat Methods 5(1):16–18

Schuster SC (2008b) Next-generation sequencing transforms today's biology. Nature Publishing Group doi: 10.1038/NMETH1156

Schuster SC et al (2008) Method of the year, next-generation DNA sequencing. Functional genomics and medical applications. Nat Methods 5:11–21

Sheth BP, Thaker VS (2014) Plant systems biology: insights, advances and challenges. Planta. doi:10.1007/s00425-014-2059-5

Tabbasam N, Zafar Y, Mehboob-ur-Rahman (2014) Pros and cons of using genomic SSRs and EST-SSRs for resolving phylogeny of the genus Gossypium. Plant Syst Evol 300:559–575

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27(9):522–530

Visendi P, Batley J, Edwards D (2014) Genomics of plant genetic resources. In: Tuberosa R, Graner A, Frison E (eds) Next generation sequencing and germplasm resources. Springer, New York, pp 369–390

Vives MC, Velazquez K, Pina JA, Moreno P, Guerri J, Navarro L (2013) Identification of a new enamovirus associated with citrus vein enation disease by deep sequencing of small RNAs. Phytopathology 103:1077–1086

Wang Y, Haung H, Ma Y, Fu J, Wang L, Dai S (2014) Construction and de novo characterization of a transcriptome of: analysis of gene expression patterns in floral bud emergence. Plant Cell Tiss Org Cult 116(3):2970309

Waugh R, Flavell AJ, Russell J, Thomas WB, Ramsay L, Comadran J (2014) Exploiting Barley genetic resources for genome wide association scans (GWAS). In: Tuberosa R, Graner A, Frison E (ed) Genomics of plant genetic resources, Springer NY, pp 237–254

Yockteng AR, Almeida AMR, Yee S, Andre T, Hill C, Specht CD (2013) A method for extracting high-quality RNA from diverse plants for next-generation sequencing and gene expression. Appl Plant Sci 1(12):1–6

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, Mccown B et al (2012) Using next-generation sequencing approaches for the isolation of simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208

Zhu C et al (2008) Status and prospects of association mapping in plants. Plant Gen J 1(1):5–20

Plant Proteomics: An Overview

M. Asif Shahzad, Aimal Khan, Maria Khalid, and Alvina Gul

Contents

M.A. Shahzad • A. Khan • M. Khalid • A. Gul (✉)
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_12

Abstract The proteins encoded in a plant have a significant role in its survival and adaptation to external stresses. The cell wall, being the outermost layer, helps in defense against pathogens by production of glycoside hydrolases and proteases that degrade the pathogen external wall. The cell membrane assists in the movement of different molecules into and out of the cell. Different cells communicate with each other with the help of specific signals. Osmotic and salt concentrations are maintained by the embedded ion pumps in the cell membrane. The chloroplast, the only photosynthetic apparatus present in plants, leads to production of energy and also utilizes sunlight for the process of photosynthesis. A number of complex reactions, cycles, and pathways are present in the chloroplast. The mitochondria, also called the powerhouses of the cells, are rich in energy-producing cycles that are required for most of the activities of plants. In the matrix and cristae, a number of enzymes are active continuously. The mitochondrial membrane assists in the survival of the mitochondrion as an independent organelle. The nucleolus, the hub of all the protein-encoding genome, contains many processes. When we begin the analysis of a protein, protein extraction is the first issue. The plant possesses a cell wall that is a critical barrier which should be overcome. Many detergents and other chemicals are applied to break the bonding present in the cell wall, and we then extract our target protein, which is separated using gel electrophoresis. Two-dimensional sodium dodecyl-sulfate-polyacrylamide gel electrophoresis (2D SDS-PAGE) facilitates the reaction by separating the proteins with respect to isoelectric point as well as molecular weight. Target proteins are visualized and then digested in gel to process it further for identification of the protein. A mass spectrometer is applied for this purpose to characterize each protein on the basis of charge to mass ratio, leading to unambiguous results. Bioinformatics tools are also used for confirmation of our target protein.

Keywords Cell wall proteome • Lectin • Glycoside hydrolyses • Proteases • Membrane-anchored proteins • Chloroplast proteomics • Lipid metabolism • Carbon metabolism • Oxidative stress and its response • Stroma • Thylakoid membrane • Nucleolus proteomics

1 Introduction

Proteomics is the study of the protein population in a tissue, cell, or subcellular compartment. A proteome is a set of proteins that is expressed in an organism by its genome (Wasinger et al. 1995). Nowadays, the sequence of nucleotides that is obtained from a genome project is transformed to computer files so that better knowledge of those obtained sequences can be retrieved. Thus, from these kinds of projects, functional genes in the genome, the proteins encoded by them, and control of the expression of genes are identified (Lockhart and Winzler 2000). Half the nucleotide sequences do not match in homology with already known proteins, which is coming from genomics research (Maheshwari et al. 2001).

In many of the eukaryotic and prokaryotic organisms whose genome has been completely sequenced many of the protein functions are still unknown. So, the solution to this problem may be the use of "microarray chip" technology (Somerville and Somerville 1999). The functions of a protein not only depends on its nucleotide sequence but are also dependent on posttranslation modification and its interaction with other proteins. If a new open reading frame (ORF) is discovered, then the genomic program tries to find the function of that newly discovered ORF. An ORF results in protein synthesis, and proteins are responsible for the control of biological function. One estimation is that as a result of posttranslational modification, 300,000 human proteins are considered in the PTM product.

Proteomics as shown in Fig. 1 has become an essential tool for understanding the functions of biological processes at the molecular level (Hakeem et al. 2012). The journal *Plant Proteomics* publishes novel and important research articles in the field of proteomics that examine the function and interactions of proteins from plant systems.

2 Cell Wall Proteome

Before the 1980s, the cell wall was considered to be a rigid and static structure but recent research proves the cell wall is dynamic and flexible instead (Cassab 1998). The cell wall is composed of carbohydrates (polysaccharides), lignin (40 %), hemicellulose, pectin, and proteins (10 %). Many of the proteins present in the cell wall are important in the stability of the cell. These cell wall polymers bind covalently and noncovalently to form the functional cell wall.

It has been revealed that the cell wall contains only 5 % to 10 % proteins linked by covalent and noncovalent bonding, forming a network mostly containing proline-rich proteins. Cell wall proteomics as shown in Fig. 2 are divided into nine functional classes by studying the *Arabidopsis* plant. Proteins that act on the cell wall represent 25 % of cell wall proteomics, which include polysaccharide lyases (PLs), carbohydrate esterases (CEs), and many others. The second class represents oxidoreductases, which includes muticopper oxidases, berberine bridge enzymes, and blue copper-binding proteins. The cell wall proteomics has many other functions

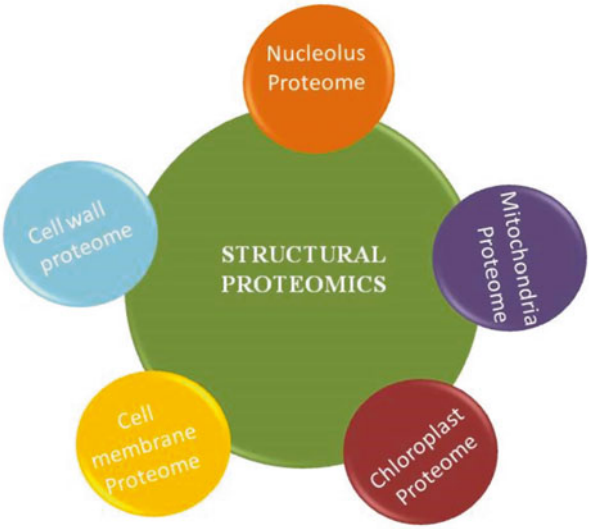


Fig. 1 Structural proteomics

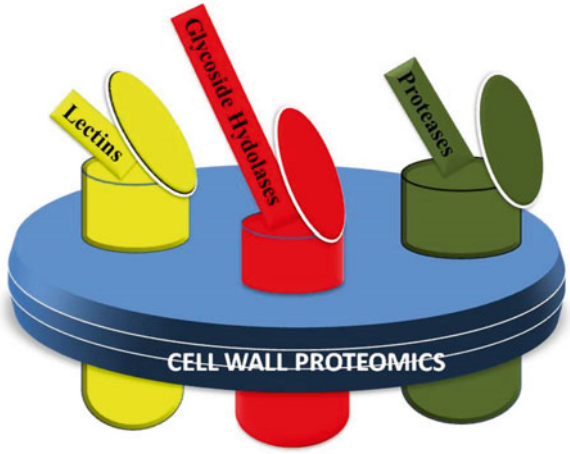


Fig. 2 Cell wall proteomics

such as protein turnover and protein maturation. One of the classes of cell wall proteomics involves signaling, which includes AGP proteins and other proteins such as involved in the transmembrane. Another class of proteins is involved in lipid metabolism and cuticle formation, such as lipid transfer proteins. Leucine-rich repeats (LRR) and glycine-rich proteins (GRPs) have been classified under structural proteins. Other classes include proteases, proteins acting on carbohydrates, miscellaneous proteins, and proteins with unknown functions (Albenne et al. 2013).

The class of cell wall proteomics is concerned with interacting domains, most probably with the carbohydrate-binding domain. Leucine-rich repeats are involved in protein–protein interactions. Those proteins that are not included in any other class have been grouped into miscellaneous proteins such as germins and purple acid phosphatases. There are some puzzling proteins whose functions are still unknown, categorized into the puzzling class of proteins. The function of one eighth of the cell wall proteome is still unknown (Albenne et al. 2013).

One of the structural proteins is hydroxyproline-rich glycoproteins (HRGPs), which are the translator product of the RSH (root-shoot-hypocotyl) gene. In *Arabidopsis thaliana*, the RSH gene is located on chromosome 1. Arabinogalactan proteins (AGPs) are glycoproteins that have been found to be functional in many processes including plant growth, development, embryogenesis, and cell proliferation (Knox 1995). The gene for AGP is on chromosome number 5. AGPs, the largest family, has the highest level of glycosylation of all the HRGPs, and are helpful in different stages of plant growth and plant development (Ellis et al. 2010).

AGPs are also involved in plant response to biotic and abiotic stresses (Gaspar et al. 2004). With the help of bioinformatics, 85 AGP genes were identified in *Arabidopsis thaliana* (Showalter et al. 2010). Another protein, proline-rich cell wall protein (PRP), is vital in specific cell wall structures during the development of the plant and is also involved in defense mechanisms against external damage and infection by pathogens (Ye et al. 1991). The PRP gene is located on chromosome 4 in *Arabidopsis*. Proline-rich proteins are the least glycosylated and thought to insert in the functional mature cell wall. Lectins are the glycoproteins. This carbohydrate protein is important in defense mechanisms against pathogen attack; lectin is important in protection of the cell against pathogens, and in symbiosis and assembly of cell wall polysaccharides. Another function of lectins is to recognize self-endogenous and exogenous ligands (Sharon and Lis 2004).

2.1 Glycoside Hydrolases

This large group of enzymes is involved in the hydrolysis of glycosidic bonds. They are important in plant cell wall metabolism, defense, signaling, and movement of storage reserves and reorganization of carbohydrates. A total of 75 glycoside hydrolases have been identified, subdivided into four groups. The first group of glycoside hydrolases consists of enzymes that are involved in organization of cell wall glycans in the period of growth and development (Minic and Jouanin 2006). Specific substrates for this small group of enzymes have also been identified; most of the substrates are pectin and xylologlucans. These two substrates are soluble in water and thus their enzymes, present outside the cell walls, are called exo-glycoside hydrolase. The multifunctionality of plant glycoside hydrolases makes the cell wall more complex by effectively modifying them even without the involvement of a large number of enzymes.

Some glycoside hydrolases are also involved in defense mechanisms against pathogens. These enzymes include chitinases (GH18, GH19, GH17), which are involved in activity against fungi (Schlumbaum et al. 1986). Plants secrete chitinases and β-1,3-glucanase in spaces within the cell to stop the growth of fungi by destroying their cell wall (Jach et al. 1995). Some glycoside hydrolases are multifunctional, such as β-D glucosidases (GH1), which have many functions such as lignification, signaling, defense, and hydrolysis of secondary metabolites (Xue et al. 1995). Some glycosidase hydrolases are thought to be involved in glycoprotein posttranslational modifications (PTMs) (Kotake et al. 2005).

2.2 Proteases

Proteases are essential in the functioning of enzymes and also protect from different microbes (Schaller 2004). Proteases comprise two sets of enzymes: the endopeptidases, that act upon the internal part of peptide chains, and the exopeptidases, which disrupt peptide bonds present on the termini of peptide chains (Barrett 1994). Research in proteomics has shown that proteases are much diversified in manner; there are multiple enzymes performing diverse functions as in the case of subtilases, carboxy peptidases, aspartases, and a number of other enzymes (Jamet et al. 2006). Enzymes of the subtilases family are especially interesting because they are vital in the production of peptide hormones and other growth factors from earlier peptides. Subtilases are involved in performing three main functions: control of development, protein turnover, and also as components of signaling cascades.

3 Cell Membrane Proteome

The plant cell membrane as shown in Fig. 3 is the outer layer of the protoplasm beneath the cell wall. As in most cell membranes, this is formed by the specific orientation of protein and phospholipid molecules. Cell membranes vary from 7.5 to 10 nm in thickness and consist of about 60% proteins and 40% phospholipids. Different species contain membranes consisting of specific types of polar lipids in specific concentrations that are probably genetically determined. Davson and Danielli (1935) proposed that membranes are made up of a central region consisting of phospholipids and an outer denser region composed of proteins (Danielli and Davson 1935).

The cell membrane is a particular compartment that is involved in structure formation as well as acting as a signaling mediator with extracellular content for the in-and-out movement of signals and materials. External stress results in intracellular reorganization of plants (Buchanan et al. 2000). So, a better knowledge of cell membrane proteins will facilitate logics to improve plant defense. Cell membranes help in controlling a number of critical functions such as metabolites, transport of ions, endocytosis, and cell division and differentiation. All these functions make use

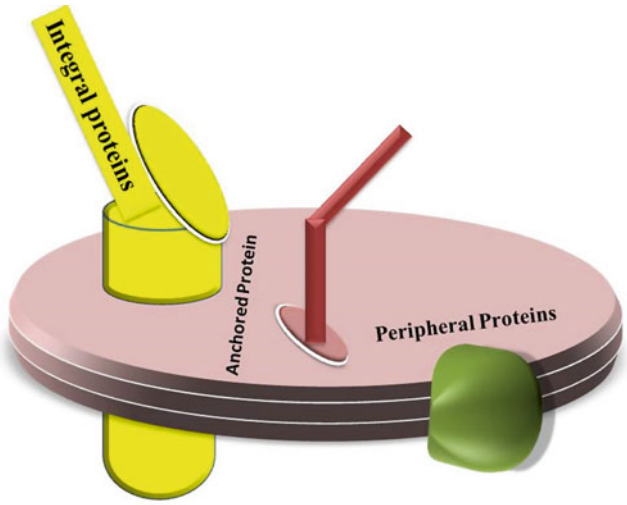


Fig. 3 Cell membrane proteomics

of a number of proteins that have very diverse structure and functions. In *Arabidopsis*, more than 27,000 proteins have been identified (Wortman et al. 2003).

Almost 25 % of proteins were grouped as membrane proteins (Schwacke et al. 2003). The plasma membrane is one of the most complicated systems of the cell that bears proteins and other components which are different with respect to type of cell, developing stage, and external environmental stress.

Cell membrane proteins are of three types:

- Transmembrane proteins
- Membrane-anchored proteins
- Peripheral membrane proteins

3.1 Transmembrane Proteins

Transmembrane proteins are those proteins that pass through the membrane into and out from the cell. The outer domain of the transmembrane protein may be a ligand-binding domain while its inner domain is called the effector domain (Marmagne et al. 2004). Receptors and transporters are the examples of transmembrane proteins, which are 30 % of the total cell membrane proteins (Ward 2001). In *Arabidopsis thaliana*, the total number of transmembrane proteins is considered to be between 4000 (15 %) and 8000 (30 %).

In *Arabidopsis thaliana*, 100 plasma membrane proteins were identified in a cell suspension culture in nonpolar proteins. It was prognosticated that 50 % of the transmembrane protein had domains (Marmagne et al. 2004).

Analysis of purely refined plasma membranes that are taken from leaves and petioles by using the technique named mass spectrometry led to the characterization of proteins as integrated and bordering proteins linked with the plasma membrane. About 238 plasma membrane proteins have been identified, and about 114 are expected to have some membrane domains or be phosphatidylinositol anchored. Of the 238 proteins identified, one third were categorized according to their function. Some families are involved in transport (17 %), signal transduction (16 %), membrane trafficking (9 %), and stress responses (9 %). About 25 % of the proteins that have been characterized in this analysis still lack information about any known function, and it is assumed that half of these somewhat resemble integral membrane proteins. In *Arabidopsis* more than 600 genes encode for protein kinase receptors, which are transmembrane proteins.

3.2 Membrane-Anchored Proteins

Many proteins of plant cells are affixed to membranes with the help of a covalent bond to glycosyl phosphatidylinositol (GPI). These types of proteins do not have a transmembrane domain and intracellular domain and thus are present entirely on the external side of the plasma membrane. GPI-anchored proteins include: enzymes associated to membrane, linkage molecules, initiation antigens, differentiation markers, protozoan coat components, and other miscellaneous glycoproteins. Many kinds of anchored proteins have been discovered in plants. Proteins that are translated in the vicinity of the plasma membrane involve signal transduction molecules such as GTPases and protein kinases. On the other hand, proteins that are anchored at the intracellular membrane include those proteins that help in regulation of vesicular movement (Fu and Yang 2001).

3.3 Peripheral Membrane Proteins

These proteins are lightly bonded to the membrane, most of the time being bonded noncovalently to the extended parts of integral membrane proteins. Peripheral membrane proteins are mostly the various proteins of multicomponent complexes such as photosynthetic proteins and H^+-ATPase in plants.

4 Chloroplast Proteomics

The chloroplast as shown in Fig. 4 is a plant cell organelle that originated from Cyanobacteria. They are involved in essential metabolic and major biosynthetic functions that include photosynthesis and amino acid biosynthesis. Most structural

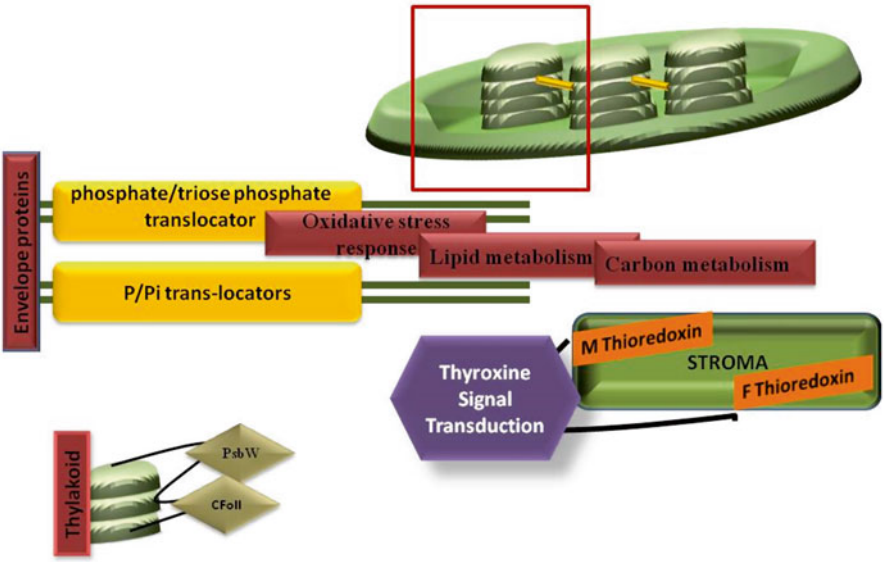


Fig. 4 Chloroplast proteomics

proteins that are involved in the construction of chloroplasts are transcribed by nuclear genes and transported into the chloroplast after being translated in the cytosol (Kleffmann et al. 2004). The three major structural and functional regions of chloroplasts are thylakoids, which are a highly organized membrane network composed of flat, compressed vesicles, the stroma, which is an amorphous matrix mostly rich in hydrophilic proteins and ribosomes, and the chloroplast envelope, a pair of outer membranes with an intermembrane space between them. The chloroplast is the site of carbon dioxide reduction and conversion into carbohydrates, amino acids, and fatty acids. Chloroplasts also bring about nitrite and sulfate reduction, converting them into amino acids (Douc and Joyard 1990; Flugge 2000).

Plastids within the plant cell involve a series of specific envelope proteins such as metabolite transporters, ion channels, pumps, and permease and pore proteins. A major feature of chloroplasts is that they resemble all other plastids and other self-existing organelles, which is why chloroplast formation requires the expression of both nuclear and plastid genes. Envelope membrane proteins take part in controlling the expression of nuclear and plastids genes and also in the specific makeup of chloroplast protein (Joyard et al. 1998).

Varying in size from 120 to 160 kb, the chloroplast genome codes for only 120 proteins; some RNA molecules are also encoded that are involved in transcription and translation of chloroplasts or in the code for small subunits of four complexes which are involved in photosynthesis (Sugita and Sugiura 1996). It is prognosticated that the number of different proteins present in the chloroplast is about 2000 to 5000, so most of the chloroplast proteins are the product of the nuclear genome. Those proteins that are transcribed by nuclear genes are produced as precursors in

the cytosol and are then transported to the chloroplast with the attached N-terminal peptide which is later degraded by proteases as it enters in the vicinity of the chloroplast.

Most of the chloroplast proteins are not actually transcribed by its own genome; rather, they are the product of nuclear genes, and then undergo translation in the cytoplasm and the protein is transported into this organelle. Transportation of these proteins is brought about by the protein import machinery that consists of an outer membrane, the Toc complex, and the inner membrane, the Tic complex (Schnell et al. 1994; Hiltbrunner et al. 2001).

Some polar lipids are also involved in the structural formation of plastid membranes (Douc and Joyard 1990). Some of the lipids are metabolized in the plastid membrane, resulting in many lipid-derived signaling molecules, which take part in the development and defense of the plant (Joyard et al. 1998). There are three membranes and six distinct compartments making up the structure of the chloroplast, which makes protein analysis much more difficult.

4.1 Chloroplast Envelope Proteins

The chloroplast envelope provides a complex network of transport systems that manipulate the in-and-out movement of proteins through the membrane. It also imports many ions and metabolites and also helps in movement out of the chloroplast. The envelope membrane has specific and unique biochemical machinery that depicts the stage of development of the plastid and also the metabolic requirements of different tissues.

It is difficult to identify the proteins especially in the case of plastid envelope membrane where the lipid content is very high (Douc and Joyard 1990). On the other hand, some transporter proteins, which are also embedded in a high content of lipid, cause problems in protein analysis. Because of the problems in the identification and analysis of proteins, very few proteins have been identified and assigned a function.

The first envelope protein to be characterized was phosphate/triose phosphate translocator (Flugge 2000). After that discovery, many substrate-specific outer membrane channels and some inner membrane metabolic translocators such as dicarboxylate and sugar were identified (Flugge 2000). The greatest molecular characterization of genes responsible for production envelope proteins was the characterization of several parts of the envelope proteins import machinery, that is, the Tic/Toc complexes.

To explore new envelope transporters, the cellular proteome analysis strategy was developed to characterize most of the nonpolar envelope proteins. This strategy relies on the utilization of highly refined and categorized membrane fractions, mining of nonpolar proteins with organic compounds, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) separation, and recurrent mass spectrometry analysis. To analyze that large amount of data, a tool based on BLAST

was designed to discover protein, expressed sequence tags (ESTs), and plant genome databases. Of the 54 characterized proteins, only 27 were new envelope proteins that had multiple α-helical transmembrane regions, which may be considered as envelope transporters (Myriam et al. 2002).

The present study of proteins has enabled us to find those with similar attributes among both already known and novel characterized envelope membrane transporters. These properties were utilized to walk through the complete *Arabidopsis* genome and helped us in making a virtual chloroplast envelope protein database. As a result, both protein analysis and in silico approaches helped us in identification of more than 50 candidates that were previously undiscovered as plastid envelope protein transporters. The function of proteins clears up the target of investigation that helps in understanding the chloroplast metabolism in a convenient way (Myriam et al. 2002).

4.2 Function of Envelope Proteins

Some envelope proteins are involved in the transport of ions and metabolites. A series of transport systems takes place in the chloroplast envelope (Seigneurin Berny et al. 1999). Triose P/Pi translocator is an important membrane protein present in the inner membrane of the chloroplast. The triose phosphate translocator that transports inorganic phosphate, 3-phosphoglycerate (3-PGA), and triose phosphate helps in the movement of photoassimilates out of the chloroplasts during the day. During the day, triose phosphates are expelled from the chloroplast stroma in counter exchange with inorganic phosphate (Pi), generated during sucrose synthesis in the aqueous part of the cytoplasm. Involved in photosynthetic acclimation, a light response results in increased tolerance to high-intensity light (Walters et al. 2003).

The transported phosphate is then used for ATP production in the light-dependent reaction. ATP produced as a result of this import is then used for other reactions in the citric acid cycle. The translocator protein exports carbohydrates produced during photosynthesis. The H^+ transporter is a low-affinity H^+/Pi chloroplast transporter that is involved in inorganic phosphate upward movement in the green parts of plants when the plant is Pi sufficient, which is required for Pi repeated translocation during Pi deficiency (Versaw and Harrison 2002). These reactions are dependent upon the tissue, are always expressed in young green tissues, and are also present in both auto- and heterotrophic tissues. It is also expressed in the root stele.

The ADP translocator is important in the import of ADP (Neuhas et al. 1997). The Pi transporter is a high-affinity transporter involved in the movement of external inorganic phosphate. It acts as an H^+ phosphate regulator in low- and high-Pi conditions. It is sensitive to the level of arsenate (Muchhal et al. 1996). In *Arabidopsis thaliana*, this protein is encoded by the PHT1-4 gene present on chromosome 2. It is tissue specific and is mostly expressed in roots, in tissues connecting the lateral roots to the primary root, and is also present in flowers, in senescing anther filaments, and in the abscission zone at the base of siliques. It is expressed in hydathodes and

axillary buds, and in some senescing leaves after Pi starvation, which is localized in all cells of undifferentiated root segments, including root tips and root hairs and in the epidermis, cortex, and stellar regions of mature root segments (Okumura et al. 1998). Some proteins have unpredictable functions: the purposes of these proteins are still unknown, and scientists are trying to determine these functions. Examples are HP45, HP34, and some other transporters as well.

- Amino acid transporters are also present in plants; the specific function of these transporter is still to be confirmed.
- Antiporters (Na^+/H^+) are also found in plants and are vital in the homeostasis of the plant body (Wang et al. 2002).

4.3 Lipid Metabolism

The chloroplast membrane takes part in lipid metabolism, produces lipid derivative growth factors, and forms defense compounds as a biological response to external stimuli. Many enzymes ais in lipid metabolism: acetyl Co-A carboxylase, acyl Co-A synthetase, and desaturases. During proteomic analysis of the chloroplast membrane of *Arabidopsis*, it was proved that chloroplast membranes are involved in lipid metabolism, thus giving more support to proteomics by considering it at the molecular level; for example, 2-lysophosphatidate acyl transferase is involved in glycerolipid biosynthesis.

4.4 Carbon Metabolism

The chloroplast is a region where acetyl Co-A carboxylase enzymes are attached with the inner envelope membrane and are believed to have a fatty acid biosynthetic machinery. The fatty acids that are synthesized in the chloroplast are either used or transported to the cytosol (Rolland et al. 1997). The chloroplast of cotton has enzymes for the Calvin cycle that are found free or membrane bound; those enzymes that are membrane bound have more significant activities compared to the free form. Examples of membrane-bound enzymes include glyceraldehyde phosphate dehydrogenase, phosphoglycerate kinase, and RuBisCo (Babadzhanova et al. 2002).

4.5 Oxidative Stress and Its Response

Plants experience a wide range of environmental stresses including light, drought, nutrient, and temperature changes that can cause oxidative stress by forming oxygen radicals. These radicals can cause major damage to membrane components

such as lipids and proteins. In such a situation, the large amount of fatty acid hydroperoxide within the membranes can be metabolized in the glutathione cycle. In *Arabidopsis thaliana* many proteins important in the oxidative stress response are present in the envelope membranes, for example, phospholipids, hydroperoxide, glutathione peroxidase (PHGPx), ascorbate peroxidase (APx), and superoxide dismutase (SOD).

The active repair mechanism present in plants effectively tackles the damage caused to a membrane protein under oxidative stress. Two protease families in *Arabidopsis* envelope membranes, the ATP-dependent Clp family and the ATP-dependent FtsH family, are involved in the removal of damaged protein from the envelope membrane. Clp proteins are also involved in the proper functioning of import machinery (Jarvis and Soll 2002).

4.6 Stroma

The chloroplast stroma is not only involved in the citric acid cycle but is also vital in the synthesis of proteins encoded by the organelle. In addition, the stroma also carries out starch and tetrapyrol synthesis. A recent application of proteomics in plants is the study of signal transduction in the chloroplast in which different new pathways were identified for thioredoxin-mediated signaling. In electron transport, the enzymes of the stroma are regulated by the activation of thioredoxin by protein ferredoxin thioredoxin reductase. Until now two chloroplast thioredoxins, m and f, have been identified as involved in signal transduction (Mottohashi et al. 2001). Thioredoxin interacts with its target proteins and forms an intermolecular disulfide bridge between the target proteins and thioredoxin. Then the formed disulfide bridge is reduced, but thioredoxin uses a second cysteine residue nearby, and the reduced target proteins now undergo a conformational change and are activated. If mutation occurs in cysteine being replaced by serine, then the target protein cannot be reduced, resulting in cross-linkage to the thioredoxin (Mottohashi et al. 2001).

4.7 Thylakoid Membrane

A thylakoid membrane is a site for photosynthesis with the help of photosynthetic pigments that are integrated in the membrane, being involved in adaptation to environmental changes. It has alternating dark and light bands of 1 nm. The thylakoid membrane of higher plants basically consists of phospholipids and galactolipids, which are irregularly organized on and across the membranes (Sprague 1987). The lipid synthesis for thylakoid membranes is brought about in the endoplasmic reticulum (ER) and the inner membrane of the plastid envelope, and movement from the inner membrane to the thylakoids is accomplished with the help of vesicles (Benning et al. 2006).

4.8 Thylakoid Lumen

The thylakoid lumen is the aqueous layer that is enclosed by the thylakoid membrane structure. It is involved in phosphorylation, which is an important part of the only photosynthetic process in plants. During daylight, these undergo a light-dependent reaction. For the proper functioning of the lumen, its PH is kept acidic, to pH 4, by the movement of H^+ ions into the lumen. This acidic pH is essential for proper carbon fixation (Lee and Kugrens 1999).

The thylakoid lumen is the least well characterized compartment as compared to other well-characterized parts. It is a continuous space that is present just beneath the thylakoid membrane. With the help of electron microscopy, the thylakoid lumen is seen as densely packed, which makes it even more difficult to characterize. The lumen helps in mediating the electron transport chain and other events related to it. It has also been reported to assist in producing the proton gradient that leads to ATP synthesis and also form currents by the ion channels (Pottosin and Schönknecht 1995).

It was discovered that all the nuclear genome encodes for luminal proteins and that these are produced as precursors in the cytoplasm. The amino-terminal end of these peptides helps in the movement of these proteins into the chloroplast stroma and across the thylakoid membrane into the lumen because they have specific amino-terminal bipartite transit peptides. According to this characteristic, bipartite transit peptides are used as markers for luminal proteins. On the other hand, not all synthesized chloroplast proteins are moved into the luminal space. The bipartite transit-formed peptides PsbW and CFoII are important proteins of the thylakoid membrane (Robinson et al. 1996).

4.9 Chloroplast Biogenesis

The mature chloroplast contains about 3000 proteins (Leister 2003). Metabolism inside the plastids is well organized but the functions of the proteins present are unknown or poorly described. Plastid proteins are encoded by both nuclear and plastid genes (Goldschmidt 1998). The plastids proteins translated from nuclear DNA are encoded in 80S ribosomes and then transported into the organelle. On the other hand, proteins that are encoded from the plastids genome are translated in 70S ribosomes. The chloroplast DNA only codes fewer than 100 proteins in higher plants, whereas the nuclear genome encodes about 95 % of different proteins present in the chloroplast proteome (Martin and Herrman 1998). Chlorobiogenesis is induced by certain environmental and internal signals between plastids and the nuclear genome.

5 Mitochondrial Proteomics

The mitochondria are membrane-bounded organelles responsible for the production of energy, involved in oxidation reactions and the transfer of electrons through the electron transport chain for the production of ATP. A mitochondrion is involved in the metabolism of amino acids and lipids and in the synthesis of nucleotides, vitamins, and cofactors, and it is also involved in the photo-respiratory pathway. Mitochondria are made up of hundreds of different proteins: most of the mitochondrial structural proteins are encoded in the nucleus, and during the formation of mitochondria those proteins are imported through complex enzymes. Another important reaction that occurs in mitochondria is phosphorylation: thus, there are two reactions occurring in mitochondria, oxidation and phosphorylation, and when these two reactions are combined, they are called oxidative phosphorylation, which is a source for the production of ATP in the cell.

5.1 Structure of Mitochondria

Mitochondria have outer and inner membranes as shown in Fig. 5 which basically have two major structural components: phospholipid and proteins (Alberts et al. 1994). The outer and inner mitochondrial membranes have different properties, and this double-membrane organization results in five distinct parts being present in the mitochondrion.

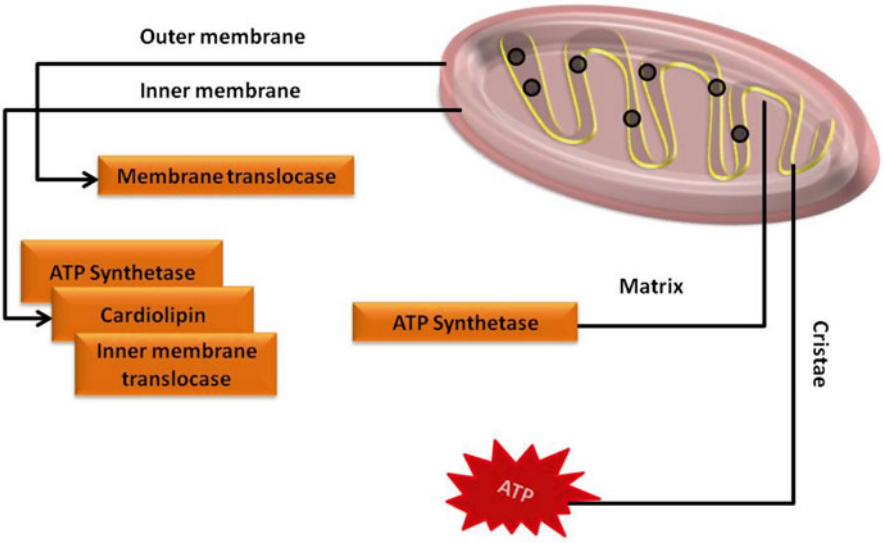


Fig. 5 Mitochondria proteomics

5.2 Outer Membrane

The outer mitochondrial membrane, which is largely made up of channel proteins called porins, surrounds the entire mitochondrion; it has a specific protein: phospholipid ratio similar to that of the eukaryotic plasma membrane (about 1:1 by weight). Porins facilitate molecules of 5 kDa or less in molecular weight to move freely across the membrane (Alberts et al. 1994). Larger proteins can be transported into the mitochondrion only when there is a signal sequence at their N-terminal that attaches to a peptide containing many subunits called membrane translocase, which then effectively moves them through the membrane (Herrmann and Neupert 2000).

Destruction of the outer membrane allows the proteins that are present in the space between the outer and inner membranes to move into the cytosol, causing the immediate death of a cell (Chipuk et al. 2006). The mitochondria-associated ER membrane (MAM) is a complex structure formed as a result of the attachment of the mitochondrial outer membrane to the endoplasmic reticulum membrane. This structure is critical in the mitochondria and with endoplasmic reticulum intracellular calcium signaling, aiding the movement of lipids across the endoplasmic reticulum and mitochondria.

5.3 Inter-Membrane Space

The inter-membrane space is present between the outer membrane and the inner membrane. The concentration gradient of ions and sugars (small molecules) in the space is equal to the concentration gradient of cytosol because the outer membrane is freely permeable to small molecules (Alberts et al. 1994). On the other hand, proteins that are larger in size or have high molecular weight should have a unique signaling sequence to be moved across the outer membrane, so the concentration gradient of the protein content of this space is different from the protein content of the aqueous part of the cytoplasm. Cytochrome *c* is one of the proteins attached to the inter-membrane space (Chipuk et al. 2006).

5.4 Inner Membrane

The inner mitochondrial membrane includes a number of proteins that perform five different kinds of functions. First are the proteins that perform specific reduction–oxidation reactions during oxidative phosphorylation; ATP synthase, which is involved in the production of ATP in the matrix of the mitochondria; specialized movement proteins that control metabolite movement across the matrix; and proteins in the intake machinery, mitochondria fusion, and fission protein.

There are 151 proteins present in the inner membrane, with higher protein content as compared to that of phospholipid. The inner membrane has more than 20 % of the total proteins in mitochondria (Alberts et al. 1994). In addition, the inner membrane is populated with an uncommon phospholipid named cardiolipin. This cardiolipin was first discovered in the heart of cows in 1942. It is mostly present in the mitochondrial as well as the plasma membrane of microbes (McMillin and Dowhan 2002). Cardiolipin consists of four fatty acids instead of two, and these aid in making the inner membrane impermeable to block undesired molecules from penetrating into the mitochondria. The inner membrane lacks specific structures called porins, which is why they are impermeable to all molecules. Most of the ions are transported across the matrix with the help of membrane transporters. Protein movement across the matrix is also carried out with the help of the translocase of the inner membrane (TIM) complex or with the help of Oxa1. After the action of a number of enzymes involved in the electron transport chain on the inner membrane, formation of the membrane potential results.

5.5 Cristae

The mitochondrial membrane has many compartments on its inner side, and each compartment is known as a crista. Cristae increase the surface area of the mitochondrial membrane, resulting in enhanced ability to produce ATP in a very small space. Mitochondria in liver cells have an inner membrane that is greater in size, almost fivefold larger than the outer membrane. The size of the inner membrane depends on the demand of ATP: the greater the demand for ATP, the greater the size of inner membrane will be. Muscle cells have a greater number of these compartments because of their high metabolic rate. These cristae are integrated into the tiny round bodies known as F1 particles, complex and irregularly organized folds attached to the inner membrane that can result in chemical osmotic disturbance (Mannella 2006).

5.6 Matrix

The matrix enclosed by the inner membrane has more than 70 % of the total proteins of the mitochondria (Alberts et al. 1994). The matrix has the majority of enzymes, as already discussed; the matrix thus is the site for many reactions including ATP formation through ATP synthase. The matrix is a complex part that does not contain only a single type of complexes. It has a number of enzymes, some tRNA, and hundreds of copies of the mitochondrial genome. The matrix mostly takes part in pyruvate oxidation and the tricarboxylic acid (TCA) or Krebs cycle. The mitochondrion is an independent compartment that has the ability to make its own RNA and many proteins (Alberts et al. 1994).

5.7 Mitochondria-Associated ER Membrane (Mam)

The mitochondria-associated ER membrane (MAM) is an important factor in the physiology of the cell and also the osmotic balance. Initially, there were some ER vesicles that appeared during cell fractionation, which was considered a problem in fractionating other components, but soon after it was discovered that these vesicles are actually the MAM, an important membranous structure that constitutes one fifth of the total membrane content of the mitochondrial outer membrane. The distance between ER and mitochondrion is about 25 nm, and the binding is maintained with the help of protein-binding complexes (De Brito and Scorrano 2010).

It has also been reported that the MAM are mainly rich in enzymes that help in the exchange of phospholipid and also for channels related to Ca^{2+} signaling (De Brito and Scorrano 2010). It is also involved in the regulation of lipids storage. Evolutionary analysis of MAM gives evidence that it was also involved in controlling calcium signaling (Hayashi et al. 2009).

5.8 Krebs Cycle

The Krebs cycle is an important metabolic process that helps in the production of the energy that is consumed by plants in performing a number of functions. A chain of reactions that make use of oxygen result in the oxidation of acetate, converting it into CO_2 and energy. This cycle is also a source of precursor molecules for other biochemical reactions. One example is ADP, which is a reducing agent. Because of is importance in pathways, it is also inferred that it might be one of the earliest pathways of cellular metabolism and might have been formed abiogenetically (Lane 2009).

5.9 Enzymes

Many enzymes have been are identified that cause the reactions to proceed, including pyruvate dehydrogenase complex, citrate synthase, isocitrate dehydrogenase, 2-oxyglutarate dehydrogenase, succinyl Co-A, succinate dehydrogenase, fumarase, and malate dehydrogenase. These enzymes make the TCA cycle happen.

5.10 Electron Transport Chain

The electron transport chain involves five complexes, namely, complex 1, 2, 3, 4, and 5:

- Complex 1 consists of five subunits (NADH oxidoreductase).
- Complex 2 consists of one subunit (succinate dehydrogenase).

- Complex 3 consists of three subunits (UQ-cytochrome oxidoreductase).
- Complex 4 consists of one subunit (cytochrome *c* oxidase).
- Complex 5 consists of five subunits (ATP synthase complex).

6 Nucleus Proteomics

Nuclear proteins are complex and multifunctional; some are mobile in nature such as transcription factor enzymes for processing of ribosomal RNA and DNA-repair enzymes. Some plant proteins are crucial in protein functions such as zinc finger proteins, glycine dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, transaldolase, actin, and malate dehydrogenase. Transcription regulators are the most dominant class of nuclear proteomics, such as RNA polymerase and TF 2A. Another class of proteins such as MADS box, RCC2, and Heat box are used for transcriptional control (Albenne et al. 2013; Narula et al. 2013). The nucleus matrix proteins as shown in Fig. 6 help in gene localization and expression. DNA methyltransferase and histone deacetylase help in mediating histone acetylation and modulating methylation of DNA by transposable elements silencing. Nuclear proteins assist in regulation of growth and development by transcriptional reprogramming such as ERF, MYB, Whirly, and WRKY factors to cause alteration in transcription factors (Casati 2012; Albenne et al. 2013).

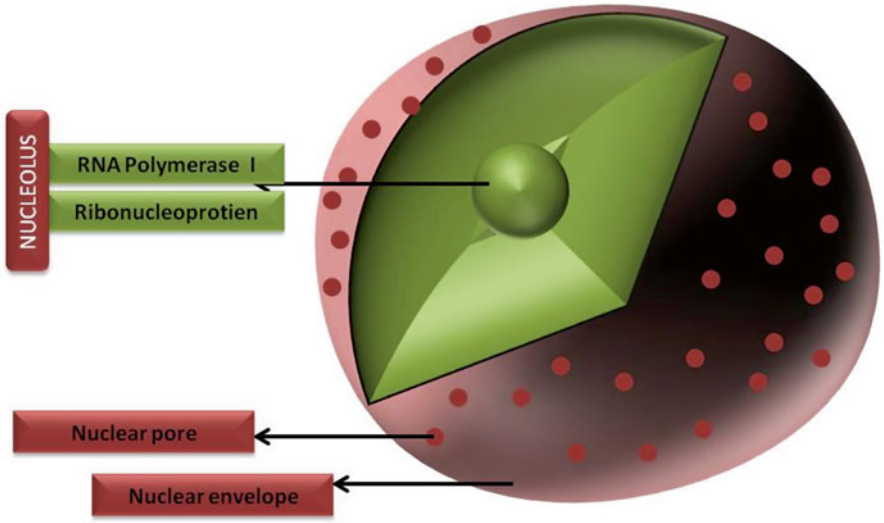


Fig. 6 Nucleus proteomics

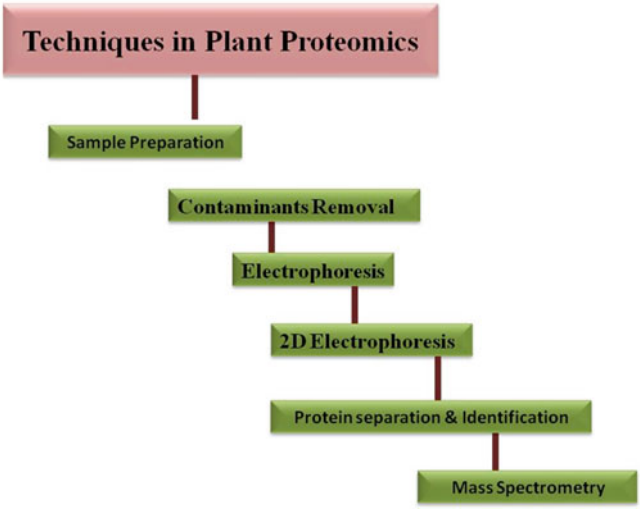


Fig. 7 Techniques in plant proteomics

Monocot nuclear proteins include SBT2 chromatin-binding proteins, RF2B, RING zinc finger proteins, and gypsy-like retroposon (Choudhary et al. 2009; Albenne et al. 2013). Some proteins that have been found in proteome analyses in the Leguminosae family are involved in different cellular functions such as embryogenesis, development, and identification of cells or organs, whereas CHP-rich zinc finger protein, nucleolin, WPP domain proteins, and MYB transcription regulators belong to Solanaceae transcription factors (Abe et al. 2003). Nuclear proteins also involve those involved in protein metabolism such as methyl transferase, ornithine methyl transferase, and homocysteine methyl transferase, which are important in DNA and RNA metabolism (Dahl et al. 2008; Albenne et al. 2013).

Many proteins are involved in cell-cycle regulation; the chromosome box proteins, BUB3 and RCC2 proteins, are important in cell division. CDC5 is an important cell-cycle protein in ciliary motility, trafficking, and mitosis. The Ran cycle is an important component represented by mago nashi proteins, Ran, Ras GTPase, and Ran GTPase. Karyopherins are nuclear proteins important in nuclear trafficking. Besides structural proteins and other proteins, the nucleus contains several enzymes such as polymerase, ligase, RNA processing enzymes, and gyrase. Another important class of nuclear proteins is histone variants that have an important function in maintenance of genome and stability (Ma et al. 2011; Albenne et al. 2013). Another class included mRNA processing proteins such as LSM2, nucleolar RNA processing protein (NRAP), and paraspeckle protein important in protein folding, signal transductors, and in developmental pathways (Caudron-Herger and Rippe 2012). For the maintenance of nuclear proteins and their stability are involved many other nuclear proteins such as different chaperons, proteasome subunit, HSP71, protein sulfide isomerase, DnaJ, and glutathione S-transferase.

Nuclear structural proteins are very diverse in plants. Myosin and actin are the most important, and the oldest proteins have been categorized in structural proteins and are active in DNA-dependent RNA polymerase and arbitrate in exporting RNA from the nucleus (Caudron-Herger and Rippe 2012). Some of the structural proteins are important in nuclear structure and the function of the genome such as tubulin, annexin A1, viscialin, titin, and spectrin. Some of the nuclear proteins are important in disease resistance, mRNA export, and chromosome scaffolding: NUF1, NUP88, NUP82, and nucleoporins.

In the eukaryotic nucleus, many nuclear bodies are involved in DNA replication and gene expression (Lamond and Earnshaw 1998). The nucleolus is a subcompartment of the nucleus and is a site for transcription and processing subunits before exporting to the cytoplasm (Lofantaine and Tollervey 2001). The nucleolus contains RNA polymerase I transcription enzymes and ribonucleoprotien particles, which are required for the modification of nucleotide bases and to process pre-RNA (Filipowicz and Pojacic 2002).

The nucleolus has ribosomal proteins, factors required for rRNA folding and for the export of ribosomal subunits from the nucleopore to the cytoplasm (Fatica and Tollervey 2002). The nucleolus has other functions as well, such as telomerase, cell cycle, cell growth, and aging (Lamond and Earnshaw 1998). The nucleolus is considered to act as a sensor for cellular stresses (Rubbi and Milner 2003). Structure of a plant nucleolus under the transmission electron microscope shows four different regions:

- Fibrillar centers (FC)
- Dense fibrillar components (DFC)
- Granular component (GC)
- Nucleolar cavity (NC)

The fibrillar is a small colorful structure present in the center, which is surrounded by a region of viscous colored material known as dense fibrillar component (DFC); this DFC is further surrounded by a region known as granular components (GC). The nucleolar cavity function is still unknown (Koberna et al. 2002).

Plant nucleoli have an active rDNA transcription unit that is spread throughout the entire nucleolus. DFC makes up 70 % of the nucleolar volume, which is surrounded by GC, whereas FC is not required for transcription.

7 Techniques in Plant Proteomics

The identification of proteins performing a specific function is opening new aspects in the field of molecular biology. Recent discoveries related to the proteins present in the cell wall (Robertson et al. 1997) and the plasma membrane have revolutionized that field (Santoni et al. 1998).

Another property of the plant cell is the presence of a cell wall, which is the outermost layer. Young plant cells are enclosed by a primary cell wall; some plants also do have a secondary wall that lies between cell and primary cell wall. The first step in protein analysis is disruption of the cell wall to access the proteome of that specific cell or tissue part. Many chemical and physical means are applied to disrupt the cell wall, such as lysis buffer, sonification, and high-speed blending (Islam et al. 2004).

Extraction of cell wall proteins (CWPs) is a difficult task; some loosely bound proteins have been reported, and there is no effective procedure for CWP extraction (Jamet et al. 2006). Proteomics involves a number of strategies and methodologies to analyze a specific protein, including protein separation and its identification. Subcellular proteomics and their functions are shown in (Table 1).

7.1 Sample Preparation

Preparing the sample is important for accurate results for protein analysis. Many problems appear; for example, in each cell there are a number of abundant proteins that may be present in the form of complexes, which makes it difficult to target our desired protein. Sample preparation is crucial when we looking for comparative proteomics as we look for even small differences between our controls and experimental results (Freeman and Hemby 2004). In a single cell, it is estimated that there might be ten copies of transcription factors but on the other hand there can be about 1 million copies of an abundant protein. This problem can be overcome by removing the most abundant protein, which will reduce the complexity of the whole sample.

Homogenization helps in making our sample uniform in terms of both composition and structure. It can be done by five different methods: mechanical, or by using ultrasonics, applying pressure, performing freeze-thaw, and osmotic and detergent lysis. Protein solubilization has special importance as it affects the efficiency of our expected final results and also affects the success of the whole experiment. Isolated proteins are often in insoluble form so we need to break the interactions present between them to make it soluble. It usually involves the breakage of interactions such as disulfide bonds, Van der Waals forces, and ionic and hydrophobic interactions (Rabilloud et al. 1997). An important point to be kept in mind while breaking these interactions is that there should be no modification or aggregation while solubilizing proteins as this can result in protein loss. Thus, we mostly use chaotropes such as urea/thiourea, detergents, reducing agents, tributyl phosphine, and protease inhibitors as sample buffer (Gorg et al. 2004). Chaotropes break hydrogen bonds and also disrupt hydrophilic interactions, which enables proteins to unfold. Reagents used mostly include urea, a neutral chaotropic agent used at high concentrations (5–9 M) to disrupt the secondary structure of proteins (Rabilloud et al. 1997).

Table 1 Subcellular proteomics and their functions

Subcellular organs	Proteins	References
Cell wall proteins	Structural proteins, oxidoreductases, protein related to lipid metabolism, miscellaneous proteins, proteins with unknown functions, proteases, proteins acting on carbohydrates, proteins with interacting domains	Albenne et al. (2013)
Cell membrane proteins	Signaling mediator, metabolite, transport of ions, endocytosis, and cell division and differentiation. Transmembrane protein as membrane trafficking and stress responses. Membrane-anchored protein as membrane trafficking, and stress responses. Peripheral protein mostly photosynthetic proteins and H^+-ATPase	Marmagne et al. (2004) Fu and Yang (2001)
Chloroplast proteins	Metabolite transporters, ion channels, pumps, and permease and pore proteins, thioredoxin proteins, chloroplast envelope proteins, stroma proteins, thylakoid proteins	Benning et al. (2006) Mottohashi et al. (2001)
Mitochondrial proteins	Translocase, mitochondria-associated ER membrane (MAM), translocase of inner membrane (TIM), succinyl Co-A, succinate dehydrogenase, fumarase	De Brito and Scorrano (2010) Hayashi et al. (2009)
Nucleolus proteins	Glycine dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, transaldolase, actin and malate, MADS box, RCC2	Filipowicz and Pojacic (2002)

Detergents break hydrophobic interactions, enabling the protein extraction. Detergents have been classified on the base of hydrophilic group, ionic, nonionic, and zwitterionic. Ionic includes anionic sodium dodecyl sulfate (SDS). Nonionic are uncharged, which includes octyl glucoside (Triton X-100). Zwitterionic are both positively and negatively charged groups with a net charge of zero: this includes CHAPS, 3-[(3-cholamidopropyl) dimethylammonio]-2-hydroxy-1-propanesulfonate (CHAPSO). Reductants break the disulfide bonds between cysteine residues, so they promote unfolding of proteins. Mostly, sulfhydryl reducing agents such as dithothreitol (DDT) and dithioerythritol (DTE) are used.

7.2 Contaminants Removal

pH and ionic strength affect the solubility of proteins. To maintain pH at a suitable level, buffers and salts are added to the sample. These buffer and salts often affect the further procedures in protein separation and spectrometric analysis, so they need to be removed before going further in our experiments (Visser et al. 2005). Salt added for preventing proteins from being precipitated interfere in 2-D electrophoresis. Salts are mostly removed by ultracentrifugation, gel filtration, and precipitation with TCA (Simpson 2004).

Detergents can be removed by using dialysis, that is, gel filtration chromatography. Lipids are also present in samples as some proteins are attached to lipids to form complex structures. This interaction with lipids reduces the solubility of proteins. A centrifugal filter device is used with the CHAPS, which allows effective lipid salt removal.

7.3 Protein Separation and Identification

Proteins in a cell vary from each other in their size and sequence of amino acids. There is not even one single method that can separate protein alone. Mostly procedures are optimized for a specific sample. Separation of protein is based upon the charge, size, and confirmation.

7.4 Electrophoresis

Polyacrylamide gel electrophoresis in SDS was first explored in 1949 (Rosenfeld et al. 1992). Separation is done under an electric field. As proteins carry a negative charge, the electric field forces its movement toward the other end. Protein migration depends upon its charge, size, and shape (Hale et al. 2004).

7.5 2D Gel Electrophoresis

The most widely used electrophoresis for separation of protein is by 2D PAGE (Gorg et al. 1999). For the past 20 years, it has also been used for the separation of a mixture of complex proteins with large molecular weight and sequence (Ong and Pandey 2001). In this approach, the protein is fractionated by isoelectric focusing in the first dimension in which separation is done based on the charge using a wide range of pI. In the second dimension, proteins are separated on SDS-PAGE based on their molecular weight (Gorg et al. 2000). A rapid improvement in resolution and reproducibility in the past 10 years has enhanced the popularity of 2DE (Rabilloud 2002). These improvements have resulted in the introduction of commercially available IPG systems that separate proteins within the narrow range of pI (Gorg et al. 2000), and also the availability of the latest fluorescent stains such as Sypro Ruby, which is highly sensitive and able to identify even proteins in very low abundance (Berggren et al. 2000). As a result, the 2DE is more applicable and reproducible with a higher capacity of loading and rapid identification of spot by utilizing the software (Wilkins et al. 1999). 2DE is the backbone in many protein analyses. After the protein has been separated on 2DE, it is visualized. After our target peptide is identified, it is excised from the gel with the help of the enzyme trypsin, and now mass spectrometric analysis can be done.

7.6 Identification by Mass Spectrometry

Developments made during the past few years in mass spectrometry have resulted in the assurance that it is the only method by which we can easily identify proteins; posttranslation modification can also be characterized (Packer and Harrison 1998). The latest ionization methods and other mass analyzer strategies have led to some astonishing improvements in this technology that have made it more accurate, higher in resolution, and also applicable to characterizing very large protein molecules (Costello 1999). Another major advancement in this technology is the introduction of the desorption method, allowing it to analyze large molecules without any complex fragmentation. The two most common methodologies used for protein identification and characterizations are matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) and electrospray ionization (Liang et al. 1996). Selection of the type of mass spectrometry is based upon such factors as the nature and source of the protein. In many of the proteome analyses, we always look to MALDI-TOF first. It is based on the mass-to-charge ratio of the peptide fragments that are produced by enzymatic cleavage of the parent protein using the trypsin enzyme. That fragmentation leads to unique peptides known as the peptide mass fingerprint (PMF) (Griffin et al. 2001). Bioinformatics strategies are also involved when a PMF is obtained. PMF along with the values of pI and molecular mass is compared to the theoretical values already present in the database, which helps in

removal of redundant data. These experimentally obtained values are compared to that reported in the database. After comparison by means of some algorithm, it gives results and computes scores.

In some cases, the PMF leads to some doubts in scores that can give a clear result. It happens when comparison of experimental results is done with the one already reported, and it gives more than two possible results on the basis of scores. In that situation we look for only one approach, that is, de novo sequencing, by a tandem mass spectrometer. A number of spectrometers can help in the analysis of proteins (Aebersold and Goodlett 2001). Data that are obtained by one or multiple MS in analysis of protein leads to characterization with a very high level of accuracy (Blackstock and Weir 1999). This methodology also gives information related to the posttranslation modifications in addition to the amino acid sequence (Bardor et al. 1999).

8 Future Perspectives

Plant proteomics is progressing continuously at a great pace in many aspects, and is also involved in developments in model systems, as in the case of yeast and *Escherichia coli* (Mann et al. 2001). With the advances of the technological era, revolutionary improvements in protein analysis and methodologies have made progress even more rapid. New analytical techniques are being introduced or are about to be applied in this field, which will make it easier to analyze even small samples with much greater accuracy and specificity. The future is expected to bring a lab-on-chip that will be able to able to analyze nano quantities (Nelson et al. 2000) of sample involving nano-separation with high resolution. Within a few years, developments will lead to the characterization of posttranslational modification and other protein complexes in an easy and specific way (Gavin et al. 2002). The development of a number of bioinformatics tools that are related to proteins will make data comparison and retrieval easier, hence making our proteomic approach rapid, and also allowing storage of vast amounts of data (Reiser et al. 2002).

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (1994) Molecular biology of the cell, 5th edn. Garland, New York

Abe H, Urao T, Seiki M, Shinozaki Y (2003) *Arabidopsis* AtMYC2 and AtMYB2 function as transcription regulator in abscisic acid signaling. Plant Cell 15:68–73

Albenne C, Canut H, Jamet E (2013) Plant cell wall proteomics: the leadership of *Arabidopsis thaliana*. Front Plant Sci 4:111

Aebersold R, Goodlett DR (2001) Mass spectrometry in proteomics. Chem Rev 101:269–295

Barrett AJ (1994) Classification of peptides. Methods Enzymol 244:1–15

Benning C, Xu C, Awai K (2006) Non-vesicular and vesicular lipid trafficking involving plastids. Curr Opin Plant Biol 9(3):241–247

Buchanan BB, Gruissem W, Jones RL (2000) Biochemistry and molecular biology of plants. American Society of Plant Physiologists, Rockville, MD

Babadzhanova MP, Babadzhanova MA, Aliev KA (2002) Free and membrane bound multienzyme complexes with Calvin cycle activities in cotton leaves. Russ J Plant Physiol 49:592–597

Berggren K, Chernokalskaya E, Stenberg TH, Kemper C, Lopez MF, Diwu Z, Haugland RP, Patton WF (2000) Background free, high sensitivity staining of proteins in one and two dimensional sodium dodecyl sulphate–polyacrylamide gels using a luminescent ruthenium complex. Electrophoresis 21:2509–2521

Blackstock WP, Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol 17:121–127

Bardor M, Loutelier BC, Marvin L, Cabanes MM, Lange C, Lerouge P, Faye L (1999) Analyses of plant glycoproteins by matrix assisted laser desorption ionization mass spectrometry: application to the N glycosylation analyses of bean phytohemagglutinin. Plant Physiol Biochem 37:319–325

Chipuk JE, Hayes BL, Green DR (2006) Mitochondrial outer membrane permeabilization during apoptosis: the innocent bystander scenario. Cell Death Differ 13(8):1396–1402

Cassab GI (1998) Plant cell wall proteins. Annu Rev Plant Physiol Plant Mol Biol 49:281–309

Casati P (2012) Recent advances in maize nuclear proteomic studies reveal histone modifications. Front Plant Sci 3:278

Costello CE (1999) Bioanalytical applications of mass spectrometry. Curr Opin Biotechnol 10:22–28

Caudron-Herger M, Rippe K (2012) Nuclear architecture by RNA. Curr Opin Genet Dev 22:179–187

Choudhary MK, Basu D, Datta A, Chakraborty S (2009) Dehydration responsive nuclear proteome of rice (*Oryza sativa* L.) illustrates protein network, novel regulators of cellular adaptation and evolutionary perspective. Mol Cell Proteomics 8:1579–1598

Danielli JF, Davson H (1935) A contribution to the theory of permeability of thin films. J Cell Comp Physiol 5(4):495

Dahl KN, Ribeiro AJS, Lammerding J (2008) Nuclear shape mechanics and mechanotransduction. Circ Res 102:1307–1318

Douc R, Joyard J (1990) Biochemistry and function of the plastid envelope. Annu Rev Cell Biol 6:173–216

De Brito OM, Scorrano L (2010) An intimate liaison: spatial organization of the endoplasmic reticulum–mitochondria relationship. EMBO J 29(16):2715–2723

Ellis M, Egelund J, Schultz CJ, Bacic A (2010) Arabinogalactan proteins: key regulators at the cell surface? Plant Physiol 153:403–419

Fu Y, Yang Z (2001) Rop GTPase: a master switch of cell polarity developments in plants. Trends Plant Sci 12:545–547

Flugge UI (2000) Transport in and out of plastids: does the outer envelope membrane control the flow? Trends Plant Sci 5:135–137

Freeman WM, Hemby SE (2004) Proteomics in protein expression profiling in neuroscience. Neurochem Res 29(6):1065

Filipowicz W, Pojacic V (2002) Biogenesis of small nucleolar ribonucleoproteins. Curr Opin Cell Biol 14:319–327

Fatica A, Tollervey D (2002) Making ribosomes. Curr Opin Cell Biol 14:313–318

Gaspar YM, Nam J, Schultz CJ, Lee LY, Gilson PR, Gelvin SB, Bacic A (2004) Characterization of *Arabidopsis* lysine rich arabinogalactin-protein AtAGP17 mutant (rat1) that results in a decreased efficiency of agrobacterium transformation. Plant Physiol 135(4):2162–2171

Goldschmidt CM (1998) Coordination of nuclear and chloroplast gene expression in plant cells. Int Ref Cytol 177:115–180

Gorg A, Obermair C, Boguth G, Harder A, Schiebe B, Wildgruber R, Weiss W (2000) The current state of two dimensional electrophoresis with immobilized pH gradients. Electrophoresis 21:1037–1053

Gorg A, Weiss W, Dunn MJ (2004) Current two dimensional electrophoresis technologies for proteomics. Protemics 4:3665

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature (Lond) 415:141–147

Gorg A, Obermair C, Boguth G, Csordas A, Diaz JJ, Madjar JJ (1999) Recent developments in two dimensional electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures. Electrophoresis 20:712–717

Griffin TJ, Goodlett DR, Aebersold R (2001) Advances in proteome analysis by mass spectrometry. Curr Opin Biotechnol 12:607–612

Hakeem KR, Chandna R, Ahmad P, Ozturk M, Iqbal M (2012) Relevance of proteomic investigations in plant stress physiology. OMICS J Integr Biol 16(11):621–635

Herrmann JM, Neupert W (2000) Protein transport into mitochondria. Curr Opin Microbiol 3(2):210–214

Hayashi T, Rizzuto R, Hajnoczky G, Su TP (2009) MAM: more than just a housekeeper. Trends Cell Biol 19(2):81–88

Hale JE, Butler JP, Gelfanova V, You JS, Knierman MD (2004) A simplified procedure for the reduction and alkylation of cysteine residues in protein prior to proteolytic digestion and mass spectral analyses. Anal Biochem 333:174–181

Hiltbrunner A, Bauer J, Alvarez HM, Kessler F (2001) Protein translocon at the *Arabidopsis* outer chloroplast membrane. Biochem Cell Biol 79:629–635

Islam N, Lonsdale M, Upadhyaya NM, Higgins TJ, Hirano H, Akhrust R (2004) Protein extraction from mature rice leaves for two dimensional gel electrophoresis and its application in proteome analyses. Proteomics 4:1903

Jarvis P, Soll J (2002) Toc, tic, and chloroplast protein import. Biochim Biophys Acta 1590:177–189

Jach G, Gornhardt B, Mundy J, Logemann J, Pinsdorf E, Leah R, Schell J, Maas C (1995) Enhanced quantitative resistance against fungal disease by combinatorial expression of different barley antifungal proteins in transgenic tobacco. Plant J 8:97–109

Jamet E, Canut H, Boudart G, Pont-Lezica RF (2006) Cell wall proteins: a new insight through proteomics. Trends Plant Sci 11(1):33–39

Joyard J, Teyssier E, Mege C, Berny SD, Marechal E, Block MA, Dorne AJ, Rolland N, Ajlali G, Douce R (1998) The biochemical machinery of plastid envelope membranes. Plant Physiol 118:715–723

Kotake T, Dina S, Konishi T, Kaneko S, Igarashi K, Samejima M, Watanabe Y, Kimura K, Tsumuraya Y (2005) Molecular cloning of a β-galactosidase from radish that specifically hydrolyses β-(1→3)- and β-(1→6)-galactosyl residues of arabinogalactan protein. Plant Physiol 138(3):1563–1576

Knox JP (1995) The extracellular matrix in higher plants: developmentally regulated proteoglycans and glycoproteins of plant cell surface. FASEB J 9:1004–1012

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjölander K, Gruissem W, Baginsky S (2004) The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol 14(5):354–362

Koberna K, Malinsky J, Pliss A, Masata M, Vecerova J, Fialova M, Bednar J, Raska I (2002) Ribosomal genes in focus: new transcripts label the dense fibrillar components and form clusters indicative of Christmas trees. J Cell Biol 157:743–748

Lee RE, Kugrens P (1999) Acidity of thylakoid lumen in plastids makes sense from an evolutionary perspective. Photosynthetica 37(4):609–614

Leister D (2003) Chloroplast research in the genome age. Trends Genet 19:47–56

Liang X, Bai J, Liu YH, Lubman DM (1996) Characterization of SDS-PAGE separated proteins by matrix assisted laser desorption/ionization mass spectrometry. Anal Chem 68:1012–1018

Lockhart DJ, Winzler EA (2000) Genomics, gene expression and DNA arrays. Nature (Lond) 405:827–835
Lane N (2009) Life ascending: the ten great inventions of evolution. Norton, New York
Lamond AI, Earnshaw WC (1998) Structure and function in the nucleus. Science 291:843–847
Lofantaine DLJ, Tollervey D (2001) The function and synthesis of ribosomes. Nat Rev Mol Cell Biol 2:514–520
Martin W, Herrman RG (1998) Gene transfer from organelles to the nucleus: how much, what happens and why? Plant Physiol 118:9–17
Minic Z, Jouanin L (2006) Plant glycoside hydrolases involved in cell wall polysaccharide degradation. Plant Physiol Biochem 44:435–449
Ma KW, Flores C, Ma W (2011) Chromatin configuration as a battle field in plant bacteria interaction. Plant Physiol 157:537–543
Marmagne A, Rouet MA, Ferro M, Rolland N, Alcon O, Joyard J, Garin J, Barbier-Brygoo H, Ephritikhine G (2004) Identification of new intrinsic proteins in *Arabidopsis* plasma membrane proteome. Mol Cell Proteomics 3:675–691
Myriam F, Daniel S, Rolland R, Thierry V, Daphné S, Didier G, Jérôme-Garin-Joyard J, Norbert R (2002) Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc Natl Acad Sci USA 9(17):11487–11492
Muchhal US, Pardo JM, Raghothama KG (1996) Phosphate transporter from higher plant *Arabidopsis thaliana*. Proc Natl Acad Sci USA 93:10519–10523
Mottohashi K, Kondoh A, Stumpp MT, Hisabori T (2001) Comprehensive survey of proteins targeted by chloroplast thioredoxin. Proc Natl Acad Sci USA 98:11224–11229
McMillin JB, Dowhan W (2002) Cardiolipin and apoptosis. Biochim Biophys Acta 1585 (2-3):97–107
Mannella CA (2006) Structure and dynamics of the mitochondrial inner membrane cristae. Biochim Biophys Acta 1763(5-6):542–548
Mann M, Hendrickson RC, Pandey A (2001) Analyses of proteins and proteomes by mass spectrometry. Annu Rev Biochem 70:437–473
Maheshwari SC, Maheshwari N, Sopory SK (2001) Genomics: DNA chips and a revolution in biology. Curr Sci 80:252–261
Nelson RW, Nedelkov D, Tubbs KA (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. Electrophoresis 21:1155–1163
Narula K, Datta A, Chakraborty N, Chakraborty S (2013) Comparative analyses of nuclear proteome: extending its functions. Front Plant Sci 4:100
Neuhas HE, Thom E, Mohlmann T, Steup M, Kampfenkel K (1997) Characterization of novel eukaryotic ATP/ADP translocator located in the plastid envelope of *Arabidopsis thaliana* L. Plant J 11:73–82
Okumura S, Mitsukawa N, Shirano Y, Shibata D (1998) Phosphate transporter gene family of *Arabidopsis thaliana*. DNA Res 5:261–269
Ong SE, Pandey A (2001) An evaluation of the use of two dimensional gel electrophoresis in proteomics. Biomol Eng 18:195–215
Pottosin II, Schönknecht G (1995) Ion channel permeable for divalent and monovalent cations in native spinach thylakoid membranes. J Membr Biol 148:143–156
Packer NH, Harrison MJ (1998) Glycobiology and proteomics: is mass spectrometry the Holy Grail? Electrophoresis 19:1872–1882
Robertson D, Mitchell GP, Gilroy JS, Gerrish C, Bolwell GP, Salabas AR (1997) Differential extraction and protein sequencing reveals major differences in patterns of primary cell wall proteins from plants. J Biol Chem 272:15841–15848
Reiser L, Mueller LA, Rhee SY (2002) Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems. Plant Mol Biol 48:59–74
Rubbi CP, Milner J (2003) Disruption of nucleolus mediates stabilization of p53 in response to DNA damage and other stresses. EMBO J 22:6068–6077

Rabilloud T, Adessi C, Giruadel A, Lunardi J (1997) Improvement of the solubilization of proteins in two dimensional electrophoresis with immobilized pH gradients. Electrophoresis 18(3-4):307

Rosenfeld J, Capdevielle J, Guillemot JC, Ferrara P (1992) In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. Anal Biochem 203: 173–179

Rolland N, Drone AJ, Amoroso G, Sultemeyer DF, Joyard J, Rochaix JD (1997) Disruption of the plastid ycf10 open reading frame affects uptake of inorganic carbon in the chloroplast of *Chlamydomonas*. EMBO J 16:6713–6726

Rabilloud T (2002) Two dimensional gel electrophoresis in proteomics: old, old fashioned, but it stills climb up the mountain. Proteomics 2:3–10

Robinson C, Knott TG (1996) Importing, sorting and assembly of photosynthetic proteins in higher plant chloroplasts. In: Andersson B, Salter AH, Barber J (eds) Molecular genetics of photosynthesis. Oxford University Press, Oxford, pp 145–159

Sharon N, Lis H (2004) History of lectins: from hemagglutinins to biological recognition molecules. Glycobiology 14:53R–62R

Showalter AM, Keppler B, Lichtenberg J, Gu DZ, Welch LR (2010) A bioinformatics approach to the identification, classification, and analyses of hydroxyproline-rich glycoproteins. Plant Physiol 153:485–513

Schlumbaum A, Mauch F, Vogeli U, Boller T (1986) Plant chitinases are potent inhibitors of fungal growth. Nature (Lond) 324:365–367

Schaller A (2004) A cut above the rest: the regulatory function of plant proteases. Planta (Berl) 220:183–197

Schwacke R, Schneider A, Van Der Graaff E, Fischer K, Catoni E, Desimone M, Frommer WB, Flugg UI, Kunze R (2003) ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins. Plant Physiol 131:16–26

Sprague SG (1987) Structural and functional consequences of galactolipids on thylakoid membrane organization. J Bioenerg Biomembr 19(6):691–703

Seigneurin Berny D, Rolland N, Garin J, Joyard J (1999) Technical advance: differential extraction of hydrophobic proteins from chloroplast envelope membranes: a subcellular-specific proteomic approach to identify rare intrinsic membrane proteins. Plant J 19:217–228

Schnell DJ, Kessler F, Blobel G (1994) Isolation of components of the chloroplast protein import machinery. Science 266:1007–1012

Sugita M, Sugiura M (1996) Regulation of gene expression in chloroplast of higher plants. Plant Mol Biol 32:315–326

Santoni V, Rouquie D, Doumas P, Mansion M, Boutry M, Dehais P, Sahnoun I, Rossignol M (1998) Use of proteome strategy for tagging proteins present at the plasma membrane. Plant J 1998(16):633–641

Simpson RJ (2004) Purifying proteins for proteomics. A laboratory manual. CSHL Press, New York

Somerville C, Somerville S (1999) Plant functional genomics. Science 285:380–383

Versaw WK, Harrison MJ (2002) A chloroplast phosphate transporter, PHT2;1, influences allocation of phosphate within the plant and phosphate starvation responses. Plant Cell 14: 1751–1766

Visser NF, Lingeman H, Irth H (2005) Sample preparation for peptides and proteins in biological matrices prior to liquid chromatography and capillary zone electrophoresis. Anal Bioanal Chem 382:535

Ward JM (2001) Identification of novel families of membrane proteins from the model plant *Arabidopsis thaliana*. Bioinformatics 17:560–563

Wortman JR, Haas Brian J, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Chunhui Y, Mulu A, Whitelaw CA, White OR, Christopher DT (2003) Annotation of *Arabidopsis* genome. Plant Physiol 132:461–468

Walters RG, Shephard F, Rogers JJ, Rolfe SA, Horton P (2003) Identification of mutants of *Arabidopsis* defective in acclimation of photosynthesis to the light environment. Plant Physiol 131:472–481

Wang HL, Postier BL, Burnap RL (2002) Polymerase chain reaction-based mutageneses identify key transporters belonging to multigene families involved in Na^+ and pH homeostasis of *Synechocystis* sp. Mol Microbiol 44:1493–1506

Wasinger VC, Cordwenn SJ, Cerpapoljak A, Yan OX, Gooley AA, Wilakins MR, Duncan MW, Harris KL, Humphrey SI (1995) Product with gene product mapping of the Mollicutes: *Mycoplasma genitalium*. Electrophoresis 16:1090–1094

Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL, Hochstrasser D (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. Plant Mol Biol 289:645–657

Xue J, Jorgensen M, Pihlgren U, Rask L (1995) The myrosinase gene family in *Arabidopsis thaliana*: gene organization, expression and evolution. Plant Mol Biol 27:911–922

Ye ZH, Song YR, Marcus A, Varner JE (1991) Comparative localization of three classes of cell wall proteins. Plant J 1:175–183

Proteomics of Bamboo, the Fast-Growing Grass

Tuan Noraida Tuan Hamzah, Khalid Rehman Hakeem, and Faridah Hanum Ibrahim

Contents

Abstract Bamboo is a vital non-timber plant in the world. Annually, it contributes up to 5 billion US dollars to the forest production. It is one of the fastest-growing plants on the earth. Due to its fast growth and abundant biomass, it is now considered as a good candidate for many important uses including production of biofuels. The feature of its fast growth is under scientific study at various levels. Earlier histological observations revealed that the cell division that occurs in the initial stage of the growth contributed heavily to the rapid growth of bamboo culms, followed by cell elongation during middle and late stages. In the current *Omics era*, proteomics is providing a novel dimension to understand the physiological as well as developmental processes in any organism. Currently, we are studying the growth

T.N.T. Hamzah • K.R. Hakeem (✉) • F.H. Ibrahim
Faculty of Forestry, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia
e-mail: kur.hakeem@gmail.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_13

characteristics of bamboo using proteomics as a molecular tool. This chapter discusses some of the latest proteomics studies related to bamboo growth and development.

Keywords Proteomics • Bamboo • Cell elongation • Cell division • Bamboo biomass

1 Introduction

With continued rapid development of global economy and constant increase in population, overall demand for natural resources will likely increase in the future, necessitating a sustainable supply of natural resources for upcoming generations. Food, shelter, clothing, and energy are some of the basic things required for human survival since time immemorial. Plants with fast growth and abundant biomass are, thus, required to fulfill any of these basic requirements. Bamboo is globally known for its rapid growth rate and also high yield renewable source (He et al. 2013). It is used as a raw material for paper and furniture making, producing various tools like chopsticks and tableware; as packaging materials; for production of biofuels and obtaining medicine and a variety of health care products (Hakeem et al. 2014). The fiber from bamboo is used for cloth making and for manufacturing other natural polymers (Xiang 2010; Saiter et al. 2013). Planting bamboo on steep slopes previously terraced for agricultural production has helped to stabilize the slopes and reduce runoff erosion in many areas, absorbing water from heavy rains that would cause flooding, and to provide shelter and protection for various animals (Moberg and Persson 2011). Other than culms, bamboo shoots are also consumed by people mostly in tropical countries. Annually, the consumption of the bamboo shoots reaches up to two million tonnes (Yang et al. 2008).

Bamboo contributes hugely to the economic growth of many countries. China is reported to have both the largest and the fastest growing bamboo sector involving more than ten million bamboo farmers, providing 35 million jobs (Buckingham et al. 2011; Yiping and Henley 2010; Hogarth and Belcher 2013; Xiang 2010) and generating a market value of over $10.5 billion (INBAR 2014). In the USA, the number of suppliers of bamboo flooring rose from less than 10 in the late 1990s to about 200 by 2005, with imports in 2005 approximating 45 million ft^2 (Malin and Boehland 2006). Numerous studies have documented improving household income and alleviating poverty through increased bamboo production (Perez et al. 2003; Marsh and Smith 2006; Booth 2013; Hogarth and Belcher 2013). Malaysia is now enhancing the forest plantation program, and bamboo is one of the potential non-timber species to be commercially planted (Hakeem et al. 2015).

Bamboo is in high demand for its fast-growing culms, and thus may provide a sustainable supply of natural resources for upcoming generations. There are a few scientific studies available to understand the growth and development of bamboo. However, the molecular mechanism behind its fast growth is yet to be explored in

detail. Other than hormones involved in cell growth and elongation, it is suggested that there must be other factors contributing to the rapid growth of bamboo culms. However, only quantitative data on bamboo shoot proteins have been generated so far (Nirmala et al. 2008; Waikhom et al. 2013).

To understand more the growth behavior and physical characteristics of bamboo, a lot of studies have focussed on anatomy of bamboo culms and generic growth style of bamboos (Lee and Chin 1960; Murphy and Alvin 1992; Lin et al. 2002). Apart from that, a study has been carried out to document the sequential elongation of internodes from base to top (Jiang 2002). A few studies before have successfully identified some internode elongation-associated genes in other plants from the Gramineae family, for instance EUI1, ACO1, SNORKEL1 and SNORKEL2, OsGLU1, SSD1, and CENL1 (Luo et al. 2006; Iwamoto et al. 2010; Hattori et al. 2009; Zhou et al. 2006; Asano et al. 2010; Ruonala et al. 2008). Despite the success in sequencing a set of cDNAs (Peng et al. 2010), ESTs (Zhou et al. 2010) and also generation of a monoclonal antibody bank (Wu et al. 2006), the molecular mechanism responsible for the rapid internode elongation of the culms remains elusive.

In the *Omics era*, proteomics approach has been used to know the molecular basis of an organism, not only its function but also the structure of the proteins, and genes involved in the process (Phizicky et al. 2003). Two-dimensional gel electrophoresis (2-DE), firstly introduced in the 1970s (Klose 1975; O'Farrell 1975), based technique is commonly used for protein separation followed by mass spectrometry (MS) for protein identification (Mann et al. 2001). Proteomics has quite recently been used to understand the various developmental processes in bamboo. It helped in revealing the pathways involved in the rapidly growing culms of Moso bamboo (He et al. 2013) and also understanding the morphological changes occurring during the floral transition (Kaur et al. 2015).

In this chapter, we discuss the various proteomics studies on some unique characteristics of bamboo.

2 Bamboo

Bamboos, giant woody grasses, fall under the monocotyledon group of angiosperms (Chapman 1996; Abd Latif et al. 1990). Referring to Grass Phylogeny Working Group (GPWG) (2001) Bambusoidea (bamboos) is one of the subfamilies of the grass family, Poaceae. It is separated into two tribes: Bambuseae (woody bamboos) and Olyreae (herbaceous bamboos). Apart from Bambusoidea, Centothecoideae, Arundinoideae, Pooideae, Chloridoideae, and Panicoideae are also the subfamilies of Poaceae. Bamboo comprises around 1200–1500 species and about 60–70 genera distributed all over the world (Wang and Shen 1987).

Bamboo is a non-timber species; however, bamboo is known as the most significant forest plant in the world (Peng et al. 2013a). Economically, around 2.5 billion people are depending on it, with trade value worth more than 2.5 billion US dollars annually (Lobovikov et al. 2007; Peng et al. 2013a, b). Having a distinct strength, a

rapid growth rate, and easy readjustment make bamboos the most significant forest resources. Unlike other plants, most bamboo species take about 2–4 months to reach their ultimate height of 15–30 m, and mature within 3–8 years of time (Chang and Wu 2000).

2.1 Botanical Description

A bamboo comprises two major parts: rhizomes and culms. Rhizomes are known to be the underground portion of the stem, while culms constitute the upper part. Most woody material is confined to the latter part. Despite its hard structure, culms actually lack bark, and the presence of silica produces a hard, smooth outer surface of the bamboo (Tewari 1992). Culms are endowed with a branching system, sheaths, foliage, flowers, fruits, and seedlings. These features are key in distinguishing bamboos from one another. Bamboos have a fast growth rate and may produce high yield output. The fast-growing behavior of bamboo results from the expansion of the individual internodes that are readily available in the bud (Magel et al. 2006). Within only 4–6 months, a bamboo might reach its maximal height, with 15–18 cm of daily increment. Culms require about 2–6 years of time to mature (Wong 1995).

2.2 Distribution

Bamboos are distributed all over the globe, with half of the species distributed in Asia Pacific, and the least distribution is in Africa (Bystriakova et al. 2003). Herbaceous bamboos are densely found in Brazil, Paraguay, Mexico, Argentina, and West Indies with around 110 species (Judziewicz et al. 1999). It is reported from New World, in Brazil, the bamboos constituted about 89 % of genera and also 65 % of the species (Filgueiras and Goncalves 2004). There are about 1290 woody bamboo species, grouped into three classes: (1) paleotropical woody bamboo (tropical and subtropical areas of Africa, Madagascar, India, Sri Lanka, southern China, southern Japan, and Oceania); (2) neotropical woody bamboos (Southern Mexico, Argentina, Chile, West Indies); and (3) north temperate woody bamboos (mostly in north temperate zones and some at high elevation habitats in Africa, Sri Lanka, Madagascar, and India).

2.3 Ecological Studies

Many bamboos prefer a habitat with a warm climate, abundant moisture, and productive soil, though few bamboos do grow in cold climes (Wang and Shen 1987). According to Grosser and Liese (1971), tropics and subtropics are the best

environment and condition for bamboos to grow well, though some taxa prefer the temperate climate. Most of the smaller bamboos are discovered in high elevations or temperate latitudes, while larger bamboos are mostly found in the tropic and subtropic areas (Lee et al. 1994). Bamboos could adapt to various kinds of habitat, excluding alkaline soils, deserts, and marshes; bamboos can grow in many types of soil: plains, hilly, and high altitude montane areas (Wang and Shen 1987).

2.4 Application of Bamboo

Traditionally, bamboo is widely used as a building material (Abd Latif et al. 1990). The products from bamboo stem or culm vary from household appliances to industrial products. Bamboo panels also have been used worldwide. Bamboo fibers are much longer than wood fibers. Bamboo panels are broadly used in recent construction. Food containers, chopsticks, boats, charcoal, and pulp and paper are few examples of the bamboo products. Besides the stem, extracts from bamboo also have been exploited for hair and skin ointment, and also for treating asthma. In Asia, bamboo shoots are cooked and eaten by people. Commonly, juvenile bamboo shoots are consumed as a vegetable or pickled, or can be processed by fermentation or deep frying, as shredded chips and canned into more presentable forms (Choudhury et al. 2011; Waikhom et al. 2013). A study on *Dendrocalamus giganteus* by Nirmala et al. (2008) demonstrated that bamboo shoots are a good source of proteins. Furthermore, Nirmala et al.(2008) found that fermented and fresh shoots of *D. giganteus* contained 2.17 and 3.11 g of proteins, respectively. Furthermore, bamboo shoots contain a high level of phytosterols, which actively act in lowering the blood cholesterol, and a high level of cellulosic content, a significant appetizer (Nirmala et al. 2011); they have anti-fatigue activity (Akao et al. 2004), high levels of antioxidant activity, microminerals, macrominerals, and high protein level per gram of dry weight (Waikhom et al. 2013). Notwithstanding the rich dietary and therapeutic traits reported for bamboo shoots of a few bamboo species (Akao et al. 2004; Waikhom et al. 2013), some species are rich in toxic cyanogenic-like taxiphyllin, which is significantly associated with a neurological disorder called Konzo (Nzwalo and Cliff 2011; Schwarzmaier 1997; Waikhom et al. 2013). Apart from that, bamboo shoot ash has been used to polish jewels. Due to greenhouse gas emissions and energy shortage problems, biofuels have been introduced as an alternative (Farrell et al. 2006; Salas Fernandez et al. 2009). Table 1 enlists some important uses of bamboo.

3 Proteomics

Proteomics is an analytical study of proteome (the protein complement of genome), involving its structures and functions (Pandey and Mann 2000; Patterson and Aebersold 2003; Phizicky et al. 2003; Hakeem et al. 2012). Scientists' interest is no

Table 1 Uses of bamboos

Parts	Uses	References
Skin and stratiform wall of bamboo culm	Mats, baskets, fans, slippers, curtains, fences, hats, ribbons, woven bamboo products, ropes, straps	Yuming et al. (2004)
Culm	Furniture, rafters, frames for lifting, scaffoldings, beanpoles, mine props, musical instruments	
Culm and wall materials	Houses, bridges, rafts, walls, floors, stools, construction materials, water channels, storage buckets	
Leaves	Treat cough and lung inflammation	
Lignocellulose	Bioethanol	Littlewood et al. (2013)
Fiber of bamboo culm	Polymer composite material	Zakhikhani et al. (2014)
		Okubo et al. (2004)
		Thwe and Liao (2003)
Fiber	Automotive parts	Davoodi et al. (2011)
Bamboo strips	Raw material for concrete formwork	Xiao and Ma (2012)
Culms and strips	Laminated floor tile	Dalcacio and Wiedemann (2010)

longer restricted to observing the phenotype of organisms, but expands to revealing the genes responsible for a phenotype (Scott and Ruedi 2003). In 1994, Marc Wilkins firstly introduced the term "Proteome" , to describe the entire complement of proteins expressed by a genome, cell, tissue or organism (Wilkins 2009). In 1995, Wasinger et al. coined the term "proteome," in a large scale genomic and cDNA sequencing project dealing with regulation and functions of sequenced genes. Since then, various high-throughput RNA measurement tools have been introduced, for instance, differential display, transcript imaging, DNA microarrays for transcriptome analysis. Unfortunately, these techniques do not offer insight into protein quality and also quantity, since mRNA production abundance does not guarantee the exact amount of protein (Gygi et al. 1999). Furthermore, many proteins undergo posttranslational modifications (PTMs), for instance, removal of signal peptides, phosphorylation, and ubiquitination that alter the protein's function and different subcellular localization (Gray et al. 2014).

Initially, the primary objective of a proteomic study is to identify the diversity of protein variants in cells and also tissues. Recently, the study has been extended to various functional aspects of proteins, for instance, posttranslational modifications, protein–protein interactions, activities and structures of proteins (Ohkmae 2003). Aiming for protein profiling, since the last two decades, proteomics has been associated with two-dimensional gel electrophoresis (2DE) for protein separation, and mass spectrometry (MS) for protein identification (Shevchenko et al. 1996; Wilkins et al. 1996). With time, these techniques have been developed upon and incorporated with other techniques aiming for high yield of proteins and easier identification of proteins functionally and structurally. In 1990s, a mass spectrometry based method called matrix-assisted laser desorption (MALDI-TOF) was developed (Yates 1998).

3.1 Techniques in Proteomics

Various techniques have been introduced in the proteomics area, including protein separation by 2D-PAGE and protein identification by MS. A summary of the techniques used in proteomics is given below.

3.1.1 Two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE)

For 40 years now, 2D-PAGE has been broadly used for big scale protein separation in proteomics area (Klose 1975; O'Farrel 1975). In this technique, proteins are separated in the first dimension, according to isoelectric point using isoelectric focusing (IEF) and its molecular weight in the second dimension using sodium dodecyl sulfate polyacrylamide gel-electrophoresis (SDS-PAGE) (Chen et al. 2010). The combination of these two methods has resulted in separation of thousands of proteins using only a single gel, recording two vital physical properties, the subunit molecular mass and the isoelectric point (Chen et al. 2010). The separated proteins are subjected to subsequent analyses including western blotting, gel visualization using pre-electrophoresis fluorescence labeling, post-electrophoresis involving the Coomassie blue staining, silver staining, differential expression analysis, and protein identification by mass spectrometry or Edman degradation method. The spot belonging to the protein of interest is excised from the gel and protease is used to digest the protein before proceeding with MS for protein profiling (Mann et al. 2001).

The instruments and materials for 2D-PAGE are readily available, and also easily applied, making 2D-PAGE as the most preferred technique by plant researchers for protein separation (Lee and Cooper 2006). Applying this technique has saved scientists time consumption in protein discovery. After some time, however, scientists were faced with limitations in the resolution of 2D-PAGE. Usually, 1000–2000 proteins can be displayed but not with plants, where the highly abundant ribulose bisphosphate carboxylase/oxygenase complex (RuBisCO) proteins supersede lots of low abundant proteins that would supposedly be clearly resolved (des Francs et al. 1985; Kim et al. 2001).

In due course, researchers have utilized immobilized pH gradient strips (IPGs) along with IEF which has improved the 2DE techniques in visualizing proteins on 2D gel, in a way that reduces the sample complexity and also overcomes the RuBisCo abundance (Gorg 1991). This has increased the number of gels to be tested, eventually increasing labor, cost, and also time. However, if the samples are first to be pre-fractionated or enriched, amount of proteins that may be displayed and analyzed will be higher (Stasyk and Huber 2004). Not only that, the proteins can also be separated on narrow-range or ultranarrow-range immobilized pH gradient strips, aiming to produce high amount of proteins for display and analysis purpose (Corthals et al. 2000; Görrg et al. 2009). Using this technique, for some reason, 2D-PAGE is not capable of resolving basic, hydrophobic, and also membrane spanning proteins (Santoni et al. 2000).

Nonetheless, by incorporating thiourea, acetonitrile, or detergents in 2D sample buffer, these hydrophobic proteins have been successfully analyzed (Görrg et al. 2009; Nouwens et al. 2000). An effective 2D separation of alkaline proteins can be achieved by combining various technologies, for example, by adding isopropanol to 2D rehydration buffer combined with application of pH gradients up to pH 12 (Hoving et al. 2002); also by using nonequilibrium pH gradient electrophoresis (NEPHGE), an alternative of IEF method, where the proteins do not accumulate at their isoelectric points and therefore less likely to precipitate, allowing basic and low abundance proteins to be resolved (Klose 1975; O'Farrell 1975).

3.1.2 Mass Spectrometry (MS)

Mass spectrometry (MS) analyses benefit the users with its high sensitivity and fast speed (Chen et al. 2010). Along with the development of technology, MS also has been innovated with desorption techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption (MALDI) aiming for the ionization of the peptide, without losing its structures. As for the target protein identification, peptide mass fingerprinting (PMF) or tandem mass spectrometric (MS/MS) data obtained from proteolytic digestion of peptide mixtures can be looked up against protein databases, in order to obtain the identity of the target protein (Mann et al. 2001).

A feature of PMF is that it is unique for each protein. Once it is produced, it will be matched with the theoretically derived PMF data that have been calculated for each entry in the database. Once the PMF of the target protein is matched with a specific protein candidate in the database, the identification is achieved (Perkins et al. 1999). Adding up to the PMF data, MS/MS spectra analysis revealed the structural information that is related to the peptide sequence, in which a benefit in identifying the target protein (Gygi and Aebersold 2000; Mann et al. 2001). In order to acquire the PMF data, a peptide mixture is first analyzed in normal MS mode. Then, during the MS/MS mode, the selected peptide ion or the parent ion is fragmented, resulting in the daughter ion. In the second part of the analysis, the daughter ion is separated and an MS/MS spectrum generated. The MS/MS spectra obtained for the target proteins are compared with the calculated spectrum for the whole peptides in the database to achieve profiling of the protein.

3.1.3 Multidimensional Protein Identification Technology (MudPIT)

Another advancement of proteomics techniques involving MS is MudPIT, a kind of high performance liquid chromatography (HPLC), for separating any basic, hydrophobic, and membrane-spanning proteins. According to a study done by Washburn et al. (2001), they have found that MudPIT can separate membrane-spanning proteins a lot more than 2D-PAGE. In 2D gels, only after separation the target protein is digested, allowing a high yield analysis of protein by MS (Wolters et al. 2001),

while by using the HPLC method, the samples are first digested before being separated in the MudPIT analysis (Fig. 1).

Separated peptides are then subsequently eluted into mass spectrometer and analyzed (Washburn et al. 2001). This method offers high throughput besides the advantage that the user would not have to wait for the analysis to finish. MudPIT is known to be compatible with quantitative and differential comparative analysis that is most possibly and commonly done using protein labeling (de Godoy et al. 2006; Gustavsson et al. 2005) in vivo or in vitro, that is, by labeling the peptides (Chen et al. 2006; Hu et al. 2006; Chen et al. 2005). One downside of MudPIT is that it requires multiple analyses in order to evade any random sampling errors associated with all MudPIT-style assays (Liu et al. 2004).

Not only 2DE-PAGE but MudPIT also has its own problems despite its high throughput ability for protein separation. Firstly, the sample being loaded on mass spectrometer directly, the usage of detergents to isolate hydrophobic proteins must be avoided, since they are already ionized, and they may mask the less-easily ionized peptides (Drexler et al. 2006). Fortunately, MS-compatible detergents are now available, and they should be incorporated into protein extraction protocols (Cadene and Chait 2000). Secondly, since thousands of tandem mass spectra are generated, use of software such as Sequest and Mascot is required (Eng et al. 1994; Perkins et al. 1999), or de novo sequencing programs (Shevchenko et al. 2002), in order to infer amino acid sequence information from the spectra. Thirdly, the protein information must be regenerated due to the digestion of the original mixture into peptides. Reassembly of peptides into proteins might take a long time since some proteins may share similar sequences.

3.1.4 One-Dimensional Liquid Chromatography Tandem Mass Spectrometry (1D-Gel-LC-MS/MS)

1D-Gel-LC-MS/MS is an emerging method for MudPIT-style separations, which is most preferred by researchers who cannot acquire the customized resources needed for MudPIT. Those more handy with gels may be fascinated by a technique combining reverse phase (RP) liquid chromatography with 1D gel separations. Using this method, proteins are firstly separated according to size on standard polyacrylamide gels (Breci et al. 2005; Laemmli 1970), or isoelectric point on IPG strips which is frequently applied in first dimensional separation in 2D-PAGE (Cargile et al. 2004). The 1D-gel separation step replaces one of the MudPIT techniques, that is, SCX separation, which reduces the sample complexity. As a protein separation is finished, the strip containing the protein is extracted and divided into multiple slices. The subsequent steps for the gel slices are similar to those spots which are excised from 2D gels. Later on, the peptides are separated on an RP column that is unified and reusable and which is coupled to a standard HPLC pump. Using MS/MS, the RP eluent is further analyzed. Since the gel slices constitute many different proteins which are later digested into peptides, a protein reassembly is necessary. The main convenience of this technique is that it offers the same solution as MudPIT, with

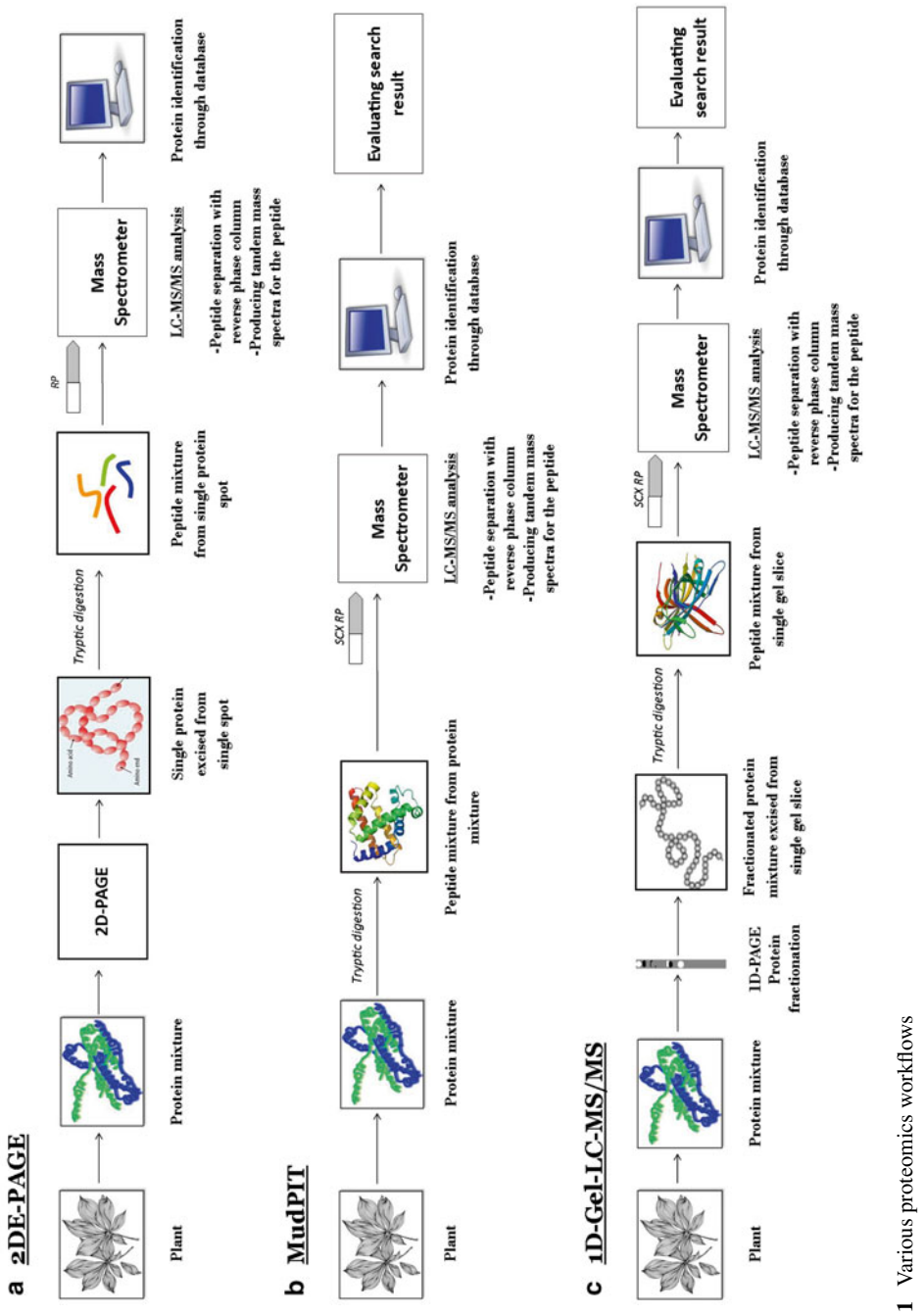


Fig. 1 Various proteomics workflows

easily acquired resources (Breci et al. 2005). Plant researchers have used this method for various proteomics studies, for instance subcellular, organelle, and membrane proteomics (Millar and Heazlewood 2003; Ferro et al. 2002; Maltman et al. 2002; Marmagne et al. 2004). A study by Lee and Cooper (2006) also proved that 1D gel separation followed by cleanup methods can efficiently diminish detergents that are not suited with MudPIT.

3.1.5 Difference Gel Electrophoresis (DIGE)

DIGE is a vital method for comparing proteomic research where it is exceptionally duplicable, precise, and sensitive. In the past, a conventional 2D gel generally involved the separation of a sample on a single gel, in which a dramatic variability due to gel-to-gel variations was prevalent. Then, perfectly suited for the time, DIGE was developed (Unlu et al. 1997) in order to improve duplicability when samples are being compared. In order to identify significant biological changes, in DIGE, two or more protein samples undergo the separation process on the same gel.

In short, the control and test samples are firstly labeled with for instance Cy3 and Cy5 fluorescent dyes, which are charge- and weight-matched, but are with different spectra; this can be done using minimal or saturation technique (Marouga et al. 2005). These dyes are both highly sensitive and offer a broad range of detection. Same amounts of protein samples which are labeled distinctly as mentioned above are mixed up and subject to 2DE separation on the same gel. This is to ensure that the same proteins from different samples co-migrate and that the fluorescence images of them are superimposed, thereby enabling more precise differential expression analysis.

During the image analysis, protein expression ratios between the samples are defined by comparing the standardized intensities of every protein spot from the Cy3 and also Cy5 channels. Lastly, the proteins with the distinctly changed level of expressions, also referred as proteins of interest, can be excised from the gel, manually or automatically, and processed for protein identification by MS (Marouga et al. 2005; Timms and Cramer 2008; van den Bergh and Arckens 2004). In the DIGE experimental design, a base component acquired is the incorporation of an internal standard, which commonly constitutes an equal mixture of the whole test samples (Alban et al. 2003). The adoption of internal standards has led to a gigantic increase in the incisiveness of this method, making it an essential method in protein quantification and also spot matching (Karp et al. 2007; Timms and Cramer 2008; van den Bergh and Arckens 2004).

Researchers are still debating on the DIGE technique, choosing the best way to perform the statistical analysis for the different comparison in reducing the figure of false positives (Karp et al. 2007; Urfer et al. 2006). Among other analyses, false discovery rate (FDR) technique is the most prominent statistical analysis approach (Karp et al. 2007). Overall, 2DE (including DIGE) incorporated with MS for identifying proteins is widely used for differential expression analysis (Jorrin-Novoa et al. 2009). Nonetheless, any examined proteome changes will have to be endorsed and

observed frequently for the sake of the proteins' molecular role in the process of interest (Minden et al. 2009).

4 Proteomics of Bamboo

Proteomics is an impressive tool to study all-out changes in protein synthesis patterns occurring during development and also in response to various environmental stimuli (Hakeem et al. 2012). Previous proteomics research showed that the fast growth features of bamboo culms is contributed to and regulated by a number of metabolic processes (Peng et al. 2013a). In one transcriptome analysis of *Dendrocalamus latiflorus*, several genes have been found to be responsible for encoding various functions of some key enzymes involved in regulation of plant hormones, lignin biosynthesis, growth and development (Finnie et al. 2006). As the genome of the Moso bamboo (*Phyllostachys edulis*) has been successfully sequenced and published, the identification and determination of the molecular regulation system of all functional genes in the bamboo becomes more feasible and reliable (Magel et al. 2006).

4.1 Analysis of Shoots and Rapidly Growing Culms of Bamboo

Cui et al. (2013) used proteomics tools, viz., two-dimensional electrophoresis (2-DE) combined with mass spectrometry (MS), in studying the protein expression during internode elongation process throughout the growth season of Moso bamboo (*Phyllostachys edulis*). The culms were collected from nine developmental stages. Each culm was divided into three portions, viz., (1) basal, (2) middle, and (3) top internodes. Overall 258 spots were obtained, among them 213 proteins were successfully identified. Among these 213 proteins, there were 189 unique proteins included, involved in various metabolic as well as physiological functions, viz., metabolism, energy, cell structure, signal transduction, and protein synthesis.

Recently Waikhom et al. (2015) studied the profile of crude proteins of edible bamboo shoots of 13 bamboo species and identify the prominent proteins. Standard one-dimensional SDS-gel polyacrylamide electrophoresis (SDS-PAGE) was performed in this study, which revealed high level peptide banding polymorphism within the range of 66.5–29.10 kDa and 20.10–15.50 kDa. From their study, Waikhom et al. have observed that the major proteins of bamboo shoots have a low molecular weight ranging from 20.10 to 15.50 kDa. Using this technique, abundant peptides could mask others and compromise resolution in SDS-PAGE, the approach allows for many samples on the same gel and subsequent accurate analysis of band patterns, an advantage over two-dimensional SDS-PAGE.

In their subsequent study, Waikhom et al. (2015), based on MALDI-TOF/TOF MS/MS results, identified the predominant bands as histone like proteins. Histones are highly alkaline proteins that are located in the nucleus and associated with DNA

to form chromatin. Five core histones are H1, H2A, H2B, H3, and H4, and have been extremely sustained in evolution, and under stress conditions, they are modified (Pawlak and Deckert 2007). So far, it has been shown that histones undergo various covalent modifications such as acetylation, methylation, phosphorylation, and ubiquitination, and these modifications control chromatin functions mediated by histones (Kouzarides 2007).

In an another study by Peng et al. (2013a, b), which was aimed to define some key genes involved in the mechanism of fast growth in bamboo shoots, using Illumina sequencing platform and the Moso bamboo genome database as a reference, the transcriptional changes that occur during shoot development have been reported. Transcriptome analysis indeed is necessary and vital in portraying the functional elements of the genome and affirming the molecular contents of cells and tissues (Wang et al. 2009; Wei et al. 2011). The genome reference is decisive for mapping in the transcriptome database as with it, the data will be more accurate and number of genes also will be sharply reduced. A previous study has reported that most of the transcription families play a critical role in plant growth, development and immunity (Severing et al. 2012).

4.2 Proteomic Study of Sporadic Flowering in Bamboo Species

The most distinctive feature in bamboo flowering boom is the prolonged vegetative phase that may last for decades (Janzen 1976; Sharma et al. 2014). This distinct characteristic feature has attracted the attention of many scientists to explore more beyond the genome of the bamboo to figure out why it does behave like that (Zhao et al. 2014). Bamboo sporadic flowering is hallmarked by an uneven distribution of flowers on fertile branches and fertile clumps (Waikhom et al. 2014), while in some cases flowering may lead to the death of whole clumps (Janzen 1976; John and Nagauda 2002). This makes flowering boom as one of the possible causes to the extinction of many woody bamboo species (Louis et al. 2014). With the advancement of technology, this is nothing much for the scientists, as for now, few studies have been done on bamboo biology and also on bamboo flowering (Louis et al. 2014). These include the description of syntenic genes between bamboo and other Poaceae plants (Gui et al. 2010), identification of the genes related to the programmed cell death after gregarious flowering in bamboo (Rai and Dey 2012), analysis of pollen structure from sporadic flowering of *Bambusa vulgaris* and *Dendrocalamus manipureanus* (Waikhom et al. 2014), and transcriptome analysis of bamboo floral tissues (Zhang et al. 2012; Gao et al. 2014). The establishment of a highly curated Moso bamboo database marked a great achievement in the field of bamboo biology (Zhou et al. 2014). Nonetheless, using scanning electron microscopy (SEM) provided evidence of differences in the development and structure of pollens in *B. vulgaris* and *D. manipureanus* (Waikhom et al. 2014). Recently, Louis et al. (2014) have used proteomics approach to explore the proteome in the floral transition, from the initiation to the advanced stages of sporadic flowering boom in

B. vulgaris and *D. manipureanus*, which aimed to find common and unique proteins involved in sporadic flowering boom in *B. vulgaris* and *D. manipureanus*. It has been suggested that sporadic boom is an energy-intensive process, associated with stress elements, mobile genetic elements, and signal transduction cross talk elements (Louis et al. 2014).

In their study, Louis et al. (2014) have identified SF-assemblin proteins in *B. vulgaris*. SF-assemblin is associated with microtubule formation during the cell cycle. Similarly, polyubiquitin has also been identified. Polyubiquitin is a biological catalyst responsible for modifying several proteins into some other thing, for example, translocation, assembly, or degradation. Also, some of the identified proteins are believed to be involved in signal transduction and energy-related processes.

According to Louis et al. (2014), the genes identified in transcriptome data are not often translated into functional proteins. Some important passenger genes are retrogenes whose mobility allows them to reach their transcriptional maximum in pollens (Abdelsamad and Pecinka 2014). Additionally, the FLOWERING LOCUS T (FT) is known to act as a mobile floral activator that is produced in leaf tissue and moves across to shoots (Louis et al. 2014). The importance of mobile elements is highlighted in a few previous studies (Liu et al. 2004; Abdelsamad and Pecinka 2014; Tamaki et al. 2014). Sporadic flowering is an active process which depletes energy, thus explaining the prompt death of bamboo clumps (Rai and Dey 2012). With regard to culms, shoots, and flowers of bamboo, a few proteins have been identified involved in the behavior of respective parts (Table 2).

In a recent study by Lin et al. (2010) and Liu et al. (2012), genes belonging to WD repeat-containing protein family, the MADS box family, and zinc finger protein family have been identified to be associated with bamboo flowering. On the other hand, members of MYB, WRKY, TGA, and NAC family have been found to act in a way that they respond to plant hormones, implying that they respond to the plant's environment (Tran et al. 2004).

5 Conclusion and Future Perspectives

Proteomic studies are helpful in understanding the various developmental as well as physiological processes of an organism. In the current chapter, we discuss some of the proteomic studies in understanding the growth and development of bamboo including the mysteries of its fast growth in elongating culms and shoots, their vegetative phase and also their sporadic flowering. Although the omics studies are widely gaining momentum in other fields of science, there is still a lot to understand and study in terms of plant growth and development. The current chapter provides a short snap of the latest proteomics updates in understanding the various processes of bamboo growth and development. However, proteomic analysis alone could not be an effective tool to understand the mechanism of pathways involved in the growth and development of bamboo; rather the integration of other omics fields could provide us with a clearer picture. Future studies will integrate the knowledge from all

Table 2 Proteins isolated and identified from various parts of bamboos using proteomics tools

Parts	Protein name	Functions	References
Culms	Elongation factor 1-delta	Translational elongation	Cui et al. (2013)
	S-adenosyl methionine synthetase	Synthesis of amino acid and hormones	
	Fructose-bisphosphate aldolase	Glycolysis	
	UDP-glucose phosphorylase	Synthesis of saccharides	
	Succinate dehydrogenase	Hormone biosynthesis	
	Malate dehydrogenase	Hormone biosynthesis	
	Phosphogluconate dehydrogenase	Hormone biosynthesis	
	Fructokinase	Regulating photosynthesis	
	Granule bound starch synthase	Starch and cell wall biosynthesis	
	Alpha-1,4-glucan-protein synthase	Starch and cell wall biosynthesis	
	UDP-glucose 6-dehydrogenase	Catalyzes the synthesis of UDP-glucuronate	
	Anthranilate synthase component I family protein	Tryptophan synthesis	
	Putative threonine synthase	Threonine synthesis	
	Arginase	Arginine synthesis	
	Argininosuccinate synthase	Arginine synthesis	
	N-acetyl-gamma-glutamyl-phosphate reductase	Arginine synthesis	
	Ketol-acid reductoisomerase	Isoleucine, valine, and leucine synthesis	
	Phospho-2-dehydro-3-deoxyheptonate aldolase	Chorismate biosynthesis	
	Hypothetical protein OsJ_26466	Fatty acid biosynthesis	
	3-oxoacyl-[acyl-carrier-protein] synthase	Fatty acid biosynthesis	
	Putative acetyl-CoA C-acyltransferase	Fatty acid beta oxidation	
	Fructose-bisphosphate aldolase	Enzymes of glycolytic pathway	
	Enolase	Enzymes of glycolytic pathway	
	NADP-dependent glyceraldehyde-3-phosphate-dehydrogenase	Enzymes of glycolytic pathway	
	Alcohol dehydrogenase	Enzymes of glycolytic pathway	
	Putative phosphoglycerate mutase	Enzymes of glycolytic pathway	
	Phosphoglycerate kinase	Enzymes of glycolytic pathway	
	Putative aconitate hydratase	Enzymes of TCA cycle	
	RuBisCO large subunit binding protein	Photosynthesis	
	Oxygen-evolving enhancer protein	Photosynthesis	
	NADH-ubiquinone oxidoreductase	Oxidative phosphorylation	
	Thioredoxin	Oxidative phosphorylation	
	ATPase	Oxidative phosphorylation	

(continued)

Table 2 (continued)

Parts	Protein name	Functions	References
Shoots	NADPH-producing dehydrogenase	Oxidative pentose phosphate pathway	Shoots
	6-phosphogluconate dehydrogenase	Pentose phosphate pathway	
	RNA-binding glycine-rich protein	Organ growth	
	Putative GAMYB-binding protein	Starch degradation	
	Transcription factor BTF3	Apoptosis	
	Putative legumin	Storage protein biogenesis	
	Protein disulfide isomerase-like 1-4	Storage protein biogenesis	
	Superoxide dismutase	Antioxidant enzymes	
	Peroxide dismutase	Antioxidant enzymes	
	Catalase	Antioxidant enzymes	
	Glutathione *S*-transferase	Antioxidant enzymes	
	USP family protein	Plant growth	
	Fructose bisphosphate aldolase	Formation of pyruvate	
Shoots	Phosphophenol pyruvate carboxylase	Carbohydrate metabolism	Waikhom et al. (2015)
	Histone H4	Forms nucleosome core	
	Histone H3	Forms nucleosome core	
	Histone H2A	Forms nucleosome core	
	RNA helicase	Unwind RNA	
Flowers	Protein kinase ADK-like protein	Signal transduction mechanism	Louis et al. (2014)
	Synaptobrevin-related protein	Intracellular trafficking and secretion	
	Small ribosomal protein 4	Ribosomes/RNA structures and translation	
	AtRAD3	Cycle regulation	
	SF-assemblin	Microtubular reestablishment	
	Putative cytosolic malate dehydrogenase	Carbohydrate metabolism	
	MLA6 protein-like	Stress/defense	
	Thaumatin-like protein	Stress/defense	
	Glutathione transferase-like protein	Stress/cellular detoxification	
	Putative tethering factor SEC34	Protein metabolism	
	Putative polyprotein	Intergrass-like protein	
	Putative gag-pol precursor	LTR-retrotransposons	
	Putative flavin-containing monooxygenase FMO-1	NAD(P)-binding domain	
	Chloroplast ATP-binding protein	Transport	
	Putative polyubiquitin (UBQ)	Regulate protein turnover	
	Ribulose-1,5-bisphosphate carboxylase	Carbohydrate metabolism	
	Putative fructokinase	Carbohydrate metabolism	
	Ribosomal protein subunit 4	RNA structures and translation	
	Rim2 protein	Transposases	
	Mitochondrial carrier-like protein (MCP)	Mitochondrial support	
	Polynucleotide phosphorylase	Metabolism of nucleic acid	

the omics data to come up with a big picture in understanding the various developmental processes of the fast-growing grasses including bamboo.

Acknowledgement The authors are thankful to Universiti Putra Malaysia for providing the Putra Grant (GB-IBT/2013/9418400) to support the bamboo proteomics research at the Faculty of Forestry.

References

Abd Latif M, Wan Tarmeze WA, Fauzidah A (1990) Anatomical features and mechanical properties of three Malaysian bamboos. J Trop For Sci 2(3):227–234

Abdelsamad A, Pecinka A (2014) Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming. Plant cell

Akao Y, Seki N, Nakagawa Y, Yi H, Matusumoto K, Ito Y, Ito K, Funaoka M, Maruyama W, Naoi M, Nozawa Y (2004) A highly bioactive lignophenol derivative from bamboo lignin exhibit a potent activity to suppress apoptosis induced by oxidative stress in human neuroblastoma SH-SY5Y cells. Bioorg Med Chem 12:4791–4801

Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, Currie I (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics 3(1):36–44

Asano K, Miyao A, Hirochika H, Kitano H, Matsuoka M, Ashikari M (2010) SSD1, Which encodes a plant-specific novel protein, controls plant elongation by regulating cell division in rice. Proc Jpn Acad Ser B Phys Biol Sci 86:265–273

Booth A (2013) Potential of bamboo to alleviate poverty in rural china remains untapped: expert. Center for International Forestry, January

Breci L, Hattrup E, Keeler M, Letarte J, Johnson R, Haynes PA (2005) Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing. Proteomics 5, 2018–2028, doi:10.1002/pmic.200401103

Buckingham K, Jepson P, Wu L, Rao V, Jiang S, Liese W, Lou Y, Fu M (2011) The potential of bamboo is constrained by outmoded policy frames. Ambio 40(5):544–548

Bystriakova N, Kapos V, Lysenko I, Stapleton CMA (2003) Distribution and conservation status of forest bamboo biodiversity in the Asia-Pacific Region. Biodivers Conserv 12:1833–1841

Cadene M, Chait BT (2000) A robust, detergent-friendly method for mass spectrometric analysis of integral membrane proteins. Anal Chem 72:5655–5688

Cargile BJ, Bundy JL, Stephenson JL (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. J Proteome Res 3(5):1082–1085

Chang ST, Wu JH (2000) Green-color conservation of ma bamboo (Dendrocalamus latiflorus) treated with chromium-based reagents. J Wood Sci 46:40–44

Chapman GP (1996) The biology of grasses. Department of Biochemistry and Biological Sciences, Wye College, University of London, UK CAB International, London, UK, pp 14–19

Chen Y, Mant CT, Farmer SW, Hancock RE, Vasil ML, Hodges RS (2005) Rational design of alpha-helical antimicrobial peptides with enhanced activities and specificity/therapeutic index. J Biol Chem 280:12316–12329

Chen R, Pan S, Yi EC, Donohoe S, Bronner MP, Potter JD, Goodlett DR, Aebersold R, Brentnall TA (2006) Quantitative proteomic profiling of pancreatic cancer juice. Proteomics 6(13): 3871–3879

Chen CY, Hsieh MH, Yang CC, Lin CS, Wang AY (2010) Analysis of the cellulose synthase genes associated with primary cell wall synthesis in Bambusa oldhamii. Phytochemistry 71: 1270–1279

Choudhury D, Sahu JK, Sharma GD (2011) Bamboo shoot based fermented food products review. J Sci Ind Res 70:199–203

Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC (2000) The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis 21:1104–1115

Cui K, He C, Zhang J, Duan A, Zeng Y (2013) Temporal and spatial profiling of internode elongation associated protein expression in rapidly growing culms of bamboo. J Proteome Res 11:2492–2507

Dalcacio R, Wiedemann EJ (2010) Product design in the sustainable era. TASCHEN, Germany

Davoodi MM, Sapuan SM, Ahmad D, Aidy A, Khalina A, Jonoobi M (2011) 32 Concept selection of car bumper beam with developed hybrid bio-composite material. Mater Des 10:4857–4865

de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol 7:R50

Drexler D, Barlow DJ, Falk P, Cantone J, Hernandez D, Ranasinghe A, Sanders M, Warrack B, McPhee F (2006) Development of an on-line automated sample clean-up method and liquid chromatography–tandem mass spectrometry analysis: application in an in vitro proteolytic assay. Anal Bioanal Chem 384:1145

des Francs CC, Thiellement H, de Vienne D (1985) Analysis of leaf proteins by two-dimensional gel electrophoresis. Plant Physiol 78(1):178–182

Farrell AE, Plevin RJ, Turner BT, Jones AD, O'hare M, Kammen DM (2006) Ethanol can contribute to energy and environmental goals. Science 311:506

Ferro M, Salvi D, Riviere-Rolland H, Vermat T, Seigneurin-Berny D, Grunwald D, Garin J, Joyard J, Rolland N (2002) Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc Natl Acad Sci U S A 99:11487

Filgueiras TS, Santos-Goncalves AP (2004) A checklist of the basal grasses and bamboos in Brazil (Poaceae). Bamboo Sci Cult 18:7–18

Finnie C, Bak-Jensen KS, Laugesen S, Roepstorff P, Svensson B (2006) Differential appearance of isoforms and cultivar variation in protein temporal files revealed in the maturing barley grain proteome. Plant Sci 170(4):808–821

Gao J, Zhang Y, Zhang C, Qi F, Li X, Mu S, Peng Z (2014) Characterization of the floral transcriptome of Moso bamboo (Phyllostachys edulis) at different flowering developmental stages by transcriptome sequencing and RNA-seq analysis. PLoS One 9(6):e98910

Gorg A (1991) Two-dimensional electrophoresis. Nature 349:545–546

Görrg A, Dres O, Luck C, Weiland F, Weiss W (2009) 2-DE with IPGs. Electrophoresis 30:122–132

Grass Phylogeny Working Group (2001) Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Miss Bot Gar 88:373–457

Gray SM, Cilia M, Ghanim M (2014) Circulative, "nonpropagative" virus transmission: an orchestra of virus-, insect-, and plant-derived instruments. Adv Virus Res 89:141–199

Grosser D, Liese W (1971) On the anatomy of Asian bamboos, with special reference to their vascular bundles. Wood Sci Technol 5:290–312

Gui YJ, Wang Y, Wang S, Wang SY, Hu Y, Bo SP, Chen H, Zhou CP, Ma NX, Zhang TZ, Fan LJ (2010) Insights into the bamboo genome: syntenic relationships to rice and sorghum. J Integr Plant Biol 52(11):1008–1015

Gustavsson N, Greber B, Kreitler T, Himmelbauer H, Lehrach H, Gobom J (2005) A proteomic method for the analysis of changes in protein concentrations in response to systemic perturbations using metabolic incorporation of stable isotopes and mass spectrometry. Proteomics 5(14):3563–3570

Gygi SP, Rochon Y, Br F, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730

Gygi SP, Aebersold R (2000) Mass spectrometry and proteomics. Curr Opin Chem Biol 4(5):489–494

Hakeem KR, Chandna R, Ahmad P, Ozturk M, Iqbal M (2012) Relevance of proteomic investigations in plant stress physiology. OMICS 16(11):621–635

Hakeem KR, Jawaid M, Rashid U (eds) (2014) Biomass and bioenergy. Springer International, Geneva

Hakeem KR, Ibrahim S, Ibrahim FH, Tombuloglu H (2015) Bamboo biomass: various studies and potential applications for value-added products. In: Hakeem KR, Jawaid M, Alothman OY (eds) Agricultural biomass based potential materials. Springer International, Switzerland, pp 231–243

Hattori Y, Nagai K, Furukawa S, Song XJ, Kawano R, Sakakibara H, Wu J, Matsumoto T, Yoshimura A, Kitano H (2009) The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. Nature 460:1026–1030

He C, Cui K, Zhan J, Duan A, Zeng Y (2013) Next-generation sequencing-based mRNA and microRNA expression profiling analysis revealed pathways involved in the rapid growth of developing culms in Moso bamboo. BMC Plant Biol 13:119

Hogarth N, Belcher B (2013) The contribution of bamboo to household income and rural livelihoods in a poor and mountainous county in Guangxi, China. Int For Rev 15(1)

Hoving S, Gerrits B, Voshol H, Muller D, Roberts RC, van Oostrum J (2002) Preparative two-dimensional gel electrophoresis at alkaline pH using narrow strip immobilized pH gradients. Proteomics 2:127–134

Hu J, Fu R, Nishimura K, Zhang L, Zhou HX, Busath DD, Vijayvergiya V, Cross TA (2006) Histidines, heart of the hydrogen ion channel from influenza A virus: toward an understanding of conductance and proton selectivity. Proc Natl Acad Sci 103:6865–6870

INBAR (2014) Bamboo: a strategic resource for countries to reduce the effects of climate change. Policy synthesis report. INBAR, Beijing, China

Iwamoto M, Baba-Asai A, Kiyota S, Hara N, Takano M (2010) ACO1, a gene for aminocyclopropane-1-carboxylate oxidase: effects on internode elongation at the heading stage in rice. Plant Cell Environ 33:805–815

Janzen DH (1976) Why bamboos wait so long to flower. Annu Rev Ecol Evol Syst 7(347):391

Jiang ZH (2002) Bamboo and rattan in the world. Liaoning Science and Technology Publishing House, Shenyang

John CK, Nadgauda RN (2002) Bamboo flowering and famine. Curr Sci 82:261–262

Jorrin-Novoa JV, Maldonadoa AM, Echevarria-ZOmenoa S, Valledorb L, Castillejoa MA, Cutoa M, Valeroa J, Sghaiera B, Donosoa G, Redondoa I (2009) Plant proteomics update (2007–2008): second generation proteomic techniques, an appropriate experimental design, and data analysis to fulfill MIAPE standards, increase plant proteome coverage and expand biological knowledge. J Proteomcs 72:285–314

Judziewicz EJ, Clark LG, Londono X, Stern MJ (1999) American bamboos. Smithsonian Institution Press, 392 pp

Karp NA, Kathryn S, Lilley KS (2007) Design and analysis issues in quantitative proteomics studies. Proteomics 7:42–50

Kaur D, Dogra V, Thapa P, Bhayyacharya A, Sood A, Sreenivasulu Y (2015) In vitro flowering associated protein changes in Dendrocalamus hamiltonii. Proteomics 15:1291–1306

Kim ST, Cho KS, Jang YS, Kang KY (2001) Two-dimensional electrophoretic analysis of rice proteins by polyethylene glycol fractionation for protein arrays. Electrophoresis 22(10):2103–2109

Klose J (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: a novel approach to testing for induced point mutations in mammals. Humangenetik 26:231–243

Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705

Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227:680–685

Lee AWC, Xuesong B, Perry NP (1994) Selected physical and mechanical properties of giant timber bamboo grown in South Carolina. Forest Prod J 44(9):40–46

Lee CL, Chin TC (1960) Comparative anatomical studies of some Chinese bamboos. Acta Bot Sin 9:76–95

Lee J, Cooper B (2006) Alternative workflows for plant proteomic analysis. Mol Bio Syst 2: 621–626

Lin JX, He XQ, Hu YX, Kuang TY, Ceulemans R (2002) Lignification and lignin heterogeneity for various age classes of bamboo (Phyllostachys pubescens) stems. Physiol Plant 114(2): 296–302

Lin XC, Chow TY, Chen HH, Liu CC, Chou SJ, Huang BL, Kuo CL, Wen CK, Huang LC, Fang W (2010) Understanding bamboo flowering based on large-scale analysis of expressed sequence tags. Genet Mol Res 9:1085–1093

Littlewood J, Wang L, Tumbull C, Murhy RJ (2013) Techno-economic potential of bioethanol from bamboo in China. Biotechnol Biofuels 6:173

Liu J, He Y, Amasino R, Chen X (2004) siRNAs targeting an intronic transposon in the regulation of natural flowering behaviour in Arabidopsis. Genes Dev 18923:2873–2878

Liu MY, Qiao GR, Jiang J, Yang H, Xie LH (2012) Transcriptome sequencing and De Novo analysis for ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina Platform. PLoS One 7(10):e46766

Lobovikov M, Paudel S, Piazza M, Ren H, Wu J (2007) World bamboo resources: a thematic study prepared in the Framework of the Global Forest Resources Assessment 2005. Food and Agriculture Organization of the United Nations, Rome, pp 1–73

Louis B, Waikhom SD, Roy P, Bhardwaj PK, Singh MW, Sharma KC, Talukdar NC (2014) Invasion of Solanum tuberosum L. by Aspergillus terreus: a microscopic and proteomics insight on pathogenicity. BMC Res Notes 7:350

Luo AD, Qian Q, Yin HF, Liu XQ (2006) EUI1, encoding a putative cytochrome P450 monooxygenase, regulates internode elongation by modulating gibberellin responses in rice. Plant Cell Physiol 47:181–191

Magel E, Kruse S, Lutje G, Liese W (2006) Soluble carbohydrates and acid invertases involved in the rapid growth of developing culms in Sasa palmata (Bean) Camus. Bamboo Sci Cult 19(1):23–29

Malin N, Boehland J (2006) Bamboo in construction: is the grass always greener? AIA Architect, April

Maltman DJ, Simon WJ, Wheeler CH, Dunn MJ, Wait R, Slabas AR (2002) Proteomic analysis of the endoplasmic reticulum from developing and germinating seed of castor (Ricinus communis). Electrophoresis 23:626–639

Mann M, Hendrickson RC, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem 70:437–473

Marmagne A, Rouet MA, Ferro M, Rolland N, Alcon C, Joyard J, Garin J, Barbier BH, Ephritikhine G (2004) Identification of new intrinsic proteins in Arabidopsis plasma membrane proteome. Mol Cell Proteomics 3:675–691

Marouga R, David S, Hawkins E (2005) The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem 382:669–678

Marsh J, Smith N (2006) New bamboo and pro—Poor impacts: lessons from China and potential for Mekong Countries. In: Proceedings, International Conference on Managing forests for poverty reduction: capturing opportunities in forest harvesting and wood processing for the benefit of the poor. FAO Regional Office for Asia and the Pacific

Millar AH, Heazlewood JL (2003) Genomic and proteomic analysis of mitochondrial carrier proteins in Arabidopsis. Plant Physiol 131(2):443–453

Moberg J, Persson M (2011) The Chinese Grain for Green Program—Assessment on the land reform's carbon mitigation potential. Master of Science Thesis, Chalmers University of Technology, Department of Energy and Environment

Murphy RJ, Alvin KL (1992) Variation in fiber wall structure in bamboo. Int Assoc Wood Anat Bull 13:403–410

Nirmala C, Sharma ML, David E (2008) A comparative study of nutrients components of freshly harvested, fermented and canned bamboo shoots of Dendrocalamus giganteus Munro. J Am Bamboo Soc 21:33–39

Nirmala C, Bisht MS, Sheena H (2011) Nutritional properties of bamboo shoots: potential and prospects for utilization as a health food. Compr Rev Food Sci Food Saf 10(153):165

Nouwens S, Cordwell SJ, Larsen MR, Moloy MP, Gillings M, Willcox MDP, Walsh BJ (2000) Complementing genomics with proteomics: the membrane subproteome of Pseudomonas aeruginosa PAO1. Electrophoresis 21:3797–3809

Nzwalo H, Cliff J (2011) Konzo: from poverty, cassava, and cyanogen intake to toxico nutritional neurological disease. PLoS Negl Trop Dis 5(6):e1051

O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. J Biol Chem. 250:4007–4021

Ohkmae K (2003) Proteomic studies in plants. J Biochem Mol Biol 37(1):133–138

Okubo K, Fujii T, Yamamoto Y (2004) Development of bamboo-based polymer composites and their mechanical properties. Composites A 35:377–383

Pandey A, Mann M (2000) Proteomics to study genes and genomes. Nature 405:837–846

Patterson SD, Aebersold RH (2003) Proteomics: the first decade and beyond. Nat Genet Suppl 33:311–323

Pawlak S, Deckert J (2007) Histone modifications under environmental stress. Biol Lett 44:65–73

Peng Z, Lu T, Li L, Liu X, Gao Z, Hu T, Yang X, Feng Q, Guan J, Weng Q, Fan D, Zhu C, Lu Y, Han B, Jiang Z (2010) Genome wide characterization of the biggest grass, bamboo, based on 10,608 putative full-length cDNA sequences. BMC Plant Biol 10:116

Peng Z, Zhang C, Zhang Y, Hu T, Mu S, Li X, Gao J (2013a) Transcriptome sequencing and analysis of the fast growing shoots of Moso bamboo (Phyllostachys edulis). PLoS One 8(11):e78944

Peng ZH, Lu Y, Li LB, Zhao Q, Feng Q (2013b) The draft genome of the fast-growing non timber forest species Moso bamboo (Phyllostachys heterocycla). Nat Genet 45(5):456–461

Perez M, Belcher B, Fu M, Yang X (2003) Forestry, poverty, and rural development: perspectives from the bamboo subsector. In: Hyde W, Belcher B, Xu J (eds) China's forests—global lessons from market reforms. Resources for the Future/Center for International Forestry, Washington, DC

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20(18):3551–3567

Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S (2003) Protein analysis on a proteomic scale. Nature 422:208–215

Rai V, Dey N (2012) Identification of programmed cell death related genes in bamboo. Genes 497(2):243–251

Ruonala R, Rinne PLH, Kangasjärvi J, Van Der Schoot C (2008) CENL1 expression in the rib meristem affects stem elongation and the transition to dormancy in populus. Plant Cell 20:59–74

Saiter J-M, Esposito A, Dobircau L, Turner JA, Adhikari R (2013) Synthetic polymer composites reinforced by bamboo fibers. Macromol Symp 327:114–120

Salas Fernandez MG, Becraft PW, Yin Y, Lübberstedt T (2009) From dwarves to giants? Plant height manipulation for biomass yield. Trends Plant Sci 14:454–461

Santoni V, Kieffer S, Desclaux D, Masson F, Rabilloud T (2000) Membrane proteomics: use of additive main effects with multiplicative interaction model to classify plasma membrane proteins according to their solubility and electrophoretic properties. Electrophoresis 21: 3329–3344

Schwarzmaier U (1997) Cyanogenesis of Dendrocalamus: taxiphyllin. Phytochemistry 16:1599–1600

Scott DP, Ruedi HA (2003) Proteomics: the first decade and beyond. Nature genetics 33:311–323

Severing EI, Dijk AD, Morabito G, Busscher-Lange J, Immink RG (2012) Predicting the impact of alternative splicing on plant MADS domain protein function. PLoS One 7:e30524

Sharma HR, Yadav S, Deka B, Meena RK, Biskat NS (2014) Sporadic flowering of Dendrocalamus longispathus (Kurz) Kurz in Mizoram, India. Trop Plant Res 1(1):26–27

Shevchenko A, Chernushevich I, Shevchenko A, Wilm M, Mann M (2002) "De novo" sequencing of peptides recovered from in-gel digested proteins by nanoelectrospray tandem mass spectrometry. Mol Biotechnol 20:107–118

Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci U S A 93:14440–14445

Stasyk T, Huber LA (2004) Zooming in. Fraction Strateg Proteomics 4(12):3704–3716
Tamaki S, Tsuji H, Matsumoto A, Fujita A, Shimatani Z, Terada R, Sakamoto T, Kurata T, Shimamoto K (2014) FT-like proteins induce transposon silencing in the shoot apex during floral induction in rice. Proc Natl Acad Sci U S A 112(8):E901–E910
Tewari DN (1992) A monograph on bamboo. International Book Distribution, Dehra Dun
Thwe MM, Liao K (2003) Durability of bamboo-glass fiber reinforced polymer matrix hybrid composites. Compos Sci Technol 63:375–387
Timms JF, Cramer R (2008) Difference gel electrophoresis. Proteomics 23–24:4886–4897
Tran L, Nakashima K, Sakuma Y, Simpson SD, Fujita Y et al (2004) Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought responsive Cis-element in the early responsive to dehydration stress promoter. Plant Cell 16:2481–2498
Unlu M, Morgan ME, Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 18(11):2071–2077
Urfer W, Grzegorczyk M, Jung K (2006) Statistics for proteomics: a review of tools for analyzing experimental data. Proteomics 6:48–55
Van den Bergh G, Arckens L (2004) Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics. Curr Opin Biotechnol 15(1):38–43
Waikhom SD, Louis B, Sharma CK, Kumari P, Bharat GS, Singh WM, Talukdar NC (2013) Grappling the high altitude for safe edible bamboo shoots with rich nutritional attributes and escaping cyanogenic toxicity. BioMed Res Int 2013:Article ID 289285
Waikhom SD, Louis B, Pranab R, Wakambam MS, Pardeep KB, Talukdar NC (2014) Scanning electron microscopy of pollen structure throws light on resolving Bambusa Dendrocalamus complex: bamboo flowering evidence. Plant Syst Evol 300(1261):1268
Waikhom SD, Bengyella L, Roy P, Talukdar NC (2015) Insights on predominant edible bamboo shoot proteins. Afr J Biotechnol 14(17):1511–1518
Wang D, Shen S (1987) Bamboos of China. Timber Press 17:1–17
Wang X, Elling AA, Li X, Li N, Peng Z (2009) Genome-wide and organspecific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. Plant Cell 21:1053–1069
Washburn MP, Wolters D, Yates JR (2001) Large scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19:242–247
Wei W, Qi X, Wang L, Zhang Y, Hua W (2011) Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 12:451
Wilkins M (2009) Proteomics data mining. Expert Rev Proteomics 6(6):599–603. doi:10.1586/epr.09.81.
Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. Biotechnology 14:61–65
Wolters DA, Washburn MP, Yates JR (2001) An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73:5683–5690
Wong KM (1995) The bamboos of Peninsular Malaysia. Forest Research Institute Malaysia (FRIM) in collaboration with Forest Research Centre, Forestry Department, Sabah, Malaysia. Malayan Forest Records, No. 41
Wu YJ, Chen HM, Wu TT, Wu JS, Chu RM, Juang RH (2006) Preparation of monoclonal antibody bank against whole water-soluble proteins from rapid-growing bamboo shoots. Proteomics 6:5898–5902
Xiang Z (2010) China's bamboo industry booms for greener economy. China English News, Global Edition, July 18
Xiao Y, Ma J (2012) Fire simulation test and analysis of laminated bamboo frame building. Constr Build Mater 34:257–266

Yang Y, Kanglin W, Shengji P, Jiming H (2004) Bamboo diversity and traditional uses in Yunna, China. Mount Res Dev 24(2):157–165

Yang Q, Duan Z, Wang Z, He K, Sun Q, Peng Z (2008) Bamboo resources, utilization and ex-situ conservation in Xishuangbanna, South-eastern China. J Forest Resour 19(1):79–83

Yates JR (1998) Mass spectrometry and the age of the proteome. J Mass Spectrom 33(1):1–19

Yiping L, Henley G (2010) Biodiversity in bamboo forests: a policy perspective for long term sustainability. International Network for Bamboo and Rattan (INBAR), Working Paper 59

Zakhikhani P, Zahari R, Sultan MTH, Majid DL (2014) Bamboo fibre extraction and its reinforced polymer composite material. Int J Chem Mol Nucl Mater Metall Eng 8(4):322–325

Zhang X-M, Zhao L, Larson-Rabin Z, Li D-Z, Guo Z-H (2012) De Novo sequencing and characterization of the floral transcriptome of Dendrocalamus latiflorus (Poaceae: Bambusoideae). PLoS One 7:e42082

Zhou H, Peng Z, Fei B, Li L, Hu T, Gao Z, Jiang Z (2014) BambooGDB: a bamboo genome database with functional annotation and an analysis platform. Database (Oxford). Vol. 2014

Zhou HL, He SJ, Cao YR, Chen T, Du BX, Chu CC, Zhang JS, Chen SY (2006) sGLU1, A putative membrane-bound endo-1, 4-β-D-glucanase from rice, affects plant internode elongation. Plant Mol Biol 60:137–151

Zhou MB, Yang P, Gao PJ, Tang DQ (2010) Identification of differentially expressed sequence tags in rapidly elongating Phyllostachys pubescens internodes by suppressive subtractive hybridization. Plant Mol Biol Rep 29(1):224–231

Proteomics Driven Research of Abiotic Stress Responses in Crop Plants

Xiuli Hu and Wei Wang

Contents

Abstract Wheat, maize (*Zea mays* L.), and rice are the most grown cereal crop plants worldwide. Abiotic stresses (primarily caused by drought, salinity, high and low temperatures, etc.) negatively affect crop plant growth, development, and eventually production. To understand the response of crop plants to abiotic stresses and the mechanism of stress tolerance, high-throughput proteomics approaches have been used in crop abiotic stress studies and accelerate our understanding of crop stress response at molecular, metabolic, and physiological levels. In this chapter, we outline the primary types of stresses threatening sustainable crop production, review the recent advances in crop stress proteomics, and discuss the limitations and future directions of crop stress proteomics. The knowledge gained from crop stress proteomics is useful for improving crops to cope with various stresses and to meet the food demand of the Earth's growing human population. Finally, the limitations and future development of proteomic approach in crop stress response research are discussed.

Keywords Crop • Proteomics • Abiotic stress • High temperature • Drought

X. Hu • W. Wang (✉)
College of Life Science, Henan Agricultural University, Zhengzhou 450002, China
e-mail: wangwei@henau.edu.cn

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_14

1 Introduction

The world population reached seven billion people in January 2013 and may still increase by 34 % by 2050. With the strong competition for land use, feeding the expected nine billion people will necessarily depend on increasing yields per unit area, rather than increasing crop area (FAO 2009; Gregory and George 2011). Cereal crop production will need to rise to about three billion tons (from the present 2.1 billion) (FAO 2009). However, this may be especially difficult in crops exposed to adverse environmental stresses. To meet the sustainability requirements of producing more with less agriculture area and adverse growth conditions, most challenges thus fall on agricultural practices (agronomy) and genotype improvement (breeding).

Abiotic stresses, such as drought, high temperatures, and salinity, are the major constraints that impair growth and yield of agricultural crops around the world (Hossain et al. 2012). Considerable advances in high-throughput methods of plant molecular and cell biology have enabled scientists to study the molecular events involved in plant response to stress in great detail and on a global scale (Feuillet et al. 2011; Prochnik et al. 2012). However, the molecular analysis of plant response to stress cannot be limited to the transcriptional level, because knowledge of a genomic sequence alone does not indicate how a plant interacts with the environment, and not all open reading frames correspond to a functional gene (Ribeiro et al. 2013). Proteomics approaches are critical for understanding plant mechanisms of stress tolerance.

Plant acclimation to environmental stress is also associated with profound changes in proteome composition. Since proteins are directly involved in plant stress response, the changes that occur in the cells of plants that are subjected to stress ultimately depend on protein synthesis, protein modifications and protein interaction that participate in various metabolic, signaling, biosynthetic, and degradation pathways (Baerenfaller et al. 2012; Hakeem et al. 2012; Han et al. 2014; Hu et al. 2015). In respond to a stress, crops may modulate the abundance of candidate proteins, either by increasing their expression or by synthesizing novel proteins primarily related plant defense system. Proteomics techniques provide one of the best options for the functional analysis of translated regions of the genome can significantly contribute to unravel the possible relationships between protein abundance and plant stress acclimation. Here we review our current knowledge of proteomes from the crop maize (*Zea mays* L.), wheat and *Oryza sativa* under drought and heat stress, in research topics relevant to improve crop production/yields. Based on those achievements, we consider future developments and strategic advances that crop proteomics could take to generate novel insight useful for crop improvement. The differential crop responses against each of these stresses are discussed in detail and are expected to improve crop production/yields.

2 Drought Stress in Maize, Rice, and Wheat

Drought is likely the most important environmental stresses around the world that adversely affects plant growth and development (Yang et al. 2010). The climate changes with rising temperature and increasing population pose serious challenges to crop improvement. It is believed that understanding of how plants respond to drought stress at the protein level are useful in the breeding of tolerant genotypes which would perform well under aridity conditions (Benešová et al. 2012).

2.1 Proteomics Analysis in Maize Response to Drought Stress

C_4 plants have much higher CO_2 assimilation rates than C_3 plants under certain conditions. The specialized differentiation of chloroplasts in mesophyll cell and bundle sheath cell type is unique to C_4 plants and improves photosynthetic efficiency. Maize is the most grown cereal crop in the world (963 million tons in 2014) and model with C_4 photosynthetic machinery, but is very sensitive to drought stress, especially during flowering, pollination and embryo development (Boyer and Westgate 2004). 2DE and high-throughput quantitative proteomics approaches have been employed to investigate the mechanism of maize response to stress (For review, see Zhao et al. 2013). There are a few reports about maize proteomics changes under water deficit. When two maize genotypes with contrasting sensitivity to dehydration were exposed to drought stress, the tolerant genotype maintained open stomata and active photosynthesis while sensitive genotype had a decreased stomata conductance and slightly lower relative water content. Besides, by the proteomics analysis results in leaf response to drought attained by 2D gel electrophoresis and iTRAQ analysis, drought upregulated protective and stress-related proteins (mainly chaperones and dehydrins) in both genotypes. The differences in the levels of various detoxification proteins corresponded well with the observed changes in the activities of antioxidant enzymes. The number and levels of upregulated protective proteins were generally lower in the sensitive genotype, implying a reduced level of proteosynthesis, which was also indicated by specific changes in the components of the translation machinery. These results indicated that the hypersensitive early stomata closure in the sensitive genotype leads to the inhibition of photosynthesis and, subsequently, to a less efficient synthesis of the protective/detoxification proteins that are associated with drought tolerance (Benešová et al. 2012). Specially, Five proteins, including elongation factor 1-delta (eEF1D), hydrolase (alpha/beta fold family), KDE-like protein (cylicin 1), xylanase inhibitor (TAXI-IV), and PsaK (photosystem I reaction center subunit) had higher upregulation in the tolerant genotype/higher downregulation in the sensitive genotype. In contrast, five proteins, including transaldolase 2, ribosomal protein S18, nicotinate phosphoribosyltransferase-like protein, WD-repeat protein, and sugar carrier protein C had higher upregulation in the sensitive genotype/higher downregulation in

the tolerant genotype (Benešová et al. 2012). These results provided a basis for selecting drought-tolerant maize varieties.

Proteomic analyses of growing maize leaves showed that two isoforms of caffeic acid/5-hydroxyferulic 3-*O*-methyltransferase (COMT) accumulated mostly at 10–20 cm from the leaf point under normal conditions while drought resulted in a shift of this region of maximal accumulation toward basal regions. This shift was due to the combined effect of reductions in growth and in total amounts of COMT. Several other enzymes involved in lignin and/or flavonoid synthesis were highly correlated with COMT. These results were useful to understand the impact of drought stress on the maize leaf elongation, and revealed that proteins involved in lignification and flavonoid synthesis have an important contribution to the maize leaf response to drought stress (Vincent et al. 2005).

Taken together, the proteomics studies have provided valuable information on C_4 plants protein components, photosynthesis, and other metabolic mechanisms underlying plants response to drought stress.

2.2 Proteomics Analysis in Wheat Response to Drought Stress

Drought has a great impact on wheat production because water deficit is a common occurrence in grown environments. Therefore, developing drought tolerance wheat varieties are in high demand. The mechanisms involved in the drought response have been extensively studied at the protein level using proteomics over the last few decades. However, only a few researcher studies were carried out by wheat with varieties different drought tolerance, which might be more useful to select drought-tolerant response proteins. Faghani et al. (2015) conducted a comparative proteomic analyses to monitor the stress response of two wheat genotypes with contrasting responses to drought stress. Results showed that drought stress increased the abundance of proteins related to defense and oxidative stress responses such as germin-like proteins, glutathione S-transferase, and superoxide dismutase, and proteins related to protein processing such as small heat shock proteins (sHSPs) in roots of both genotypes in response to drought stress. In addition, the abundance of proteins such as endo-1, 3-beta-glucosidase, peroxidases, S-adenosylmethionine synthase, and malate dehydrogenase was upregulated in roots or leaves of the tolerant genotype and downregulated in that of the sensitive genotype. Overall, proteins related to oxidative stress, protein processing and photosynthesis showed decreased abundance to a greater extent in the sensitive genotype. Irar et al. (2010) compared the embryo proteome from two durum wheat genotypes Mahmoudi (salt and drought sensitive) and Om Rabia3 (salt and drought tolerant). Several proteins belonging to the seed storage family, LEA-type/heat shock proteins, enzyme metabolism and radical scavengers were identified and found to accumulate in different amounts in embryos of tolerant and sensitive wheat varieties. The differential expression pattern could be used as a basis for a biochemical screen of tolerance/sensitivity to drought stress in wheat, and provided a better insight into the molecular responses of wheat plants to drought stress.

Ford et al. (2011) firstly applied shotgun proteomics study in three wheat cultivars, of which Kukri (drought intolerant), Excalibur and RAC875 (both drought tolerant). The two tolerant cultivars differ in their drought tolerance mechanisms: Excalibur shows a higher osmotic adjustment potential, higher stomatal conductance, low ABA content, and rapid recovery after stress compared to RAC875. In contrast, cultivar RAC875 stores more water-soluble carbohydrates in the stem and minimizes water loss due to waxier and thicker leaves (Izanloo et al. 2008). The characteristic leaded to the two drought tolerant varieties differing in their protein responses under drought. Excalibur lacked significant changes in proteins during the initial onset of the water deficit in contrast to RAC875 that had a large number of significant changes. All three cultivars had changes consistent with an increase in oxidative stress metabolism and reactive O_2 species (ROS) scavenging capacity with increases in antioxidant enzymes as well as ROS avoidance through the decreases in proteins involve in photosynthesis and the Calvin cycle. The findings from this proteomic study support the physiological and yield data (Izanloo et al. 2008) in three wheat cultivars response to drought stress. This highlights the importance of proteomics as a complementary tool for identifying candidate genes in abiotic stress tolerance in cereals. This study has provided potential candidates for genetic manipulation of wheat cultivars to enhance drought tolerance (Ford et al. 2011).

2.3 Proteomics Analysis in Rice Response to Drought Stress

Rice (*Oryza sativa* L.) is the primary source of food for above half of the world's population and is grown in highly diverse situations that range from flooded wetland to rain-fed dryland (Degenkolbe et al. 2009). Irrigated rice which accounts for 55 % of the world rice area provides 75 % of global rice production. The present and future food security of Asia depends largely on the irrigated rice production system: more than 75 % of the rice supply comes from 79 million ha of irrigated land (Bhuiyan 1992; Tuong et al. 2004). Water deficit is therefore a key constraint that affects rice production in different countries. Our challenge is to develop drought-tolerant rice varieties to counteract drought stress in the face of declining water availability.

Proteomics analysis dealing with rice response to drought stress to screen some proteins related to endurance drought stress were studied in two rice genotypes with contrasting susceptibility to drought stress at reproductive stage. Drought susceptible rice cultivar Zhenshan97B and tolerant rice cultivar IRAT109 had an osmotic potential of leaves reduced 78 % and 8 % after 20 day of drought treatment, respectively. Two-dimension gel analysis of proteins extracted from flag leaves showed that the expression of glycine dehydrogenase, orthophosphate dikinase, ribulose bisphosphate carboxylase (Rubisco), glycine hydroxymethyltransferase and ATP synthase was downregulated in Zhenshan97B response to drought stress, suggesting the reduction of capacity of carbon assimilation in this rice cultivar; In IRAT109 response to drought stress, transketolase, Rubisco were downregulated; however, Rubisco

activase and peptidyl-prolyl cis–trans isomerase, which might alleviate the damage on Rubisco by drought stress, were upregulated. The increased abundances of chloroplastic superoxide dismutase and dehydroascorbate reductase might provide antioxidant protection for IRAT109 against damage by drought stress (Ji et al. 2012).

Mirzaei et al. (2012) describe patterns of protein expression of rice seedlings exposed to moderate drought, extreme drought, and combined drought/recovery treatments. The analysis of label-free quantitative shotgun proteomics indicated that more proteins among identified proteins were downregulated in early stages of drought but more were upregulated by severe drought. After rewatering, these proteins were significantly downregulated, suggesting that stress-related proteins were being degraded. Besides, proteins involved in signaling and transport pathway became dominant under severe drought stress but decreased again after rewatering. Most of the nine aquaporins identified were responsive to drought, with six decreasing rapidly in abundance as plants were rewatered. Nine G-proteins appeared in large amounts during severe drought and dramatically degraded once plants were rewatered. Heat shock cognate 70 protein (HSP70 family) was highly upregulated under moderate drought but downregulated in extreme drought. However, Komatsu and Zang (2007) also observed that levels of a heat shock protein and a dnaK-type molecular chaperone ((HSP70 family) were reduced under drought stresses by the results of two-dimensional polyacrylamide gel electrophoresis. These results indicated that mechanisms of plant drought tolerance are complex, interacting, and polygenic.

3 Heat Stress in Maize, Wheat, and Rice

Heat stress is one of the most common abiotic stresses for many crops worldwide and reduces the yield and quality of crops. The average global temperature increase was approximately 1 °C in the last decade (Rasul et al. 2011). Increased climate variability and higher average temperatures are expected in many regions of world, which will cause to more frequent extreme high temperatures events. Field evidence increasingly shows that heat stress in flowering can have large negative impacts on cereal grain yields (Rattalino Edreira et al. 2011). Consideration of canopy temperature is suggested as a promising approach to concurrently account for heat and drought stress, which are likely to occur simultaneously under field conditions (For review, see Rezaei et al. 2015). Thus, there is a need to screen heat endurance crops for increasing crop productivity with increasing temperatures worldwide.

3.1 Proteomics Analysis in Maize Response to Heat Stress

Many studies have reported on the significant reduction in maize production due to the extreme weather event (combined heat wave and drought) that occurred in the USA in 2012 (Chung et al. 2014) and China in 2009, 2010, and 2014. However,

most of these studies focused on yield and a few assessed the potential effect of weather extremes on proteomics. Heat stresses usually lead to protein dysfunction. It is especially important for plant survival under heat stress to maintain proteins in their functional conformations and prevent the aggregation of nonnative proteins.

2-DE gel analysis of proteins in maize leafs under drought, heat, and combined both stresses indicated cytochrome b6-f complex iron-sulfur subunit protein, G protein, uncharacterized protein (B4G072) and three sHSPs: sHSP17.4, sHSP17.2, and sHSP26 were upregulated by heat and combined drought and heat stress; granule-bound starch synthase IIa were downregulated by heat and combined drought and heat stress (Hu et al. 2010). Ristic et al. (1999) also found that some HSPS were upregulated by heat and combined drought and heat stress in leaves of drought and heat resistant maize line. In addition, by the results of 2-DE-based proteomics, RNA interference (RNAi), co-immunoprecipitation (Co-IP) four chloroplast proteins, including ATP synthase subunit β, chlorophyll a–b binding protein, oxygen-evolving enhancer protein 1, and photosystem I reaction center subunit IV, strongly interacted with sHSP26 under heat stress, and the suppression of sHSP26 expression significantly reduced the O_2 evolution rate of photosystem II under heat stress (Hu et al. 2015). Overall, these findings demonstrate the relevance of sHSPs in protecting maize endurance to heat stress.

In maize seedlings response to heat and combined heat stress, enolase 2, *O*-methyltransferase, alcohol dehydrogenase 1, fructokinase-1, caffeoyl-CoA *O*-methyltransferase 1, APX1—Cytosolic Ascorbate Peroxidase, aquaporin PIP2-5, glutathione S-transferase 4, serine/threonine-protein kinase receptor, pathogenesis-related protein 10, VAP27-2, DNA polymerase, nucleoside diphosphate kinase, glycine-rich RNA-binding protein 2, and pathogenesis-related protein 10 had significant difference in expression revealed by 2-D gel analysis results of root proteins and were significantly upregulated by heat and combined drought and heat stress except glutathione S-transferase 4. Nevertheless, the increase of most proteins by heat stress was lower than that by combined drought and heat stress (Liu et al. 2013). The results of this study suggest that plants cope with combined drought and heat stress in a complex manner, where sHSPs play a pivotal role in this complex cellular network. Although this study is an initial proteomic investigation into the maize response to heat stress and combined drought and heat stress, this kind of study provides a good starting point in understanding maize responses to combined stress. However, further proteomics analyses should be conducted to gain a better understanding of the overall responses of plants to heat and combined drought and heat stress.

3.2 Proteomics Analysis in Wheat Response to Heat Stress

Wheat, one of the most important crops, is sensitive to heat stress, which is considered to be one of the major limiting factors for wheat production in Europe (Semenov and Shewry 2011). The optimum temperature for wheat during grain filling is

around 31 °C, and higher temperatures have been shown to significantly decrease grain yield (Zhang et al. 2013; Wang et al. 2015; for review, see Rezaei et al. 2015). Wheat plants subjected to high temperature episodes at spikelet initiation, anthesis or both stages often cause grain yield loss. Thus, improvement of wheat heat tolerance under more frequent heat stress conditions will become essential for food supply. To meet this challenge, the integration of wheat genomics, transcriptomics, and proteomics with rapidly evolving bioinformatics tools and interactive databases is required.

Previous studies carried on the proteome analysis of leaves in two wheat genotypes-tolerant cultivar '810' and the sensitive cultivar '1039' (Wang et al. 2015). Proteins related to photosynthesis, glycolysis, stress defense, heat shock and ATP production were differently expressed in leaves of the tolerant and sensitive cultivar under heat stress compared to control. For the tolerant cultivar '810', heat stress increased the abundance of proteins including BRI1-KD interacting protein 114, Rubisco activase, oxygen evolving enhancer protein 1, glyceraldehyde-3-phosphate dehydrogenase, Bp2A protein, isocitrate dehydrogenase, 2-Cys peroxiredoxin BAS1 and glycine decarboxylase P subunit, chloroplast protease, peptidyl-prolyl-cis-transisomerase and, 60 kDa chaperonin subunit CPN60, HSP70 while decreased the abundance of NADP-isocitrate dehydrogenase, Glutamine synthetase. For the sensitive cultivar '1039', heat stress increased the abundance of proteins including Rubisco activase, sedoheptulose bisphosphatase, CPN60, peroxiredoxin IIE-2, ascorbate peroxidase, dihydrolipoyl dehydrogenase 1 and glycine decarboxylase P subunit while decreased the abundance of proteins including HSP70, ATP-dependent zinc metalloprotease FtsH2, ATP synthase CF1 subunit, ER molecular chaperone, peroxiredoxin IIE-2, ascorbate peroxidase, dihydrolipoyl dehydrogenase 1 and glycine decarboxylase P subunit. Collectively, the results indicate that primarily changes in both the amount and activities of enzymes involved in photosynthesis, antioxidant activities and HSPs in leaves contributed to higher heat tolerance in the cultivar '810' compared to the heat sensitive cultivar '1039' (Wang et al. 2015).

The proteomic approach is well suited to examining the effects of heat stress on grain development and further elucidation of those aspects of protein composition that account for the unique dough-forming properties of wheat flour (For review, see Skylas et al. 2005). This strategy was pursued by analysis of wheat grain proteome under heat stress (Majoul et al. 2004). Majoul et al. (2004) observed that heat-stress during grain filling decreased the amount of several heat stress proteins including granule-bound starch synthase, glucose-1-phosphate adenyltransferase, beta-amylase, ATP synthase beta-chain. These several proteins were related metabolic pathways, starch synthesis, carbohydrate metabolism and energy metabolism. In contrast, heat stress increased the amount of five sHSPs, HSP82, elongation factors and eucaryotic translation initiation factors. The proteomic analysis was carried in immature endosperm from grains of two wheat varieties differing in heat tolerance: heat-tolerant cultivar Fang and the susceptible cultivar Wyuna. The results showed that the heat-tolerant cultivar Fang exhibited a stronger and more diverse heat-shock response than the susceptible cultivar Wyuna in terms of the changes in protein composition (Skylas et al. 2002). Seven sHSPs (16–17 kDa) were expressed in

heat-shocked Fang but not in heat-shocked Wyuna. Collectively, these sHSPs may be putative marker proteins for heat-tolerance, offering the prospect of assisting breeders in the selection of heat-tolerant cultivars.

3.3 Proteomics Analysis in Rice Response to Heat Stress

Heat stress impediment in developing stage of rice has been occurred due to the impact of global warming (Ainsworth and Ort 2010). The production of rice is known to be sensitive to increasing environmental temperature (Peng et al. 2004), and the current temperatures of rice grain filling are already approaching critical levels in many countries (For review, see Mitsui et al. 2013). Global warming inevitably affects the grain yields of rice. Furthermore, grain quality should be more susceptible to the heat stress compared with the grain yield. Grain chalking caused by heat stress during ripening stage is one of the major problems in the field of agriculture. Recent proteomic analyses revealed dynamic changes of metabolisms during rice grain development under heat stress. Some excellent reviews discuss the effects of heat stress on rice grain development (Mitsui et al. 2013; Zou et al. 2011)

To understand the responses of rice seedlings to different high-temperature stresses, 7-day-old rice seedlings were exposed to different heat stress (35, 40 and 45 °C) for 48 h. The identified proteins were sorted into nine functional groups. The two most abundant groups were photosynthesis/photorespiration-related proteins and HSPs. RuBisCO large subunit and glyceraldehyde-3-phosphate dehydrogenase, oxygen-evolving complex proteins were downregulated at each heat stress. RuBisCO activase precursor, photorespiration-related proteins, chloroplast glutamine synthetase precursor, glycine dehydrogenase, and glycine cleavage system H protein were upregulated at each heat stress. Among the identified HSPs, Chaperonin 60 and chaperonin 21 can bind to the RuBisCO large subunit and help the RuBisCO subunits assemble into a holoenzyme (Spreitzer 1999; Roy 1989). Two Heat shock cognate 70 kDa proteins and three sHSPs were specifically upregulated by 45 °C. Antistress proteins: dehydroascorbate reductase was upregulated by each high temperature. An isoform of glyoxalase I and a salt-induced protein were specifically upregulated by 45 °C while another isoform of glyoxalase I was induced by each high temperature (Han et al. 2009). Besides, Lee et al. (2007) also observed that heat stress increased the expression of HSP70, dnak-type molecular chaperone, Bip chaperonin 60 beta, putative HSP (XP_468773), ATP synthesis CF1 alpha chain, proteasome subunit alpha type, alpha 1 subunit of 20S proteasome, putative glycine dehydrogenase, Glutamate dehydrogenase, pyruvate dehydrogenase E1 alpha subunit, thiamine biosynthesis protein and 17.7 kDa, 16.9 kDa, 22 kDa three sHSPs, while downregulated Rubisco small subunit. Liao and Huang (2011) proved that heat stress also upregulated expression of the 18 kDa HSP, 17.9 kDa HSP and class 1 HSP by the proteome analysis of young rice Caryopsis. This increasing number of the upregulated, protection-related proteins indicated that more protective processes were involved in heat stress. All the results indicated that different strategies were adopted at different levels of high temperature.

4 Conclusions and Future Perspective

Crop agricultural production is severely limited by various abiotic and biotic stress factors. Proteomics results proved that drought and heat stress induce profound alterations in protein network mainly involved in signaling, protein metabolism, energy metabolism, storage proteins, lignin metabolism, and in transport proteins, chaperone proteins, protective proteins as well as proteins affecting regulation of plant growth and development. Stress-induced proteins changes revealed significant differences in their relative abundance or posttranslational modifications between two genotypes with contrasting susceptibility to stress. More attention should be paid to these key proteins and their potential role in underlying plant tolerance to a given stress which can be used as protein biomarkers of a given stress. In particular, the techniques of non gel-based quantitative proteomic method such as the iTRAQ system were adopted in recent several years which allowed relatively minor changes in the composition of a large number of proteins to be indentified and be accurately quantified within broad dynamic ranges of protein abundance, so these approaches will contribute to a detailed protein functional characterization which will surely help us to better understand the processes of plant stress acclimation and stress tolerance acquisition. It may also be possible to identify marker proteins to select for tolerance to stress. Nevertheless, such studies require validation of the protein changes by transcriptomic and biochemical (enzyme assays) studies before the transgenic experiments are initiated.

References

Ainsworth EA, Ort DR (2010) How do we improve crop production in a warming world? Plant Physiol 154:526–530

Baerenfaller K, Massonnet C, Walsh S, Baginsky S, Buhlmann P, Hennig L et al (2012) Systems-based analysis of Arabidopsis leaf growth reveals adaptation to water deficit. Mol Syst Biol 8:606

Benešová M, Holá D, Fischer L, Jedelský PL, Hnilička F et al (2012) The physiology and proteomics of drought tolerance in maize: Early stomatal closure as a cause of lower tolerance to short-term dehydration? PLoS One 7(6), e38017

Bhuiyan SI (1992) Water management in relation to crop production: Case study on rice. Outlook Agric 21:293–299

Boyer J, Westgate M (2004) Grain yields with limited water. J Exp Bot 55(407):2385–2394

Chung U, Gbegbelegbe S, Shiferaw B, Robertson R, Yun JI, Tesfaye K, Gerrit Hoogenboom G, Sonder K (2014) Modeling the effect of a heat wave on maize production in the USA and its implications on food security in the developing world. Weather Clim Extremes 5:67–77

Degenkolbe T, Do PT, Zuther E, Repsilber D, Walther D, Hincha DK, Köhl KI (2009) Expression profiling of rice cultivars differing in their tolerance to long-term drought stress. Plant Mol Biol 69:133–153

Faghani E, Javad G, Komatsu S, Mirzaei M, Khavarinejad RA, Najafi F, Farsad LK, Salekdeh GH (2015) Comparative physiology and proteomic analysis of two wheat genotypes contrasting in drought tolerance. J Proteomics 114:1–15

FAO (2009) How to feed the world in 2050. http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf
Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88
Ford KL, Cassin A, Bacic A (2011) Quantitative proteomic analysis of wheat cultivars with differing drought stress olerance. Front Plant Sci 2:1–11
Gregory PJ, George TS (2011) Feeding nine billion: the challenge to sustainable crop production. J Exp Bot 62:5233–5239
Hakeem KR, Chandna R, Ahmad P, Ozturk M, Iqbal M (2012) Relevance of proteomic investigations in plant stress physiology. OMICS 16(11):621–635
Han F, Chen H, Li XJ, Yang MF, Liu GS, Shen SH (2009) A comparative proteomic analysis of rice seedlings under various high-temperature stresses. Biochim Biophys Acta 1794(11):1625–1634
Han C, Yang PF, Sakata K, Komatsu S (2014) Quantitative proteomics reveals the role of protein phosphorylation in rice embryos during early stages of germination. J Proteome Res 13:1766–1782
Hossain Z, Nouri MZ, Komatsu S (2012) Plant cell organelle proteomics in response to abiotic stress. J Proteome Res 11:37–48
Hu XL, Li YH, Li CH, Yang HR, Wang W, Lu MH (2010) Characterization of small heat shock proteins associated with maize tolerance to combined drought and heat stress. J Plant Growth Regul 29:455–464
Hu XL, Yang YF, Gong FP, Zhang DY, Zhang L, Wu LJ, Li CH, Wang W (2015) Protein sHSP26 improves chloroplast performance under heat stress by interacting with specific chloroplast proteins in maize (Zea mays). J Proteomics 115:81–92
Irar S, Brini F, GodayA MK, Pagès M (2010) Proteomic analysis of wheat embryos with 2-DE and liquid-phase chromatography (ProteomeLab PF-2D)–A wider perspective of the proteome. J Proteomics 73(9):1707–1721
Izanloo A, Condon AG, Langridge P, Tester M, Schnurbusch T (2008) Different mechanisms of adaptation to cyclic water stress in two South Australian bread wheat cultivars. J Exp Bot 59:3327–3346
Ji K, Wang Y, Sun W, Lou Q, Mei H, Shen S, Chen H (2012) Drought-responsive mechanisms in rice genotypes with contrasting drought tolerance during reproductive stage. J Plant Physiol 169(4):336–344
Komatsu S, Zang X (2007) A proteomics approach for identifying osmotic-stress-related proteins in rice. Phytochemistry 68:426–437
Lee DG, Ahsan N, Lee SH, Kang KY, Bahk JD, Lee IJ et al (2007) A proteomic approach in analyzing heat-responsive proteins in rice leaves. Proteomics 7:3369–3383
Liao JL, Huang YJ (2011) Evaluation of protocols used in 2-D electrophoresis for proteome analysis of young rice caryopsis. Genomics Proteomics Bioinformatics 9(6):229–237
Liu TX, Zhang L, Yuan ZL, Hu XL, Lu MH, Wang W, Wang Y (2013) Identification of proteins regulated by ABA in response to combined drought and heat stress in maize roots. Acta Physiol Plant 35(2):501–513
Majoul T, Bancel E, Triboï E, Ben Hamida J, Branlard G (2004) Proteomic analysis of the effect of heat stress on hexaploid wheat grain: characterization of heat-responsive proteins from non-prolamins fraction. Proteomics 4(2):505–513
Mirzaei M, Pascovici D, Atwell BJ, Haynes PA (2012) Differential regulation of aquaporins, small GTPases and V-ATPases proteins in rice leaves subjected to drought stress and recovery. Proteomics 12:864–877
Mitsui T, Shiraya T, Kaneko K, Wada K (2013) Proteomics of rice grain under high temperature stress. Front Plant Sci 4:36
Peng S, Huang J, Sheehy JE, Laza RC, Visperas RM, Zhong X et al (2004) Rice yields decline with higher night temperature from global warming. Proc Natl Acad Sci U S A 101:9971–9975
Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M et al (2012) The cassava genome: current progress, future directions. Trop Plant Biol 5:88–94

Zhao Q, Chen SX, Dai SJ (2013) C4 photosynthetic machinery: insights from maize chloroplast proteomics. Front Plant Sci 4:1–5
Rasul G, Chaudhry QZ, Mahmood A, Hyder W (2011) Effect of temperature rise on crop growth and productivity. Pak J Meteorol 8:53–62
Rattalino Edreira JI, Budakli Carpici E, Sammarro D, Otegui ME (2011) Heat stress effects around flowering on kernel set of temperate and tropical maize hybrids. Field Crop Res 123(2):62–73
Rezaei EE, Webber H, Gaiser T, Naab J, Ewert F (2015) Heat stress in cereals: Mechanisms and modeling. Eur J Agron 64:98–113
Ribeiro M, Nunes-Miranda JD, Branlard G, Carrillo JM, Rodriguez-Quijano M, Igrejas G (2013) One hundred years of grain omics: identifying the glutens that feed the world. J Proteome Res 12:4702–4716
Ristic Z, Yang GP, Bhadula SK (1999) Two-dimensional gel analysis of 45 ku heat shock proteins from a drought and heat resistant maize line. J Plant Physiol 154(2):264–268
Roy H (1989) Rubisco assembly: a model system for studying the mechanism of chaperonin action. Plant Cell 1:1035–1042
Semenov MA, Shewry PR (2011) Modeling predicts that heat stress, not drought, will increase vulnerability of wheat. Eur Sci Rep 1:66
Skylas DJ, Cordwell SJ, Hains PG, Larsen MR, Basseal DJ, Walsh BJ, Blumenthal C, Rathmell WG, Copeland L, Wrigley CW (2002) Heat shock of wheat during grain filling: characterization of proteins associated with heat-tolerance using a proteome approach. J Cereal Sci 35:175–188
Skylas DJ, van Dyk D, Wrigley CW (2005) Proteomics of wheat grain. J Cereal Sci 41:165–179
Spreitzer RJ (1999) Questions about the complexity of chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase. Photosynth Res 60:29–42
Tuong TP, Bouman BAM, Mortimer M (2004) More rice, less water—integrated approaches for increasing water productivity in irrigated rice-based systems in Asia 2004 12th AAC, 4th ICSC Symposia papers
Vincent D, Lapierre C, Pollet B, Cornic G, Negroni L, Zivy M (2005) Water deficits affect caffeate *O*-methyltransferase, lignification, and related enzymes in maize leaves. A proteomic investigation. Plant Physiol 137:949–960
Wang X, Dinler BS, Vignjevic M, Jacobsen S, Wollenweber B (2015) Physiological and proteome studies of responses to heat stress during grain filling in contrasting wheat cultivars. Plant Sci 230:33–50
Yang S, Vanderbeld B, Wan J, Huang Y (2010) Narrowing down the targets: towards successful genetic engineering of drought-tolerant crops. Mol Plant 3(3):469–490
Zhang XX, Cai J, Wollenweber B, Liu FL, Dai TB, Cao WX, Jiang D (2013) Multiple heat and drought events affect grain yield and accumulations of high molecular weight glutenin subunits and glutenin macropolymers in wheat. J Cereal Sci 57(1):134–140
Zou J, Liu CF, Chen XB (2011) Proteomics of rice in response to heat stress and advances in genetic engineering for heat tolerance in rice. Plant Cell Rep 30:2155–2165

Proteomics in Sex Determination of Dioecious Plants

Erhui Xiong, Xiaolin Wu, Le Yang, and Wei Wang

Contents

Abstract Sex determination in plants is controlled by complex endogenous genetic programs and responses to environmental cues. The great majority of the world's flowering plants are hermaphrodite, while approximately 6 % are dioecious, having male and female flowers in individuals. Male and female plants usually have different economic value. However, most dioecious plants do not exhibit discernible sexual dimorphism before sexual maturity. Therefore, sex identification of dioecious plants at juvenile stage would greatly benefit breeding program. In this chapter, we outline the primary methods, especially proteomics, in sex determination

E. Xiong • X. Wu • L. Yang • W. Wang (✉)
College of Life Science, Henan Agricultural University, Zhengzhou 450002, China
e-mail: wangwei@henau.edu.cn

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_15

of dioecious plants, and summarize the current applications of proteomics in sex determination of plants, finally discuss the limitations and future of proteomics in sex determination of dioecious plants.

Keywords Dioecious plant • Proteomics • Sex determination

1 Introduction

The great majority of flowering plants are hermaphrodite, while approximately 6 % or 14,600 species in 960 genera and 200 families are dioecious, having male and female flowers in individuals (Ming et al 2007). So far, only a limited number of dioecious plants have evolved sex chromosomes, such as XY system in *Silene latifolia*, XA balance system in *Humulus lupulus* L., ZW system in *Fragaria* × *ananassa Duch* (Negrutiu et al. 2001; Vyskot and Hobza 2004). Recently, Akagi et al. (2014) investigated sex determinants in the Caucasian persimmon (X:Y system) and revealed a Y-specific sex-determinant candidate (OGI) displaying male-specific conservation among *Diospyros* species. A range of sex-linked molecular markers has also been summarized in a botanical briefing (Ainsworth 2000). In dioecious plants, despite increasing research efforts on a number of different plant species, there is relatively little information available on the molecular basis of sex determination.

Unlike animals, most dioecious plants do not exhibit discernible sexual dimorphism before sexual maturity and always accompany a longer growing period, such as nursery-grown *Pistacia chinensis* with the first flowering at 6–10 year-old. Male and female plants usually have different economic value. For example, *Pistacia chinesis,* as a biodiesel feedstock, the female plants have higher economic values than male plants because of high seed oil content (35–50 %) (Wang and Liu 2011); *Simmondsia chinensis* (Agrawal et al. 2007; Sharma et al. 2008) and *Hippophae rhamnoides* (Korekar et al. 2012), the female have commercial value. In addition, dioecious plants are ideal systems for the research on the evolutionary, developmental and molecular processes of sex determination. Therefore, a reliable method for sex determination in dioecious plants cultivated for fruit or seed at juvenile stage would greatly benefit breeding programs.

With the advent of post-genomic era, proteomics is becoming more and more popular. As a powerful tool for protein identification and gene function analysis, proteomics techniques have been widely used in plant science. Compared with the gene technology, it is a more direct understanding of function and regulation, and especially 2-DE-based proteome analysis was suitable for paired comparison of dioecious plants (Xiong et al. 2013).

In this chapter, we outline the primary methods, especially proteomics, in sex determination of dioecious plants, and summarize the current applications of proteomics in sex determination of plants, finally discuss the limitations and future of proteomics in sex determination of dioecious plants.

2 Sex Determination of Dioecious Plants

2.1 Dioecious Plants

Dioecious plants (dioecy) refer to a minority of species which has male and female individual organisms with unisexual flowers, which indicates these plants are incapable of self-pollination (Ainsworth 2000). Dioecy appears at a low frequency and in a scattered taxonomic distribution (Charlesworth 1985; Renner and Ricklefs 1995). In some cases, the recent origins of dioecy from a cosexual state have been documented (Darwin 1877; Westergaard 1958). In plants, only 6 % of known angiosperm species are dioecious (Ming et al. 2007), e.g., *Asparagus officinalis* L. (Bracale et al 1990), *Actinidia chinensis*, *H. lupulus* L. (Shephard et al. 1999) and *Carica papaya* L. Jaiswal et al. (1984).

2.2 Sex Determination in Dioecious Plants

The economic values of male and female dioecious plants are usually different. Sex identification of dioecious plants has an important significance. To date, the mechanism of sex determination in dioecious plants has two main types: chromosomal sex determination theory and sex determination genes theory (Ming et al. 2011).

2.2.1 Chromosome Sex Determination Theory in Dioecious Plants

Sex chromosome was first detected in locust by McClung (1902). Subsequently, plants sex chromosomes were also reported in the *Sphaerocarpos* (Allen 1917) and in flowering plants of *Elodea gigantea* (Santos 1923), *Rumex acetosa* (Kihara and Ono 1923), and *Silene dioica* (*Melandrium rubrum*) (Blackburn 1923). So far, the sex chromosomes were found in more than 40 species of higher plants (Ming et al. 2011; Aryal and Ming 2013). There are three main types of chromosome sex determination in dioecious plants, viz., active Y chromosome (XY system), X chromosome to autosomes ratio (X:A balance system) and ZZ/ZW (ZW system) (Kumar et al. 2014; Heikrujam et al. 2014).

XY System

XY system is the most familiar in dioecious plants. The decision of most female gene were located on X chromosome, and female plants are homogametic having two identical X chromosomes, while Y chromosome plays an important role of sex determination in male and males are heterogametic which contains an X and a dominant Y chromosome (Westergaard 1948). The sex determination mechanism of XY system in dioecious plants is similar to that in mammals (Westergaard 1958).

X and Y chromosomes are obviously different in size. X is smaller than Y, except in *Humulus lupulus* and *Viscum* (Parker 1990), and they are both larger than autosomal chromosomes (Ciupercescu et al. 1990; Matsunaga et al. 1994; Matsunaga and Kawano 2001). Y chromosome contains four regions: homology female region of inhibition, male promoter, male fertility zone and homologous region of X chromosome. If inhibition regions of Y chromosome is missing, it will produce the perfect flowers; promoter regions missing, the original male plants into female plant; fertility region missing, it will form a male sterile plant (Farbos et al. 1999; Matsunaga and Kawano 2001).

In addition, another digametic type of male plants has a single X chromosome and no Y chromosome (Jaquiéry et al. 2012; Kumar et al. 2014). These male plants contain one less sex chromosome than that in female plants. The form of sex chromosome in male plants is XO and in female plants is XX. This type of dioecious plants has been found to include only three species, i.e., *Zanthoxylum piperitum* L., *Vallisneria natans (Lour.) Hara*, and *Dioscorea sinuata* L.

X:A Balance System

Bridges (1925) discovered that in *Drosophila* the sex was determined by the balance between the female determining gene in the X chromosome and the male sex determining gene in the autosomes. The genders of some dioecious plants are not decided by sex chromosomes, but by the ratio between X chromosome and autosomes, i.e., X:A balance System, as in *Rumex acetosa* (Ainsworth 2000), *Humulus japonicus*, *Humulus lupulus* (Shephard et al. 1999, 2000), and *Phoenix dactylifera* (Siljak-Yakovlev et al. 1996); a 1:1 ratio of X:A results in female, a 1:2 ratio results in male, and the intermediate ratio results in intersex plants (Ainsworth et al. 1999).

In addition, sex determination in the dioecious *Melandrium* is controlled by both the X:Y chromosome system and the X:A balance system, and there exists a sex balance in the offspring (Warmke 1946; Westergaard 1948).

ZW System

Different from the type of XY system, ZW system female plants contains a Z and a dominant W chromosome (ZW) while male plants are homogametic (ZZ). To date, female heterogamety (ZW) in ZW system has been observed in *Fragaria* × *ananassa Duch* (Correns 1928), the female chromosome composition is ZW+40A and the male is ZZ+40A. In addition, The type of ZW sex determination system has also been suspected in *Salix viminalis, Silene otites* (Sansone 1938), *Populus spp.* (Yin et al. 2008; Pakull et al. 2011), and *Myrstica fragrans* (Flach 1966).

2.2.2 Sex Determination Gene Theory of Dioecious Plants

Plant traits are controlled by genes, and sex is no exception. In some dioecious plants without obvious sex chromosomes, sex is still controlled by genes; it may be due to a single locus or multiple loci either unlinked or tightly linked on autosomes (Grant et al. 1994).

Asparagus officinalis L. is a member of *Liliaceae*. Female and male plants have male and female vestigial organs, respectively. These rudimentary organs generally are not functional; however, sometimes in the male plant vigorous seeds can be harvested, and the offspring ratio of male and female plants is close to 3:1. Rick and Hanna (1943) had put forward and proved that in *Asparagus officinalis* L. the sex was controlled by a single gene, male gene combination was MM or Mm, female gene combination was mm, and M is dominant to m. In *Ecballium elaterium,* the sex was controlled by three allele genes (Sondur et al. 1996). *Mercurialis annua* was controlled by multiple unlinked loci (Janousek and Mrackova 2010; Louis 1989).

Sex is also related to plant hormones, such as auxins, cytokinins, and abscisic acid, which have been shown to be responsible for regulation of sex determination in flowering plants (Dellaporta and Calderon-Urrea 1993; Dauphin-Guerin et al. 1980; Irish and Nelson 1989). Short sunshines can promote female sex (*Cannabis sativa* and *Humulus lupulus*), low temperatures can promote female sex, and high temperatures can promote male sex (*Cannabis sativa* and *Spinacia oleracea*) (Zhao and Liu 1991).

3 Methods for Sex Determination of Dioecious Plants

3.1 Morphological Sex Determination in Dioecious Plants

Genetic diversity must be reflected in the external morphology. Morphological is a prevailing method for sex determination of dioecious plants. Bugala (1951) first identified the sex of *Populus tremula* through leaf color. Then, there are many studies of sex identification based on leaves, stems, branches, inflorescences, flowers, canopies, etc. (Dzhaparidze 1969; Geber et al. 1999), to prove the difference between male and female plants. For example in sea buckthorn, leaves of female plants are narrower than that of male plants (Li et al. 2007); in grape, the pollen morphology of female plants are irregular concave, dry and collapse, while in male plants they are elliptic and larger than that in female plants (Zhang et al. 1998). These differences can be used for determine the sex of male and female plants at vegetative stage.

As a preliminary identification method of sex determination in dioecious plants, the morphological characteristics are simple and effective. However, the morphological characteristics of the plant are limited by the developmental stage and easily

affected by external environment, Furthermore, in the majority of dioecious plants such as *Pistacia chinesis*, they are no obvious differences between male and female before the sexual organs mature. Therefore, the method of external morphological characteristics cannot be used as a reliable basis for sex identification at juvenile stage.

3.2 Physiological Sex Determination in Dioecious Plants

Numerous studies have shown that female and male plants have significant differences at physiological level, such as photosynthetic activity, respiration rate, transpiration rate, water efficiency, and phenolic contents (Dzhaparidze 1969; Geber et al. 1999). As early as in 1984, the acidic and alkaline phosphatase activity was investigated in *Carica papaya* L. vegetative tissues and reproductive tissues, and found that the enzyme activity in male plants is higher than that in female plants (Jaiswal et al. 1984). Dawson (1993) found that *Acer negundo* L. has sexual dimorphism in the growth and biomass allocation.

The same indicators in different species have different results. In *Diospyros lotus,* the male plants of Eh were higher than female plants, and this result is consistent with the view of Manoilov (1922), but get the opposite result of *Ginkgo biloba* and *Eucommia ulmoides* Oliv (Zhao and Liu 1991).

Compared with the method of morphology, this method can improve the accuracy of sex determination in dioecious plants. But these studies are most used of male and female plants after sexual maturity, and these indicators are easily affected by environmental factors and different periods of plant growth. The method of physiology can provide some reference in the identification of plant sex, which still need to improve.

3.3 Isozyme for Sex Determination in Dioecious Plants

Isozyme technique is a very powerful tool in life science research, widely used in various fields of agriculture, medicine, and biology. The sex of a plant is mainly regulated by genes. Isozymes are the direct result of gene expression in plants. Peroxidase and polyphenol oxidase, which exist widely in higher plants, play an important role in growth and development of plants (Perice and Brewbaker 1973). The differences in isozymes can be used for sex identification of dioecious plants.

As early as in 1972, Penel found that the differences of *Spinacia oleracea* gender associated with the peroxidase isozyme. Then, in *Mercurialis annua*, specific peroxidase was found in the early stage flower organogenesis of male plants (Kahlem 1976). The activity of peroxidase in different species was different. Truta et al. (2002) showed that the activity of peroxidase was greater in male plants than female plants of *Cannabis sativa* L.; however, the activity of peroxidase was greater in

female plants of *Phoenix dactylifera* L.(Qacif et al. 2007; Bekheet et al. 2008). In addition, others isozymes were also exploited, such as isocitrate dehydrogenase pattern, glutamate oxaloacetate, esterase, acid phosphatase, malate dehydrogenase, and catalase (Maestri et al. 1991; Bekheet et al. 2008; Sharma et al. 2010).

According to the theory of molecular enzymology "one gene–one isozyme subunit," isozymes are indicators in molecular level, and therefore, compared to the method of external morphology and physiology, the isozyme method can provide more reliable scientific basis for identification of plant sex. However, this method is limited to the difference between varieties, isozyme tissue-specificity, developmental stage, habitat, etc. Thus, the isozyme technique used in plant sex determination also requires in-depth research.

3.4 Markers Based on DNA Fingerprinting for Sex Determination in Dioecious Plants

Application of molecular markers provides a powerful means for the study of plant sex determining mechanism. At present, the molecular marks of male and female plant gender differences is mainly DNA fingerprinting markers, such as random amplified polymorphic DNA (RAPD), sequence-characterized amplified region (SCAR), amplified fragments length polymorphism (AFLP), etc.

3.4.1 Random Amplified Polymorphic DNA (RAPD)

Random amplified polymorphic DNA (RAPD), a kind of molecular technology based on PCR, can analyze the entire sequence of an unknown genome, introduced by Welsh and McClelland (1990) and Williams et al. (1990). Polymerase chain reaction (PCR)-based DNA markers are the valuable tools for analyses of variations between individuals regardless of their developmental stage, which is particularly useful in sex identification studies in plants (Milewicz and Sawicki 2013).

Hormaza et al. (1994) first identified a 945 bp specific band in female plants of *Pistacia vera* L. by RAPD. With the RAPD technological improved and perfected, RAPD technology has been employed largely for sex determination in dioecious species, such as 500 bp and 730 bp specific band in male plants of *Cannabis sativa* L. (Sakamoto et al. 1995), 1190 bp in female plants of *Hippophae salicifolia* D. Don (Rana et al. 2009), 980 bp and 1200 bp specific band in female plants and 650 bp, 850 bp, and 1400 bp in male plants of *Piper betle* L. (Samantaray et al. 2012). In addition, *Salix viminalis* (Alstrom-Rapaport et al. 1998), *Silene latifolia* (Zhang et al. 1998), *Actinidia chinensis* (Shirkot et al. 2002), *Ginkgo biloba* (Jiang et al. 2003) *Encephalartos natalensis* (Prakash and Vanstaden 2006), *Gracilaria changei* (Sim et al. 2007), *Carica papaya* L. (Urasaki et al. 2002, Niroshini et al. 2000), and *Commiphora wightii* (Samantaray et al. 2010).

Although developed in the early 1990s and have a problem with reproducibility, RAPD markers still enjoy the highest popularity for analysis of sex determination because they are simple, cheaper, and less time consuming.

3.4.2 Amplified Fragments Length Polymorphism (AFLP)

Amplified fragment length polymorphism (AFLP), a DNA fingerprint technology developed by Zabeau and Vos (1993) and introduced by Vos et al. (1995), can detect a large number of fragments in a single reaction. It is popular for ingenious primer design and collocation flexibility. Recently, it has been widely used to identify specific markers in dioecious plants.

In *Asparagus officinalis* L., Reamon-Buttner et al. (1998) isolated 9 sex-linked AFLP markers, subsequently cloned and sequenced. Three AFLP fragments were chosen for designing STS primers (Reamon-Buttner and Jung 2000). Semerikov et al. (2003) found four sex linked AFLP markers in female of *Salix viminalis* L. Rahman and Ainsworth (2004) found Mse1.1/Pst1.1 (primer name/restriction enzyme) in male of *Rumex rothschildianus*, Mse1.1/Pst1.3 and Mse1.1/Pst1.4 in male of *Rumex hastatulus.* In addition, sex-specific DNA fragments have been identified in *Eucommia ulmoides* Oliv. (Wang et al. 2011), *Ficus fulva* Reinw. Ex. (Parrish et al. 2004), *Uapaca kirkiana* Muell. Arg. (Mwase et al. 2007), and *Simmondsia chinensis* (Link) Schneider (Agarwal et al. 2011).

In the identification of male and female gender, AFLP not only overcomes the complications of RFLP and poor stability of RAPD technology, but also inherits the reliability of RFLP and convenience of RAPD. However, it needs long experiment process, higher laboratory and operator requirements.

3.4.3 Sequence-characterized Amplified Region (SCAR)

Sequence-characterized amplified regions (SCAR) (Paran and Michelmore 1993) converted from RAPD by specific oligonucleotide primers are markers that more stably amplify, because they can be amplified at higher annealing temperatures with locus specificity and with high reproducibility for molecular identification, and are applicable in identifying the sex of plants without flowering (Esfandiyari et al. 2011; Khadke et al. 2012).

Recently, SCAR markers have been widely used to develop sex-linked molecular markers in several dioecious plants; Jiang and Sink (1997) discovered the specific SCAR marker of SCC15 (980 bp) in male plants and can be used for early sex identification of *Asparagus officinalis* L. Mandolino et al. (1999) amplified a 390 bp region by PCR using primer OPA, SCAR specific markers in the *Cannabis sativa* L. male plant. The specific SCAR marker of SDP (225 bp) was found in *Carica papaya* L. male plant (Urasaki et al. 2002). In addition, *Salix viminalis* (Gunter et al. 2003), *Rumex nivalis* (Stehlik and Blattner 2004), *Piper longum* (Manoj et al. 2005), *Asparagus officinalis* (Gao et al. 2007), *Carica papaya* (Bedoya and Nunez 2007), *Gingko biloba* (Liao et al. 2009), and *Pistachia atlantica* (Esfandiyari et al. 2011).

SCAR markers are more independent of reaction conditions, and are usually dominant markers that can detect a single locus, making them more reproducible (Jiang and Sink 1997), but it needs complex operation and high cost.

In addition, several sex specific inter simple sequence repeat (ISSR) markers were also reported in different dioecious species such as *Humulus lupulus* L. (Danilova and Karlov 2006), *Phoenix dactylifera* L. (Younis et al. 2008), *Carica papaya* L. (Da Costa et al. 2011), and *Trichosanthes dioica* Roxb. (Nanda et al. 2013).

4 Proteomics for Sex Determination in Dioecious Plants

4.1 Proteomics

With the advent of post-genomic era, proteomics has become one of the most important aspects of life sciences. In 2001, Human Proteome Organization was established in the USA, and then Europe and Asia Pacific region also set up a regional organization of proteome research, and tried to complete the human proteome project in cooperation.

Proteomics is the large-scale study of proteins, particularly their structures and functions (Anderson et al. 1998; Blackstock and Weir 1999). In the last decade, the global-scale analysis of proteins developed very rapidly. Regardless of basic theory and technology, proteomics has evolved from the early qualitative, protein cataloging towards a quantitative approach. At present, plant proteomics is focused on model plants like Arabidopsis, rice, and maize to address the biochemical, physiological, metabolic, and developmental processes. The advances in plant proteomics have been reviewed previously (Rossignol et al. 2006).

Typically, the approach for protein separation is two-dimensional gel electrophoresis (2-DE), followed by mass spectrometry (MS) analysis. Despite rapid advances in gel-free proteomics, 2-DE coupled to MS currently remains the dominant proteomic technique (Rossignol et al. 2006). Protein samples are first separated by isoelectric focusing according to the isoelectric point in an immobilized pH gradient (IPG), followed by the second dimensional separation of sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) based on molecular weight. The use of IPG has markedly increased the reproducibility and resolution of 2-DE.

The newly developed fluorescence difference gel electrophoresis (DIGE) labels protein samples with fluorescent dyes before 2-DE, enabling accurate analysis of differences in protein abundance between pair samples within the same gel, thus avoiding gel-to-gel variance (Amme and Mock 2006). Afterwards, software (e.g., PDQuest, ImageMaster) is used to analyze protein profiles, and spots of interest are subjected to MS or N-terminal sequencing.

Besides, the latest technology of iTRAQ is the first set of multiplexed, amine-specific, stable-isotope reagents that can label all peptides in up to eight different biological samples, enabling simultaneous identification and quantitation, both relative and absolute, while retaining important posttranslational modification (PTM)

information. It has many advantages: simultaneously identifies and quantifies proteins from multiple samples; provides flexibility to multiplex up to eight different biological samples simultaneously in a single experiment; offers a simple work flow without sample fractionation for reduced-complexity samples, such as affinity pull-downs; is fully supported by ProteinPilot™ Software on all AB SCIEX proteomics LC/MS/MS platforms.

In addition, two-dimensional capillary electrophoresis and liquid chromatography–capillary electrophoresis technique combination with mass spectrometry can be used for proteomics research. All of the above methods can be used for differential proteomic analysis of dioecious plants.

4.2 Proteomics in Sex Determination of Dioecious Plants

As a powerful tool for protein identification and functional characterization, proteomics techniques have been widely used in plant biology and recently have been introduced to identify specific proteins involved in sex determination (Xiong et al. 2013).

In dioecious plants cultivated for fruit or seed, it is difficult to identify females at juvenile stage. However, females often invest more in reproduction, and less in growth and maintenance than males. This differential investment between sexes may result in distinct growth patterns and sex-specific responses to environmental stresses. Proteins are the primary products in the realization of hereditary information and reflect the genetic structure of the organism most precisely. It is expected that the differences between female and male would display at the proteome level.

The majority of dioecious plants are non-model organisms with no genome sequencing data. So the reports of proteomics in sex determination of dioecious plants were few, mainly in herbaceous plants like *Asparagus officinalis* L. (Bracale et al. 1990), woody plants like *Pistacia chinensis*, *Ginkgo biloba* L., *Actinidia chinensis*, *Actinidia kolomikta*, and *Momordica charantia* L. (Wang and Zeng 1998; Golan-Goldhirsh et al. 1998; Yang and Fu 2012; Xiong et al. 2013) (Table 1); the studies were performed in different tissues and organs.

Bracale et al. (1990) first used 2-DE to compare and analyze the differences between male and female flowers of *Asparagus officinalis* L. and found five specific proteins in male and four specific proteins in female. Due to technical limitations, the specific proteins were not identified. In 1998, Golan-Goldhirsh et al. used SDS-PAGE and western blotting to identify proteins from *Pistacia vera* between male and female plants during different periods of bud development. The results showed that the high abundance of 32 kDa inflorescence bud protein in male plants and 32 kDa and 27 kDa inflorescence bud proteins in female plants confirmed the practical application of proteome in sex determination of dioecious plants. In *Actinidia chinensis* and *Actinidia kolomikta,* an intense band of approximately 18 kDa was observed in males which was lacking in females, while an intense band of

Table 1 Summary of proteomics in sex determination of dioecious plants

Species	Organ	Methods	Inference	References
Pistacia chinensis	Inflorescence bud	SDS-PAGE Western blot Immunoblotting	An intense band of 32 kDa in male Two intense bands of 32 kDa and 27 kDa in female	Golan-Goldhirsh et al. (1998)
	Leaf	2-DE, MS/MS	The abundance of NB-ARC domain containing protein and light harvesting chlorophyll a/b-binding protein in male is higher than that in female	Xiong et al. (2013)
	Stem phloem	2-DE, MS/MS	Eukaryotic translation initiation factor 5A2, phosphoglycerate kinase 2, and an expressed protein accumulate in high abundance in female than in male Temperature-induced lipocalin level in male is higher than that in female	Xiong et al. (2013)
	Stem xylem	2-DE, MS/MS	Ascorbate peroxidase and temperature-induced lipocalin levels in female are higher than that in male	Xiong et al. (2013)
Asparagus officinalis L.	Flowers	2-DE	Five specific proteins in male Four specific proteins in female	Bracale et al. (1990)
Momordica charantia Linn.	Flower bud	CE	A specific band of 30 kDa in male A specific band of 11 kDa in female	Wang and Zeng (1998)
Ginkgo biloba Linn	Leaf	SDS-PAGE	A specific band of 106 kDa in male	Yang and Fu (2012)
	Inflorescence	SDS–PAGE	A specific band of 28 kDa in male Two specific bands of 36 kDa and 92 kDa in female	Yang and Fu (2012)
Actinidia chinensis, Actinidia kolomikta	Leaf	SDS–PAGE	An intense band of approximately 18 kDa was observed in male plants only An intense band of approximately 67 kDa was observed in female plants only	Khukhunaishvili and Dzhokhadz (2006)

approximately 67 kDa was observed only in females (Khukhunaishvili and Dzhokhadz 2006). Then, *Momordica charantia* and *Ginkgo biloba* were studied by SDS-PAGE, capillary electrophoresis, and other methods, only to find the approximate molecular weight of such proteins.

With the development of mass spectrometry technology, combined 2-DE and mass spectrometry has taken the sex identification in dioecious plants to a new level. The phenol-based protein extraction protocol has been proved to work well in various plant tissues, including woody tissues, based on phenol extraction, 2-DE, and MS technology. Xiong et al. (2013) compared and analyzed the differences between male and female vegetative tissues (leaves and xylem, phloem) of *Pistacia chinensis*. 2-DE analysis has revealed a total of ten differential protein spots between male and female plants in *Pistacia chinensis*, of which seven protein spots have been identified by MS/MS. Three proteins (ascorbate peroxidase, phosphoglycerate kinase 2, and temperature-induced lipocalin) might serve as molecular markers for sex determination between female and male plants. It has been revealed that the abundance of several functional stress proteins vary greatly between both sexes in *Pistacia chinensis*. The identified sex-related proteins will be useful to better understand the molecular events of sex differentiation.

4.3 Advantages and Disadvantages of Proteomics for Sex Determination

Proteins are the products of gene expression and regulation of metabolism, compared to other methods; it is more intuitive to identify the sex of dioecious plants at the protein level. In particular, the technology of 2-DE combined MS/MS and iTRAQ can find out the different proteins between male and female plants, identify the different proteins and analyze the function through protein databases. It is expected to yield a more direct understanding of function and regulation.

Although proteomics has showed a superiority for sex determination in dioecious plants, it still has some disadvantages. First, protein function analysis is based on the premise of genome sequencing, however, majority of dioecious plants are non-model organisms, unlike *Arabidopsis*, food crops like rice and maize, for which genome sequence has been completed. Thus, a major limitation in proteomic analysis of sex determination in dioecious plants is little information of gene sequences available in public databases. In most cases, the accurate identification of proteins in dioecious plants needs to perform MS/MS analysis and refer to the available expressed sequence tags (ESTs) in relevant plant species. In addition, the accuracy is also effects of homologous sequences, this needs to start from scratch sequencing, tremendously increase the cost.

Second, although the differences in a gene will be reflected at the protein level, we cannot ensure that all differences would be reflected. So we cannot carry on an overall analysis of all the differences of dioecious plants. Moreover, although the

techniques of non gel-based quantitative proteomic methods such as the iTRAQ were adopted, which allowed relatively minor changes, the low-abundance proteins are still not easy to be detected completely.

5 Conclusion

In dioecious species, the male and female plants have different value; however, most dioecious plants do not exhibit discernible sexual dimorphism before sexual maturity. A reliable method for sex determination in dioecious plants at juvenile stage followed by reasonable planting would greatly benefit breeding programs of dioecious plants cultivated for fruits or seeds. As a powerful tool for protein identification and gene function analysis, proteomics has been widely used in plant science, and it can be used for sex determination of dioecious plants. Especially in recent years, the advent of new technology, such as iTRAQ and DIGE, has greatly promoted the development of differential proteomes. In spite of this, it still has some unanswered questions, such as low-abundance proteins, cost, etc. So far, there has been little research on proteomics in sex determination of dioecious plants. Therefore, it is of great potential for sex determination through proteomics method. It is expected that the difference between female and male dioecious plants at juvenile stage would be perfectly identified by combining proteomics and DNA/RNA markers.

References

Agarwal M, Shrivastava N, Padh H (2011) Development of sex-linked AFLP markers in *Simmondsia chinensis*. Plant Breed 130:114–116

Agrawal V, Sharma K, Gupta S, Prasad M (2007) Identification of sex in *Simmondsia chinensis* (Jojoba) using RAPD markers. Plant Biotechnol Rep 1:207–210

Ainsworth C (2000) Boys and girls come out to play: the molecular biology of dioecious plants. Ann Bot 86:211–221

Ainsworth C, Lu J, Winfield M, Parker JS (1999) Sex determination by X: autosome dosage: *Rumex acetosa* (sorrel). In: Ainsworth CC (ed) Sex determination in plants. Bios Scientific Publishers, Oxford, pp 121–136

Akagi T, Henry IM, Tao R, Comai L (2014) A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. Science 346:647–650

Allen CE (1917) A chromosome difference correlated with sex differences in *Sphaerocarpos*. Science 46:466–467

Alstrom-Rapaport C, Lascoux M, Wang YC, Roberts G, Tuskan GA (1998) Identification of a RAPD marker linked to sex determination in the basket willow, *Salix viminalis* L. J Hered 89:44–49

Amme S, Mock HP (2006) Proteome analysis of cold stress response in Arabidopsis thaliana using DIGE-technology. J Exp Bot 57:1537–1546

Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19:1853–1861

Aryal R, Ming R (2013) Sex determination in flowering plants: papaya as a model system. Plant Sci 217–218:56–62
Bedoya GC, Nunez V (2007) A SCAR marker for the sex types determination in Colombian genotypes of *Carica papaya*. Euphytica 153:215–220
Bekheet SA, Taha HS, Hanafy MS, Solliman ME (2008) Morphogenesis of sexual embryos of date palm cultured in vitro and early identification of sex type. J Appl Sci Res 4:345–352
Blackburn KB (1923) Sex chromosomes in plants. Nature 112:687–688
Blackstock WP, Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. Trends Biotechnol 17:121–127
Bracale M, Galli MG, Falavigna A, Soave C (1990) Sexual differentiation in *Asparagus officinalis* L. Sex Plant Reprod 3:23–30
Bridges CB (1925) Sex in relation to chromosomes and genes. Am Nat 59:127–137
Bugala W (1951) Sex determination of poplars from the colour of leaf. Forestry Abstr 52:13
Charlesworth D (1985) Distribution of dioecy and self-incompatibility in angiosperms. In: Greewoog PJ, Harvey PH, Slatkin M (eds). Evolution: Essays in honour of John Maynard Smith. Cambridge University Press, Cambridge, pp 237–268
Ciupercescu DD, Veuskens J, Mouras A, Ye D, Briquet M, Negrutiu I (1990) Karyotyping Melandrium album, a dioecious plant with heteromorphic sex chromosomes. Genome 33:556–562
Correns C (1928) Bestimmung, Vererbung und Verteilung des Geschlechtes bei den hoheren Pfl anzen. Hanbuch Vererb 2:1–138
Da Costa FR, Pereira TNS, Gabriel APC, Pereira MG (2011) ISSR markers for genetic relationships in Caricaceae and sex differentiation in papaya. Crop Breed Appl Biotechnol 11: 352–357
Danilova TV, Karlov GI (2006) Application of inter simple sequence repeat (ISSR) polymorphism for detection of sex specific molecular markers in hop (*Humulus lupulus* L.). Euphytica 151:15–21
Darwin C (1877) The different forms of flowers on plants of the same species. Murray, London
Dauphin-Guerin B, Teller G, Durand B (1980) Different endogenous cytokinins between male and female *Mercurialis annua* L. Planta 144:124–129
Dawson TD (1993) Gender specific physiology, carbon isotope discrimination and habitat distribution in box elder acer negundo. Ecology 74:798–815
Dellaporta SL, Calderon-Urrea A (1993) Sex determination in flowering plants. Plant Cell 5:1241–1251
Dzhaparidze LI (1969) Sex in plants. Part 2. Biochemical and physiological sex differences in dioecious plants. Problem of influencing sex formation. Academy of Sciences of the Georgian SSR, Institute of Botany (translated from Russian by Israel Program for Scientific)
Esfandiyari B, Davarynejad GH, Shahriari F, Kiani M, Mathe A (2011) Data to the sex determination in Pistachia species using molecular markers. Euphytica 185:227–231
Farbos I, Veuskens J, Vyskot B, Oliveira M, Hinnisdaels S, Aghmir A, Mouras A, Negrutiu I (1999) Sexual dimorphism in white campion: deletion on the Y chromosome results in a floral asexual phenotype. Genetics 151:1187–1196
Flach M (1966) Nutmeg cultivation and its sex problems. Eng Sum Meded Landh Hoogesh 66:1–85
Gao WJ, Li RL, Li SL, Ding CL, Li SP (2007) Identification of two markers linked to the sex locus in dioecious *Asparagus officinalis* plants. Russ J Plant Physiol 54:816–821
Geber MA, Dawson TE, Delph LF (1999) Gender and sexual dimorphism in flowering plants. Springer, Berlin
Golan-Goldhirsh A, Peri I, Birk Y, Smirnoff P (1998) Inflorescence bud proteins of *Pistacia vera*. Trees 12:415–419
Grant S, Houben A, Vyskot B, Siroky J, Pan WH, Macas J, Saedler H (1994) Genetics of sex determination in flowering plants. Dev Genet 15:214–230
Gunter LE, Roberts GT, Lee L, Larimer FW, Tuskan GA (2003) The development of two flanking SCAR markers linked to a sex determination locus in *Salix viminalis* L. J Hered 94:185–189

Heikrujam M, Sharma K, Prasad M, Agrawal V (2014) Review on different mechanisms of sex determination and sex-linked molecular markers in dioecious crops: a current update. Euphytica 201:1293

Hormaza JI, Dollo L, Polito VS (1994) Identification of a RAPD marker linked to sex determination in *Pistacia vera* using bulked segregant analysis. Theor Appl Genet 89:9–13

Irish E, Nelson T (1989) Sex determination in monoecious and dioecious plants. Plant Cell 1:737–744

Jaiswal VS, Narayan P, Lal M (1984) Activities of acid and alkaline phosphatases in relation to sex differentiation in *Carica papaya* L. Biochem Physiol Pfianzen 1984(179):799–801

Janousek B, Mrackova M (2010) Sex chromosomes and sex determination pathway dynamics in plant and animal models. Biol J Linn Soc 100:737–752

Jaquiéry J, Stoeckel S, Rispe C, Mieuzet L, Legeai F, Simon JC (2012) Accelerated evolution of sex chromosomes in aphids, an X0 system. Mol Biol Evol 29:837–847

Jiang C, Sink KC (1997) RAPD and SCAR markers linked to the sex expression locus M in asparagus. Euphytica 94:329–333

Jiang L, You RL, Li MX, Shi C (2003) Identification of a sex associated RAPD marker in *Ginkgo biloba*. Acta Bot Sin 45:742–747

Kahlem G (1976) Isolation and localization by histoimmunology of isoperoxidases specific for male flowers of the dioecious species *Mercurialis annua* L. Dev Biol 50:58–67

Khadke GN, Bindu KH, Ravishankar KV (2012) Development of SCAR marker for sex determination in dioecious betel vine (Piper betle L). Curr Sci 103:712–716

Khukhunaishvili RG, Dzhokhadz DI (2006) Electrophoretic study of the proteins from actinidia leaves and sex identification. Appl Biochem Microbiol 42:107–110

Kihara H, Ono T (1923) The sex chromosomes of *Rumex acetosa*. Zeitschrift für Induktive Abstammungs und Verebungslehre 39:1–7

Korekar G, Sharma RK, Kumar R (2012) Identification and validation of sex-linked SCAR markers in dioecious *Hippophae rhamnoides* L. (Elaeagnaceae). Biotechnol Lett 34:973–978

Kumar S, Kumari R, Sharma V (2014) Genetics of dioecy and causal sex chromosomes in plants. J Genet 93:241–277

Liao L, Liu J, Dai Y, Li Q, Xie M, Chen Q, Yin H, Qiu G, Liu X (2009) Development and application of SCAR markers for sex identification in the dioecious species *Ginkgo biloba* L. Euphytica 169:49–55

Li C, Xu G, Zang R, Korpelainen H, Berninger F (2007) Sex-related differences in leaf morphological and physiological responses in *Hippophae rhamnoides* along an altitudinal gradient. Tree Physiol 27(3):399–406

Louis JP (1989) Genes for the regulation of sex differentiation and male fertility in *Mercurialis annua* L. J Heredity 80:104–111

Maestri E, Restivo FM, Marziani Longo GP, Falavigna A, Tassi E (1991) Isozyme gene markers in the dioecious species *Asparagus officinalis* L. Theor Appl Genet 81:613–618

Mandolino G, Carboni A, Forapani S, Faeti V, Ranalli P (1999) Identification of DNA markers linked to the male sex in dioecious hemp. Theor Appl Genet 98:86–92

Manoilov J (1922) Identification of the sexes in dioecious plants by chemical reaction. BaI App Bot Plautbreed 13:503–505

Manoj P, Banerjee NS, Ravichandran P (2005) Development of sex-associated SCAR markers in *Piper longum* L. PGR News Lett 141:44–50

Matsunaga S, Kawano S (2001) Sex determination by sex chromosomes in dioecious plants. Plant Biol 3:481–488

Matsunaga S, Hizume M, Kawano S, Kuroiwa T (1994) Cytological analysis in *Melandrium album*, genome size, chromosome size and florescence in situ hybridization. Cytologia 59:135–141

McClung CE (1902) The accessory chromosome—sex determinant. Biolo Bull-us 3:43–84

Milewicz M, Sawicki J (2013) Sex-linked markers in dioecious plants. Plant Omics J 6:144–149

Ming R, Wang J, Moore PH, Paterson AH (2007) Sex chromosomes in flowering plants. Am J Bot 94:2141–2150

Ming R, Bendahmane A, Renner SS (2011) Sex Chromosomes in land plants. Annu Rev Plant Biol 62:485–514
Mwase WF, Erik-Lid S, Bjornstad A, Stedje B, Kwapata MB, Bokosi JM (2007) Application of amplified fragment length polymorphism (AFLPs) for detection of sex-specific markers in dioecious Uapaca kirkiana Muell. Arg Afr J Biotechnol 6:137–142
Nanda S, Kar B, Nayak S, Jha S, Joshi RK (2013) Development of an ISSR based STS marker for sex identification in pointed gourd (*Trichosanthes dioica Roxb.*). Sci Hortic 150:11–15
Negrutiu I, Vyskot B, Barbacar N, Georgiev S, Moneger F (2001) Dioecious plants: a key to the early events of sex chromosome evolution. Plant Physiol 127:1418–1424
Niroshini E, Everard JMDT, Karunanayake EH, Tirimanne TLS (2000) Sex-specific random amplified polymorphic DNA (RAPD) markers in *Carica papaya*. Trop Agr Res 12:41–49
Pakull B, Groppe K, Mecucci F, Gaudet M, Sabatti M, Fladung M (2011) Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a *Populus tremula x Populus* tremuloides cross. Can J For Res 2:245–253
Paran I, Michelmore RW (1993) Development of reliable PCR based markers linked to downy mildew resistance genes in lettuce. Theor Appl Genet 85:985–993
Parker JS (1990) Sex chromosomes and sexual differentiation in flowering plants. Chromosomes Today 10:187–198
Parrish TL, Koelewijn HP, van Dijk PJ (2004) Identification of a male-specific AFLP marker in a functionally dioecious fig, *Ficus fulva Reinw.ex Bl.* (Moraceae). Sex Plant Reprod 17:l7–l22
Penel CL, Greppin H (1972) Evolution de l'activité auxines-oxydasique et peroxydasique lors de l'induction photopériodique et de la sexualisation de l'épinard. Plant Cell Physiol 13:151–156
Perice LG, Brewbaker JL (1973) Applications of isozyme analysis in horticultural science. HortSci 8:17
Prakash S, Vanstaden S (2006) Sex identification in Encephalartos natalensis (Dyer and Verdoorn) using RAPD markers. Euphytica 152:197–200
Qacif N, Baaziz M, Bendiab K (2007) Biochemical investigations on peroxidase contents of male and female inflorescences of date palm (*Phoenix dactylifera* L.). Sci Hortic 114:298–301
Rahman MA, Ainsworth CC (2004) AFLP analysis of genome difference between males and females in dioecious plant *Rumex acetosa*. J Biol Sci 4:160–169
Rana S, Shirkot P, Yadav MC (2009) A female sex associated randomly amplified polymorphic DNA marker in dioecious *Hippophae salicifolia*. Genes Genom Genomics 3:96–101
Reamon-Buttner SM, Jung C (2000) AFLP-derived STS markers for the identification of sex in *Asparagus officinalis* L. Theor Appl Genet 100:432–438
Reamon-Buttner SM, Schondelmaier J, Jung C (1998) AFLP markers tightly linked to the sex locus in *Asparagus officinalis* L. Mol Breed 4:91–98
Renner SS, Ricklefs RE (1995) Dioecy and its correlates in the flowering plants. Am J Bot 82:596–606
Rick CM, Hanna GC (1943) Determination of Sex in *Asparagus officinalis* L. Am J Bot 30:711–714
Rossignol M, Peltier JB, Mock HP, Matros A, Maldonado AM, Jorrin JV (2006) Plant proteome analysis: a 2004-2006 update. Proteomics 6:5529–5548
Sakamoto KL, Shimomura K, Kamada H, Satoh S (1995) A male-associated DNA sequence in a dioecious plant, *Cannabis sativa* L. Plant Cell Physiol 36:1549–1554
Samantaray S, Geetha KA, Hidayath KP, Maiti S (2010) Identification of RAPD markers linked to sex determination in guggal [*Commiphora wightii* (Arnott.)] Bhandari. Plant Biotechnol Rep 4:95–99
Samantaray S, Phurailatpam A, Bishoyi AK, Geetha KA, Maiti S (2012) Identification of sex-specific DNA markers in betel vine (*Piper betle* L.). Genet Resour Crop Ev 59:645–653
Sansone FW (1938) Sex determination in Silene otites and related species. J Genet 35:387–396
Santos JK (1923) Differentiation among chromosomes in *Elodea*. Bot Gaz 75:42–59
Semerikov V, Lagercrantz U, Tsarouhas V, Ronnberg-Wastljung A, Alstrom-Rapaport C, Lascoux M (2003) Genetic mapping of sex-linked markers in *Salix viminalis* L. Heredity 91:293–299

Sharma K, Agrawal V, Prasad M, Gupta S, Kumar R, Prasad M (2008) ISSR marker-assisted selection of male and female plants in a promising dioecious crop, jojoba (*Simmondsia chinensis*). Plant Biotechnol Rep 2:239–243

Sharma A, Zinta G, Rana S, Shirko P (2010) Molecular identification of sex in Hippophae rhamnoides L. using isozyme and RAPD markers. For Stud China 12:62–66

Shephard H, Parker J, Darby P, Ainsworth CC (1999) Sex expression in hop (*Humulus lupulus L.* and *H. japonicus Sieb. et Zucc.*): floral morphology and sex chromosomes. BIOS Scientific Publishers, Oxford, pp 137–148

Shephard HL, Parker JS, Darby P, Ainsworth CC (2000) Sexual development and sex chromosomes in hop. New Phytol 148:397–411

Shirkot P, Sharma DR, Mohapatra T (2002) Molecular identification of sex in *Actinidia deliciosa var.* deliciosa by RAPD markers. Sci Hortic 94:33–39

Siljak-Yakovlev S, Benmalek S, Cerbah M, Coba de la Peña T, Bounaga N, Brown S, Sarr A (1996) Chromosomal sex determination and heterochromatin structure in date palm. Sex Plant Reprod 9:127–132

Sim MC, Lim PE, Gan SY, Phang SM (2007) Identification of random amplified polymorphic DNA (RAPD) marker for differentiating male from female and sporophytic thalli of *Gracilaria changii* (Rhodophyta). J Appl Phycol 19:763–769

Sondur SN, Manshardt RM, Stiles JI (1996) A genetic linkage map of *papaya* based on randomly amplified polymorphic DNA markers. Theor Appl Genet 93:547–553

Stehlik I, Blattner FR (2004) Sex-specific SCAR markers in the dioecious plant *Rumex nivalis* (Polygonaceae) and implication for the evolution of sex chromosome. Theor Appl Genet 108:238–242

Truta E, Gille E, Toth E, Maniu M (2002) Biochemical differences in *Cannabis sativa L.* depending on sexual phenotype. J Appl Genet 43:451–462

Urasaki N, Tokumoto M, Tarora K, Ban Y, Kayano T, Tanaka H, Oku H, Chinen I (2002) A male and hermaphrodite specific RAPD marker for papaya (*Carica papaya L.*). Theor Appl Genet 104:281–285

Vos P, Hogers R, Bleeker M, Reijans M, Van Der Lee T, Homes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP, a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Vyskot B, Hobza R (2004) Gender in plants: sex chromosomes are emerging from the fog. Trends Genet 20:432–438

Wang QM, Zeng GW (1998) Study of specific protein on sex determination of *Momordica charantia* L. Acta Bot Sin 40(3):241–246

Wang D, Li Y, Li Z (2011) Identification of a male-specific amplified fragment length polymorphism (AFLP) and a sequence characterized amplified region (SCAR) marker in *Eucommia ulmoides* Oliv. Int J Mol Sci 12:857–864

Wang XR, Liu WZ (2011) Development of oil bodies in the fruit of *Pistacia chinensis*. Chin Bullet Bot 46:665–674

Warmke HE (1946) Sex determination and sex balance in Melandrium. Am J Bot 33:648–660

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Westergaard M (1948) The relation between chromosome constitution and sex in the offspring of triploid Melandrium. Hereditas 34:257–279

Westergaard M (1958) The mechanism of sex determination in dioecious flowering plants. Adv Genet 9:217–281

Williams JGK, Rubelik AR, Livak KJ, Rafalski A, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Xiong EH, Wu XL, Shi J, Wang XY, Wang W (2013) Proteomic identification of differentially expressed proteins between male and female plants in *Pistacia chinensis*. PLoS One 8, e64276

Yang JH, Fu QY (2012) The Preliminary analysis of specific protein on sex determination of Ginkgo biloba L. Sci Technol Informat 31:241

Yin T, Difazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA (2008) Genome structure and emerging evidence of an incipient sex chromosome in populus. Genome Res 18:422–430

Younis RAA, Ismail OM, Soliman SS (2008) Identification of sex-specific DNA markers for date palm (*Phoenix dactylifera L.*) using RAPD and ISSR techniques. Res J Agric Biol Sci 4:278–284

Zabeau M, Vos P (1993) Selective restriction fragment amplification: a general method for DNA fingerprinting. Eur Patent Appl EP 0534858

Zhang YH, Distilo VS, Rehman F, Avery A, Mulcahy D (1998) Y chromosome specific markers and the evolution of dioecy in the genus Silene. Genome 41:141–147

Zhao YY, Liu J (1991) Physiological and biochemical characteristics and identification of the sexes in dioecious plants. J Beijing Teachers College 12:27–33

Metabolome Analysis of Crops

Sameen Ruqia Imadi and Alvina Gul

Contents

Abstract Plant metabolomics deals with study of plants by measuring gene expression, protein interaction, and other different regulatory processes. Metabolomic studies are much closer to phenotypes than mRNA transcripts and proteins. Metabolite profiling approaches are used to study stress response of different plants. Metabolic profiling of phytomedicinal species is a powerful tool for quality control, yield optimization and to study environmental and ecological interactions. Metabolomics can be effectively used for the analysis of species lacking sequenced genome. Phenotype leans on intricate metabolic interactions of an organism and hence metabolomics is paramount for crop improvement in breeding programs. Complex polygenic inheritance of agronomic traits can be dissected through metabolic profiling. The chapter deals with application of different stresses and conditions on model crop plants and studying their effects on metabolites of these plants. A reference is made to metabolomic studies on wheat, rice, maize, sugarcane, barley, and cotton. Finally the chapter looks into the future prospects of metabolomic studies and their application for advanced research in plant sciences.

S.R. Imadi • A. Gul (✉)
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_16

Keywords Metabolite profiling • *Arabidopsis thaliana* • Crop breeding • Crop improvement • Rice • Barley

1 Introduction

If we take a look at the past, we can see that it was not long back when genomic techniques were quite ambitious to be applied on plants and other multicellular organisms. But now these techniques constitute an emerging field of systems biology. They comprise a valuable technology casting a vast and clear global picture of biological organization (Hall 2006). Plant biology like other fields of biology has undergone major transitions and transformations in previous years (Fernie 2003). Metabolomics refers to profiling of metabolites on large scale for quantitation of metabolites in organisms leading to a platform for diagnostics, gene function analysis, and systems biology. Metabolite profiling aims to identify and quantify specific class or class of chemically related metabolites through effective metabolite analysis techniques (Fernie et al. 2004).

Metabolome of a plant is a highly complex, dynamic assortment of primary and secondary compounds because it is the final downstream product of genome (Hagel and Facchini 2008). Metabolomics, proteomics, and transcriptomics techniques are collectively known as omic techniques (Davies 2010). Metabolomics has provided a new direction for the study of biological systems. This technique is observed to be a valuable tool for high throughput screening of bioactive substances (Aliferis and Chrysayi-Tokousbalides 2011). Metabolomics is a global assessment and validation of endogenous small molecules also known as metabolites in plants and other biological systems. These metabolites when analyzed can reveal the state of a biological system and is widely used for diagnosis of certain disorders and diseases (Zhang et al. 2012).

Plant metabolites are known to possess a large chemical diversity. Every plant has a complex set of metabolites. Advances in plant metabolomics provide possibilities to profiling of multiple metabolites and quantitative analysis of selected metabolites with the aid of systems biology (Oksman-Caldentey and Saito 2005). Techniques which involve the analysis of metabolite levels in plants can prove to be a very potent tool for plant research and biotechnology. By the use of these techniques specific classes of metabolites can be measured. Metabolome analysis can help to profile a wide range of low molecular weight compounds (Stitt and Fernie 2003). Liquid chromatography–mass spectrometry (LC-MS) based metabolomics is considered to be a very powerful tool for profiling of metabolites. A wide variety and range of metabolites can be distinguished and separated by use of combination of different extraction methods, separation columns, and ion detection methods (Tohge et al. 2011a; b).

Metabolomic studies rely on the multitude of small metabolites present in plants. These studies are highly associated with mass spectrometry and nuclear magnetic resonance as parallel technologies. Liquid chromatography when combined with

NMR technique makes a powerful methodology for identification of different metabolites. This will lead to better characterization of metabolomes (Moco and Vervoort 2007). Metabolomic analysis of soybean plants was conducted to get an insight into the alteration of mitochondrial functions as a result to flooding stress. It was observed that flooding causes an alteration in mitochondrial membrane proteins. It directly impairs the electron transport chain. NADH production was also seen to increase in mitochondria during tricarboxylic acid cycle as a result to flooding (Komatsu et al. 2011).

Changes in metabolome levels of plants reflect the activities of cells at functional levels (Putri et al. 2013). Metabolome analysis of matured and non-stressed leaves of *Populus euphratica* reveals that primary sugars are more strongly accumulated, sugar alcohol production pathways get activated, and secondary metabolites are quickly consumed. Physiological measurements show higher tannin and soluble Phenolic contents which is associated with enrichment of glucose and fructose (Janz et al. 2010). Metabolome analysis of pea plants under drought stress reveals a large difference in primary and secondary metabolites and alteration in metabolic pathways. Proline, valine, threonine, homoserine, myoinositol, gamma aminobutyrate, and trigonelline are observed to be present in high concentration in drought stressed plants. Concentrations of glutamate, asparagines, and malate are also seen to alter in this condition but are not specifically associated to drought stress. Metabolome analysis is used to identify non-targeted gene products in a specific biological sample (Charlton et al. 2008).

Nuclear magnetic resonance based metabolomic analysis studies reveal the quantitative trait loci in strawberry which control fruit quality and fruit development (Moing et al. 2004). *Suaeda salsa* is a halophyte plant living in saline soils. Metabolomic analysis was conducted to reveal the effects of high salt concentrations on this plant. It was observed that salt stress inhibits the growth of plant and induces significant metabolic responses. Decrease in amino acids, lactate, 4-aminobutyrate, malate, choline, and phosphocholine was seen with an accompanied increase in betaine, sucrose, and allantoin in roots. Superoxide dismutase, glutathione S-transferase, peroxidase, catalase, and glutathione peroxidase are seen to enhance their activities. This shows osmotic and oxidative stress and disturbance in energy metabolism in *Suaeda salsa* (Wu et al. 2012).

Lactuca sativa plants subjected to ultraviolet B radiations were analyzed metabolomically. It was observed that as a result of higher light conditions a group of secondary compounds is upregulated and as a result of exposure to UV-B rays the levels of these compounds are further elevated (Wargent et al. 2014).

2 Plant Metabolomics: Applications

The wealth of scientific findings has drastically increased with the induction of metabolomic reports (Fiehn et al. 2007). Metabolomic studies have defined applications in drug discovery, systems biology, molecular biology, cell biology, medical

sciences, and agricultural sciences. Metabolome analysis of medicinal plants provides with evidence based development of phytotherapeutics and nutraceuticals. Disease development, drug discovery and chemical toxicology have also flourished with the integration of metabolomics (Shyur and Yang 2008). Initially the metabolome based plant studies that were used for only metabolome analysis of plants can now contribute to stress biology study in plants. The technique is used to identify different compounds like by-products of stress metabolism, signal transduction molecules, and all molecules which are involved in plant acclimation towards stress (Shulaev et al. 2008).

One of the major applications of metabolome analysis lies in food component analysis. Using metabolomic techniques food quality can be assessed, and consumption of food can be monitored (Wishart 2008). It is a fast-growing technology which is very useful for phenotyping and diagnostic analysis of plants. The technology is also becoming a key player in functional annotation of genes and understanding of cellular responses to biological and nonbiological conditions comprehensively. Recently metabolome approaches have been used to identify the natural variance in metabolite content between individual plants which leads to crop improvement by improving the compositional quality of crops (Schauer and Fernie 2006).

Metabolomics is quickly paving its way in plant studies. It helps in many prospects including the metabolite alterations and modifications in plants under normal conditions, abiotic and biotic stresses. Metabolomic responses of plants towards temperature, water, food, light, and circadian rhythms and seasons can be measured using this technique. Toxicology of pesticides, insecticides, and antifungal chemicals can also be measured (Bundy et al. 2009). One of the most important applications of metabolome analysis of crops lies in identifying the metabolic differences between genetically modified crops and conventional crops. Metabolome analysis techniques are also applied on plants to determine the natural phytochemical variation which provides an indication about effects of geographical position, crop management and breeding programs on metabolite levels and profiles (Rischer and Oksman-Caldentey 2006).

Metabolomic techniques have helped to the insights of small biochemical molecules in bacteria, plants, and animals (Idle and Gonzalez 2007). As it is an omic technology, it has many applications in synthetic biology, medicine, medical sciences, and predictive plant modeling (Putri et al. 2013). The technique of metabolomics is widely been used in ecology and environmental sciences. Ecological metabolomics deals with mechanisms of plant resistance to herbivores, whereas environmental metabolomics deals with the effects of herbicides, pesticides, and insecticides on plants and plant tolerance towards them (Macel et al. 2010). Metabolomic techniques can be used to obtain information on metabolic status of cells and identify the effects of foreign genes on cells (Okazaki and Saito 2012).

Metabolomic approach is till date the best approach to identify genetically modified crops and transgenic plants (Kusano et al. 2011). Metabolomic techniques can be implied for making quality assessments, discriminating taxonomic relationships, plant natural products, illuminating genotype differences, measuring biochemical

content, measuring tolerance of plants to different stresses, and understanding source to sink transitions (Guy et al. 2008). It is the era to conduct metabolomic analysis on large scale with respect to metabolite measured and experiments carried out (Lisec et al. 2006).

Metabolomic approaches are continuously being used by scientists to discover new responses to genetic or environmental perturbation and to validate the hypothesis of gene products and in vivo actions of gene products (Fiehn et al. 2007). One of the major applications of metabolomics lie in pharmaceutical research and development. Metabolite profiling helps in analysis of drug safety and research on toxicological studies (Lindon et al. 2006). Metabolomics can be used in functional genomics to differentiate the plants of different origin (Kim et al. 2010). Metabolome analysis of plants is essential to understand the fundamental stress physiology and stress response mechanisms in plants (Rodziewicz et al. 2014).

3 Metabolomic Studies on Some Leading Crop Plants

3.1 Arabidopsis thaliana (A Model Plant)

Although metabolomic studies on plants are mostly performed on *Arabidopsis thaliana*, they are genome independent due to which the studies conducted on this plant can be applied to a wide range of plant species (Tohge et al. 2011a; b). Metabolite profiling of sulfur stressed *Arabidopsis thaliana* plants revealed that decrease in sulfur supply is associated with decrease in levels of proteins and chlorophylls. Decrease in concentrations of amino acids like cysteine and methionine is also observed which results in degraded proteins (Nikiforova et al. 2005). Metabolite profiling of *Arabidopsis thaliana* plants kept under sulfur and nitrogen deficiencies shows that several metabolic pathways are affected. Metabolites involved in glucosinolate metabolism are observed to be highly enhanced (Hirai et al. 2004).

210 lines of *Arabidopsis thaliana* plants which were used for targeted metabolite quantitative trait locus were subjected to metabolite profiling. It was observed that the distribution of metabolite quantitative trait loci across the genome include 11 QTL clusters. Out of these 11 clusters, 8 are associated with the epistatic network that is involved in regulation of plant metabolism (Rowe et al. 2008). Metabolite analysis of one wild type and two mutant types of *Arabidopsis thaliana* was conducted to get knowledge about alteration of metabolite accumulation in different types. The mutant which was methionine hyper-accumulator was observed to loss overall network connectivity due to over accumulation of methionine. Transparent testa4 mutants were observed to possess a new correlation between malate and sinapate. It was also seen that the levels of malate, sinapate and sinapoyl malate remain unchanged. This proves that the methionine hyper-accumulator loses the metabolic stability, whereas transparent testa4 generates metabolic network of backup pathway for lost physiological functions (Kusano et al. 2007).

Gas chromatography–mass spectrometry profiling was used to identify 433 metabolites in a sample of *Arabidopsis thaliana.* It was observed that primary metabolites vary when compared to all other metabolites in different lines. Malic acid and citrate are considered to be the most important metabolites which vary in different lines. It was also observed that glucose and fructose are the two metabolites that are involved in discriminating between the crosses (Taylor et al. 2002). Metabolomic profiling of *Arabidopsis thaliana* reveals 35 known compounds in all the isolated compounds. These compounds include six anthocyanins, eight flavonols, two dicarboxylic acids, one hydrocarbon, one steroid, two chlorophyll derivatives, three galactolipids, one apocarotenoid, one carotenoid, four phenylpropanoids with three indoles, one indole glucosinolate, and one nucleoside (Nakabayashi et al. 2009).

Metabolome analysis of *Arabidopsis* plants overexpressing the PAP1 gene which encodes for MYB transcription factor exposes that cyanidin and quercetin derivatives are highly accumulated. Eight anthocyanins are also observed among an array of putative 1800 metabolites in PAP1 overexpressing plants. It was also observed that PAP1 induces the production of glycosyltransferase, acyltransferase, glutathione S-transferase, sugar transporters, and transcription factors (Tohge et al. 2005). Metabolome analysis of *Arabidopsis* plants subjected to salt stress shows that methylation cycle for supply of methyl groups, phenylpropanoid pathway for lignin production, and biosynthesis of glycine betaine are induced. Long-term exposure to salt stress results in co-induction of sucrose metabolism and glycolysis. Reduction of methylation cycle is also observed (Kim et al. 2007).

Metabolome analysis of cadmium stressed *Arabidopsis* plants show that the stress activate carbon, nitrogen and sulfur metabolic pathways. Six different families of phytochelatins increase their accumulation in response to high cadmium concentrations. Many metabolic adaptations take place as a result to cadmium stress (Sarry et al. 2006). Flavonol synthase (FLS) are the enzymes which are associated with conversion of flavonols to dihydroflavonols in flavonoids biosynthesis pathway. Metabolite analysis of *Arabidopsis* plants reveals the presence of many members of FLS gene family but it was observed that flavonol synthase 1 (FLS1) is the only enzyme which catalyzes this pathway. FLS1 null mutants of *Arabidopsis* show high levels of anthocyanins and low levels of Flavonol glycoside. Accumulated glycosylated forms of dihydroflavonols are also seen (Stracke et al. 2009).

Metabolomics of *Arabidopsis thaliana* grown under sulfur stress show that a group of metabolites that are regulated by same mechanism, cluster together. Glucosinolate metabolism is regulated and anthocyanin biosynthesis is enhanced (Hirai et al. 2005).

3.2 Gossypium (Cotton)

In cotton, Ligon lintless 2 (Li2) is expected to control the quality of yarn and fabric. Metabolomic analysis of Li2 mutant cotton plants was conducted to examine the processes which are involved in cotton fiber accumulation. It was observed that

metabolome of mutant fibers was altered. Levels of free sugars, sugar alcohols, sugar acids, and sugar phosphates were highly decreased. Downregulation of processes which were associated with carbohydrate biosynthesis, cell wall loosening, and cytoskeletons takes place. Mutant fibers had increased accumulation of gamma aminobutyric acid. Higher nitrate assimilation is observed with high levels of 2-ketoglutarate, succinate, and malate (Naoumkina et al. 2013).

Difference between *Gossypium raimondii* and *Gossypium arboretum* was analyzed using metabolomic approaches. Metabolite profiling of these two species reveal 206 identified compounds. It was observed that 186 identified metabolites, out of a total of 206 differentiate highly in accumulation in the two species. These two species differ in protein biosynthesis and ROS management during transition stage and secondary cell wall deposition. Glycolysis, starch and sucrose metabolism, amino and nucleotide sugar metabolism, phenylpropanoid biosynthesis, galactose metabolism, nitrogen metabolism, and glutathione metabolism are some of the pathways that differ in both the species (Tuttle 2014).

3.3 Hordeum vulgare (Barley)

Shoots and roots of phosphorus-deficient barley plants were metabolomically profiled using gas chromatography–mass spectrometry technique. It was observed that mildly phosphorus deficient plants accumulate disaccharides and trisaccharides including sucrose, maltose, raffinose, and 6-ketose in shoots. Severe phosphorus deficient plants show increased level of metabolites that are related to ammonium metabolism. High level of disaccharides and trisaccharides and low levels of phosphorylated intermediates and organic acids are also observed in severely phosphorus deficient barley plants. Carbohydrate metabolism is affected by reduction in phosphorus consumption and phosphorus metabolites. These plants also showed a sharp decrease in glutamine and asparagine levels in both shoots and roots (Huang et al. 2008).

Boron toxicity in barley plants was studied using metabolomic approaches. Metabolite profiles of root and leaf tissues of intolerant commercial cultivar of barley and boron-tolerant Algerian landrace were compared. It was observed that the amplitude of metabolite changes in roots of commercial cultivar was greater than that of boron-tolerant cultivars. The sensitive cultivar showed a drastic decrease in metabolite levels in leaf tips which can be associated with gradual accumulation of boron in leaves (Roessner et al. 2006).

Storage metabolome of barley vacuole was studied, and 59 primary metabolites with known structures and 200 secondary metabolites with predicted structures were identified. These metabolites majorly comprised of amino acids, organic acids, sugars, sugar alcohols, shikimate pathway intermediates, vitamins, phenylpropanoids, and flavonoids. Out of a total of 259 putative metabolites, 12 were found to exist solely in vacuole, whereas 34 of them existed in protoplast and the rest 213 metabolites were common in both (Tohge et al. 2011a; b).

Fusarium head blight resistance in barley was studies using metabolome analysis strategy. It was observed that resistance indicator metabolites like deoxynivalenol and its detoxification product are enhanced in barley plants inoculated with Fusarium head blight. Resistance biomarker metabolites like phenylalanine, P-coumaric acid, jasmonate, and linolenic acid increase their accumulation in resistant genotypes. Fusarium head blight resistant barley species were metabolomically profiled. It was observed that there among a total of 161 metabolites, 53 are resistance related or pathogenicity related. These metabolites mainly belonged to three types of metabolic families: fatty acids, phenylpropanoids, and flavonoids (Kumaraswamy et al. 2011).

Impact of drought stress in barley plants was studied using metabolomic approaches. It was observed that 17 metabolites are highly affects as a result to drought stress. These metabolites include fructose and glucose as monosaccharides, raffinose as trisaccharide, organic acids, and biogenic amine gamma aminobutyric acid. These compounds enhance their accumulation in drought-stressed plants (Wenzel et al. 2015).

Metabolome of barley is composed of high levels of phenolics like ferulic acid, caffeic acid, sinapinic acid, and their esters (Khakimov et al. 2014). Metabolome profiling of cultivated barley species and wild species was conducted to analyze the effects of salinity. It was observed by metabolite profiling that osmotic adjustment is key mechanism for salt tolerance. In roots polyols had showed a role for induction of salt tolerance. High level of sugars and energy in roots and enhanced rate of photosynthesis was also observed in salt treated barley plants (Wu et al. 2013).

3.4 Oryza sativa (Rice)

Metabolome quantitative trait loci (mQTL) analysis of rice plants show that there are 803 mQTLs in rice genome which have uneven distribution. Most metabolites in rice are highly affected by environmental factors, which can easily be observed by their changing levels (Matsuda et al. 2012). Rice plants were subjected to metabolome analysis at milking stage to measure the effects of temperature on grain filling. It was observed that high temperature results in increased levels of sucrose and pyruvate/oxaloacetate derived amino acids. This is also associated with decrease in levels of sugar phosphates and organic acids which are involved in glycolysis and glucogenesis and tricarboxylic acid cycles. Starch deposition may be impaired by downregulation of sucrose import and degradation of sucrose and ATPs (Yamakawa and Hakata 2010).

Transgenic rice plants overexpressing YK 1 were analyzed motobolomically. It was observed that composition of organ specific and tissue specific metabolites is not altered in transgenic rice plants when compared to control plants. It was also seen that expression of quite a less number of metabolites was modified (Takahashi et al. 2005). Rice cultivars with compatible and incompatible strains of rice blast fungal pathogen *Magnaporthe grisea* were metabolomically analyzed. It was observed that alanine was largely elevated in the leaves of compatible plants. Alanine

concentration was upto 30 % more in compatible plants as compared to resistant plants. Malate, glutamine, proline, cinnamate, and some unknown sugars are also involved in fungal penetration in leaves (Jones et al. 2011).

Metabolomic analysis of rice coleoptiles in anaerobic environment was conducted to evaluate the tolerance of hypoxia. It was observed that tricarboxylic acid cycle is highly altered and diverts to synthesize glutamate, GAB and glutamine in hypoxic plants. Malate production increases which results in enhancement of glycoxylate cycle to synthesize malate. This leads to maintenance of glycolysis for energy production (Fan et al. 2003). Rice seeds were subjected to metabolite profiling. Positive correlation between shikimic acids and phenolics has been observed. Out of a total 52 identified metabolites, 45 were primary metabolites, while the rest 7 were phenolic acids (Kim et al. 2013).

With the help of metabolomic analysis, the difference between red, black, and non-colored indica and japonica rice species were identified. It was observed that among different species and varieties of rice, different levels of anthocyanins cyanidin-3-glucoside and peonidin-3-glucoside are present. Among all the varieties, black rice variety was shown to possess a high level of fatty acid methyl esters, free fatty acids, organic acids, and amino acids (Frank et al. 2012). Metabolite profiling of 8 day old transgenic rice plants producing large amount of tryptophan showed that the plants had higher levels of anthranilate, tryptamine, and serotonin. Free tryptophan was also seen in large amount. Overproduction of tryptophan does not induce any effect on the production of phenylalanine and tyrosine. Overaccumulation of free tryptophan may be due to low activity of tryptophan decarboxylase and other metabolic genes that need tryptophan (Dubouzet et al. 2007).

Metabolome analysis of zinc stressed rice plants showed low accumulation of monosaccharide sugars and increased concentrations of hydrogen peroxide, phenolics, peroxidases, and nitrogen rich metabolites in roots. High level of citrate, allantoin, and stigmasterol are also observed which are caused by high levels of hydrogen peroxide. Significant decrease in concentrations of tricarboxylic acid cycle intermediates succinate and aromatic amino acid tyrosine is seen. Bicarbonate stress applied on rice plants showed reduction in iron concentrations in shoot which results in lower iron dependent ascorbate peroxidase activity. Bicarbonate stress also results in accumulation of tyrosine and malate, fumarate, and succinate which are tricarboxylic acid cycle intermediates (Rose et al. 2012). Thirty-six compounds were isolated from rice leaves. These compounds were analyzed metabolomically, and it was discovered that five of them were flavonoids and eight were flavonolignan isomers (Yang et al. 2014).

3.5 Saccharum (Sugarcane)

In a study, sugarcane plants were subjected to metabolome analysis. It was observed that the metabolites which are correlated with level of sucrose accumulation increased down the stem. Metabolites like tricarboxylic acid cycle intermediates

and amino acids were abundant in meristem to internode 2 samples but decreased in more matured internodes down the stem. Levels of trehalose and raffinose show positive correlations with sucrose concentration (Glassop et al. 2007). Transgenic sugarcane plants modified with proteinase inhibitors were analyzed metabolomically. The levels of chlorogenic acid, syringic acid, glucose, sucrose, threonine, alanine, aspartic acid, proline, fumaric acid, succinic acid, choline, glycine, asparagines, and unidentified polyphenols were observed in transgenic stems as well as control stems. And it was concluded that transgenic plants and control plants depict no differences so transgenic sugarcane can be consumed without fear of any danger (Lira et al. 2009).

Sugarcane cultures when tested metabolomically showed significant changes in the levels of glucose, fructose, sucrose and maltose in cultures grown in embryogenic callus. Embryogenic callus, non-embryogenic callus, and non-embryogenic callus–media relationships are also shown to possess different amino acid compounds like glutamine, asparagines, threonine, alanine, leucine, and valine and organic acid derivatives which basically include lactate, malonate, choline, 4-aminobutyrate, and 2-hydroxyisobutyrate (Mahmud et al. 2014).

3.6 Solanum

Genetically modified potato tubers and conventional potato tubers when analyzed metabolomically showed that apart from intentional and targeted changes, the genetically modified potatoes are similar to traditional potatoes (Catchpole et al. 2005). Potato tubers when subjected to metabolite profiling showed compositional changes after genetic modifications. 40 genetically modified lines were analyzed metabolomically. It was observed that the compounds which are most notably affected are proline, trigonelline, and phenolic compounds (Defernez et al. 2004).

Metabolite profiling and metabolite fingerprinting of genetically modified potatoes show that their metabolite components are same as that of their corresponding conventional cultivars (Colquhoun et al. 2006). Potato tubers experience six stages in life cycle. Metabolite profile using Mass spectrometry analysis of these stages confirms that these stages can be distinguished from each other on the basis of metabolome analysis. Manipulation of source–sink relationship in potato tubers is seen to significantly affect the metabolite levels beyond the general sugar–starch metabolism (Shepherd et al. 2010).

Western flower thrip is a common pest of tomatoes. Resistance for thrip was observed by metabolomic analysis of cultivated and wild tomato species. Results obtained show that only wild tomato species are thrip resistant whereas the cultivated species of tomatoes are susceptible to thrips. Thrip-resistant tomato varieties contain a large amount of acylsugars which act as a resistant factor (Mirnezhad et al. 2010).

3.7 Triticum (Wheat)

Durum wheat grain from four different cultivars was grown in conventional as well as organic farming in three consecutive years. Metabolomic analysis was conducted on amino acids, organic acids, fatty acids, sugars, and sterols to determine the effects of genotype, environment, and genotype–environment interactions. It was observed that genotype had a very small effect on the metabolite composition genotype by environment is shown to possess large affects on metabolite composition and quality of grain (Beleggia et al. 2013).

Phloem exudate of wheat was analyzed metabolomically. 79 metabolites were found out in the phloem exudate. Out of 79, 53 were identified. As the wheat proceeds towards maturity, 39 metabolites are shown to modify their concentrations with an increase in 21 and a decrease in 18 (Palmer et al. 2014).

3.8 Zea mays

Metabolome analysis of maize plants subjected to long term nitrogen deficiency show a major deficiency in carbon metabolism and a decrease in downstream metabolic processes. Nitrogen deficiency stress is observed to cause same alterations in metabolism, as many other biotic and abiotic stresses cause. Metabolically nitrogen assimilation gets affected as a result to nitrogen deficiency stress (Amiour et al. 2012).

Metabolite profiling technique when applied to maize plants under environmental variations showed a differential expression of 15 metabolites on comparison with control maize plants (Barros et al. 2010). On the basis of metabolite profiling and metabolomic analysis of transgenic maize lines, it can be said that the transgenic lines have significant differences in metabolite levels and contents when compared to their wild relatives and wild isogenic lines. Hence transgenic maize plants can be differentiated from wild plants on the basis of metabolome analysis (Leon et al. 2009).

4 Conclusion and Future Prospects of Metabolomics in Plant Sciences

Metabolomics has enabled us to get a deeper insight into the fundamental biochemical bases of things that we eat. This technique can help in designing modified breeding programs which are aimed for better quality production, food processing strategy optimization, and improvement of bioavailability of nutrients (Hall et al. 2008; Dun et al. 2013). The post-genomic era specifically needs metabolomic

approaches to flourish. As a result of metabolomic analysis a large quantity of information can be generated but only a small amount is required. The success of research depends on the efficiency to extract meaningful information and data (Goodacre et al. 2004).

The field of metabolomics has continued its rapid growth in the last decade and is proven to be a very powerful technology for predicting and explaining complex phenotypes in diverse biological systems. The technique can be integrated with other omic approaches like genomics, transcriptomics, and proteomics to get a better understanding of global systems biology (Putri et al. 2013; Sakurai et al. 2013). Plant metabolomics is providing large data sets that enable us to pave a way towards comprehensive understanding of plant growth, development, defense, and productivity. Recent advances in metabolomics and post genomics allow the gathering of enriched information on individual cells as well as single cell types (Misra et al. 2014).

Next-generation sequencing and advanced proteomic methods may cast a challenge on metabolomics, but it should be understood that metabolomics is currently the only technique by which one can analyze the whole genome sequences. Metabolomic strategies tend to be used in genome sequence information (Tohge et al. 2014; Glauser et al. 2013). It is predicted that in future metabolomic techniques will play a major role in overpassing the gap between phenotype and genotype. Metabolomic analysis is able to assist in genome sequence annotation and thus can help in linking a gene to a function. Plants are a rich source of diverse biochemicals. It can be said that metabolomic studies can provide a better understanding for identifying and defining the unexploited biodiversity of plants (Hall 2006).

Plant research and biotechnology will flourish in future with the integration of metabolomics with transcriptomic and proteomic techniques. With the use of these techniques, plant phenotypes will be linked to gene, protein expression and metabolite synthesis and accumulation which is a challenge presently (Amiour et al. 2012). Functional genomics and systems biology techniques can be combined in future to analyze crop plants for developing the ability of worldwide stable food systems (Jogaiah et al. 2013).

State-of-the-art genomic tools can be combined with metabolome profiling to identify the major genes which can be engineered. This will lead to production of improved crop plants (Oksman-Caldentey and Saito 2005). Analysis of molecular phenotype using metabolome analysis will enhance our novel understanding of plant metabolism and its interaction with environment (Weckwerth 2011).

Metabolomic approaches can be used in future for the purpose of breeding of crops because plant traits like taste and yield are very closely related to metabolite conditions of plants (Oikawa et al. 2008). The tool of metabolomics is expected to play a key role in acceleration of understanding off mode the action of bioactive compounds (Aliferis and Jabaji 2011).

References

Aliferis KA, Chrysayi-Tokousbalides M (2011) Metabolomics in pesticide research and development: review and future perspectives. Metabolomics 7(1):35–53

Aliferis KA, Jabaji S (2011) Metabolomics—a robust bioanalytical approach for the discovery of the modes-of-action of pesticides: a review. Pesticides Biochem Physiol 100(2):105–117

Amiour N, Imbaud S, Clément G, Agier N, Zivy M, Valot B, Balliau T, Armengaud P, Quilleré I, Cañas R, Tercet-Laforgue T, Hirel B (2012) The use of metabolomics integrated with transcriptomic and proteomic studies for identifying key steps involved in the control of nitrogen metabolism in crops such as maize. J Exp Bot 63(14):5017–5033

Barros E, Lezar S, Anttonen MJ, Van Dijk JP, Röhlig RM, Kok EJ, Engel K-H (2010) Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. Plant Biotechnol J 8(4):436–451

Beleggia R, Platani C, Nigro F, De Vita P, Cattivelli L, Papa R (2013) Effect of genotype, environment and genotype-by-environment interaction on metabolite profiling in durum wheat (Triticum durum Desf.) grain. J Cer Sci 57(2):183–192

Bundy JG, Davey MP, Viant MR (2009) Environmental metabolomics: a critical review and future perspectives. Metabolomics 5(1):3–21

Catchpole GS, Beckmann M, Enot DP, Mondhe M, Zywicki B, Taylor J, Hardy N, Smith A, King RD, Kell DB, Fiehn O, Draper J (2005) Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. Proc Natl Acad Sci U S A 102(4):14458–14462

Charlton AJ, Donarski JA, Harrison M, Jones SA, Godward J, Oehlschlager S, Arques JL, Ambrose M, Chinoy C, Mullineaux PM, Domoney C (2008) Responses of the pea (Pisum sativum L.) leaf metabolome to drought stress assessed by nuclear magnetic resonance spectroscopy. Metabolomics 4(4):312–327

Colquhoun IJ, Le Gall G, Elliott KA, Mellon FA, Michael AJ (2006) Shall I compare thee to a GM potato. Trends Genet 22(10):525–528

Davies H (2010) A role for "omics" technologies in food safety assessment. Food Control 21(12):1601–1610

Defernez M, Gunning YM, Parr AJ, Shepherd LVT, Davies HV, Colquhoun IJ (2004) NMR and HPLC-UV profiling of potatoes with genetic modifications to metabolic pathways. J Agric Food Chem 52(20):6075–6085

Dubouzet JG, Ishihara A, Matsuda F, Miyagawa H, Iwata H, Wakasa K (2007) Integrated metabolomic and transcriptomic analyses of high-tryptophan rice expressing a mutant anthranilate synthase alpha subunit. J Exp Bot 58(12):3309–3321

Dun WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, Hankemeier T, Goodacre R, Neumann S, Kopka J, Viant MR (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9(1):44–66

Fan TW-M, Lane AN, Higashi RM (2003) In vivo and in vitro metabolomic analysis of anaerobic rice coleoptiles revealed unexpected pathways. Russ J Plant Physiol 50(6):787–793

Fernie AR (2003) Metabolome characterisation in plant system analysis. Func Plant Biol 30(1):111–120

Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5:763–769

Fiehn O, Sumner LW, Rhee SY, Ward J, Dickerson J, Lange BM, Lane G, Roessner U, Last R, Nikolau B (2007) Minimum reporting standards for plant biology context information in metabolomic studies. Metabolomics 3(3):195–201

Frank T, Reichardt R, Shu Q, Engel K-H (2012) Metabolite profiling of colored rice (Oryza sativa L.) grains. J Cer Sci 55(2):112–119

Glassop D, Roessner U, Bacic A, Bonnett GD (2007) Changes in the sugarcane metabolome with stem development. Are they related to sucrose accumulation? Plant Cell Physiol 48(4):573–584

Glauser G, Veyrat N, Rochat B, Wolfender JL, Turlings TCJ (2013) Ultra-high pressure liquid chromatography–mass spectrometry for plant metabolomics: a systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. J Chromatogr A 1292:151–159

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22(5):245–252

Guy C, Kopka J, Moritz T (2008) Plant metabolomics coming of age. Physiol Plant 132(2):113–116

Hagel JM, Facchini PJ (2008) Plant metabolomics: analytical platforms and integration with functional genomics. Phytochem Rev 7(3):479–497

Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol 169(3):453–468

Hall RD, Brouwer ID, Fitzgerald MA (2008) Plant metabolomics and its potential application for human nutrition. Physiol Plant 132(2):162–175

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proc Natl Acad Sci U S A 101(27):10205–10210

Hirai MY, Klein M, Fujikawa Y, Yano M, Goodenowe DB, Yamazaki Y, Kanaya S, Nakamura Y, Kitayama M, Suzuki Y, Sakurai N, Shibata D, Tokuhisa J, Reichelt M, Gershenzon J, Papenbrock J, Saito K (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. J Biol Chem 280:25590–25595

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(5):691–703

Idle JR, Gonzalez FJ (2007) Metabolomics. Cell Metab 6(5):348–351

Janz D, Behnke K, Schnitzler J-P, Kanawati B, Schmitt-Kopplin P, Polle A (2010) Pathway analysis of the transcriptome and metabolome of salt sensitive and tolerant poplar species reveals evolutionary adaption of stress tolerance mechanisms. BMC Plant Biol 10(150): doi:10.1186/1471-2229-10-150

Jogaiah S, Govind SR, Tran L-SP (2013) Systems biology-based approaches toward understanding drought tolerance in food crops. Crit Rev Biotechnol 33(1):23–39

Jones OAH, Maguire ML, Griffin JL, Jung Y-H, Shibato J, Rakwal R, Agrawal GK, Jwa N-S (2011) Using metabolic profiling to assess plant-pathogen interactions: an example using rice (Oryza sativa) and the blast pathogen Magnaporthe grisea. Eur J Plant Physiol 129(4):539–554

Khakimov B, Jespersen BM, Engelsen SB (2014) Comprehensive and comparative metabolomic profiling of wheat, barley, Oat and Rye using Gas chromatography-mass spectrometry and advanced chemometrics. Foods 3(4):569–585

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(3):415–424

Kim HK, Choi YH, Verpoorte R (2010) NMR-based metabolomic analysis of plants. Nat Protoc 5:536–549

Kim JK, Park S-Y, Lim S-H, Yeo Y, Cho HS, Ha S-H (2013) Comparative metabolic profiling of pigmented rice (Oryza sativa L.) cultivars reveals primary metabolites are correlated with secondary metabolites. J Cer Sci 57(1):14–20

Komatsu S, Yamamoto A, Nakamura T, Nouri M-Z, 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(9):3993–4004

Kumaraswamy GK, Bollina V, Kushalappa AC, Choo TM, Dion Y, Rioux S, Mamer O, Faubert D (2011) Metabolomics technology to phenotype resistance in barley against Gibberella zeae. Eur J Plant Physiol 130(1):29–43

Kusano M, Fukushima A, Arita M, Jonsson P, Moritz T, Kobayashi M, Hayashi N, Tohge T, Saito K (2007) Unbiased characterization of genotype-dependent metabolic regulations by metabolomic approach in Arabidopsis thaliana. BMC Syst Biol 1:53

Kusano M, Redestig H, Hirai T, Oikawa A, Matsuda F, Fukushima A, Arita M, Watanabe S, Yano M, Hiwasa-Tanase K, Ezura H, Saito K (2011) Covering chemical diversity of genetically-

modified tomatoes using metabolomics for objective substantial equivalence assessment. PLOS One 6(2), e16989
Leon C, Rodriguez-Meizoso I, Lucio M, Garcia-Cañas V, Ibañez E, Schmitt-Kopplin P, Cifuentes A (2009) Metabolomics of transgenic maize combining Fourier transform-ion cyclotron resonance-mass spectrometry, capillary electrophoresis-mass spectrometry and pressurized liquid extraction. J Chromatogr A 1216(43):7314–7323
Lindon JC, Holmes E, Nicholson JK (2006) Metabonomics techniques and applications to pharmaceutical research & development. Pharm Res 23(6):1075–1088
Lira T, Yariwake JH, Choi YH, Kim HK, Verpoorte R (2009)Metabonomic study of transgenic and non-transgenic sugarcane leaves based on NMR profile. Planta Med 75: DOI: 10.1055/s-0029-1234415
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
Macel M, Van Dam NM, Keurentjes JJB (2010) Metabolomics: the chemistry between ecology and genetics. Mol Ecol Resour 10(4):583–593
Mahmud I, Thapaliya M, Boroujerdi A, Chowdhury K (2014) NMR-based metabolomics study of the biochemical relationship between sugarcane callus tissues and their respective nutrient culture media. Anal Bioanal Chem 406(24):5997–6005
Matsuda F, Okazaki Y, Oikawa A, Kusano M, Nakabayashi R, Kikuchi J, Yonemaru J-I, Ebana K, Yano M, Saito K (2012) Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. Plant J 70(4):624–636
Mirnezhad M, Romero-González RR, Leiss KA, Choi YH, Verpoorte R, Klinkhamer PG (2010) Metabolomic analysis of host plant resistance to thrips in wild and cultivated tomatoes. Phytochem Anal 21(1):110–117
Misra BB, Assmann SM, Chen S (2014) Plant single-cell and single-cell-type metabolomics. Trends Plant Sci 19(10):637–646
Moco S, Vervoort J (2007) Metabolomics technologies and metabolite identification. TrAC Trends Anal Chem 26(9):855–866
Moing A, Maucourt M, Renaud C, Gaudillère M, Brouquisse R, Lebouteiller B, Gousset-Dupont A, Vidal J, Granot D, Denoyes-Rothan B, Lerceteau-Kohler E, Rolin D (2004) Quantitative metabolic profiling by 1-dimensional 1H-NMR analyses: application to plant genetics and functional genomics. Funct Plant Biol 31(9):889–902
Nakabayashi R, Kusano M, Kobayashi M, Tohge T, Yonekura-Sakakibara K, Kogure N, Yamazaki M, Kitajima M, Saito K, Takayama H (2009) Metabolomics-oriented isolation and structure elucidation of 37 compounds including two anthocyanins from Arabidopsis thaliana. Phytochemistry 70(8):1017–1029
Naoumkina M, Hinchliffe DJ, Turley RB, Bland JM, Fang D (2013) Integrated metabolomics and genomics analysis provides new insights into the fiber elongation process in Ligon lintless-2 mutant cotton (Gossypium hirsutum L.). BMC Genomics 14:155
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(1):304–318
Oikawa A, Matsuda F, Kusano M, Okazaki Y, Saito K (2008) Rice metabolomics. Rice 1(1):63–71
Okazaki Y, Saito K (2012) Recent advances of metabolomics in plant biotechnology. Plant Biotechnol Rep 6(1):1–15
Oksman-Caldentey K-M, Saito K (2005) Integrating genomics and metabolomics for engineering plant metabolic pathways. Curr Opin Biotechnol 16(2):174–179
Palmer LJ, Dias DA, Boughton B, Roessner U, Graham RD, Stangoulis JCR (2014) Metabolite profiling of wheat (Triticum aestivum L.) phloem exudate. Plant Methods 10:27
Putri SP, Nakayama Y, Matsuda F, Uchikata T, Kobayashi S, Matsubara A, Fukusaki E (2013) Current metabolomics: practical applications. J Biosci Bioeng 115(6):579–589
Rischer H, Oksman-Caldentey K-M (2006) Unintended effects in genetically modified crops: revealed by metabolomics? Trends Biotechnol 24(3):102–104
Rodziewicz P, Swarcewicz B, Chmielewska K, Wojakowska A, Stobiecki M (2014) Influence of abiotic stresses on plant proteome and metabolome changes. Acta Physiolgiae Plantarum 36(1):1–19

Roessner U, Patterson JH, Forbes MG, Fincher GB, Langridge P, Bacic A (2006) An investigation of boron toxicity in barley using metabolomics. Plant Physiol 142(3):1087–1101

Rose MT, Rose TJ, Pariasca-Tanaka J, Yoshihashi T, Neuweger H, Goesmann A, Frei M, Wissuwa M (2012) Root metabolic response of rice (Oryza sativa L.) genotypes with contrasting tolerance to zinc deficiency and bicarbonate excess. Planta 236(4):959–973

Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20(5):1199–1216

Sakurai T, Yamada Y, Sawada Y, Matsuda F, Akiyama K, Shinozaki K, Hirai MY, Saito K (2013) PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. Plant Cell Physiol 54(2):5

Sarry J-E, Kuhn L, Ducruix C, Lafaye A, Junot C, Hugouvieux V, Jourdain A, Bastien O, Fievet JB, Vailhen D, Amekraz B, Moulin C, Ezan E, Garin J, Bourguignon Dr J (2006) The early responses of Arabidopsis thaliana cells to cadmium exposure explored by protein and metabolite profiling analyses. Proteomics 6(7):2180–2198

Schauer N, Fernie AR (2006) Plant metabolomics: towards biological function and mechanism. Trends Plant Sci 11(10):508–516

Shepherd LVT, Alexander CA, Sungurtas JA, McNicol JW, Stewart D, Davies HV (2010) Metabolomic analysis of the potato tuber life cycle. Metabolomics 6(2):274–291

Shulaev V, Cortes D, Miller G, Mittler R (2008) Metabolomics for plant stress response. Physiol Plant 132(2):199–208

Shyur L-F, Yang N-S (2008) Metabolomics for phytomedicine research and drug development. Curr Opin Chem Biol 12(1):66–71

Stitt M, Fernie AR (2003) From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon–nitrogen interactions. Curr Opin Biotechnol 14(2):136–144

Stracke R, De Vos RCH, Bartelniewoehner L, Ishihara H, Sagasser M, Martens S, Weisshaar B (2009) Metabolomic and genetic analyses of flavonol synthesis in Arabidopsis thaliana support the in vivo involvement of leucoanthocyanidin dioxygenase. Planta 229(2):427–445

Takahashi H, Hotta Y, Hayashi M, Kawai-Yamada M, Komatsu S, Uchimiya H (2005) High throughput metabolome and proteome analysis of transgenic rice plants (Oryza sativa L.). Plant Biotechnol 22(1):47–50

Taylor J, King RD, Altmann T, Fiehn O (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics 18:S241–S248

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J-I, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, Noji M, Yamazaki M, Saito K (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J 42(2):218–235

Tohge T, Mettler T, Arrivault S, Carroll AJ, Stitt M, Fernie AR (2011) From models to crop species: caveats and solutions for translational metabolomics. Front Plant Sci 2(61): doi: 10.3389/fpls.2011.00061

Tohge T, Ramos MS, Nunes-Nesi A, Mutwil M, Giavalisco P, Steinhauser D, Schellenberg D, Willmitzer L, Persson S, Martinoia E, Fernie AR (2011b) Toward the storage metabolome: profiling the barley vacuole. Plant Physiol 157(3):1469–1482

Tohge T, de Souza LP, Fernie AR (2014) Genome-enabled plant metabolomics. J Chromatogr B 966:7–20

Tuttle R (2014) Integrated transcriptomics and metabolomics between 10 and 28 days post anthesis in two species of commercial cotton fiber. Plant Animal Genome XXII. Last accessed on May 7, 2016. https://pag.confex.com/pag/xxii/webprogram/Paper9894.html

Wargent JJ, Nelson BCW, Mcghie TK, Barnes PW (2014) Acclimation to UV-B radiation and visible light in Lactuca sativa involves up-regulation of photosynthetic performance and orchestration of metabolome-wide responses. Plant Cell Environ doi: 10.1111/pce.12392

Weckwerth W (2011) Green systems biology—from single genomes, proteomes and metabolomes to ecosystems research and biotechnology. J Proteome 75(1):284–305

Wenzel A, Frank T, Reichenberger G, Herz M, Engel K-H (2015) Impact of induced drought stress on the metabolite profiles of barley grain. Metabolomics 11:454–467

Wishart DS (2008) Metabolomics: applications to food science and nutrition research. Trends Food Sci Tech 19(9):482–493

Wu H, Liu X, You L, Zhang L, Yu J, Zhou D, Zhao J (2012) Salinity-induced effects in the halophyte Suaeda salsa using NMR-based metabolomics. Plant Mol Biol Reporter 30(3):590–598

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. doi:10.1371/journal.pone.0055431

Yamakawa H, Hakata M (2010) Atlas of rice grain filling-related metabolism under high temperature: joint analysis of metabolome and transcriptome demonstrated inhibition of starch accumulation and induction of amino acid accumulation. Plant Cell Physiol 51(5):795–809

Yang Z, Nakabayashi R, Okazaki Y, Mori T, Takamatsu S, Kitanaka S, Kikuchi J, Saito K (2014) Toward better annotation in plant metabolomics: isolation and structure elucidation of 36 specialized metabolites from Oryza sativa (rice) by using MS/MS and NMR analyses. Metabolomics 10(4):543–555

Zhang A, Sun H, Wang P, Han Y, Wang X (2012) Recent and potential developments of biofluid analyses in metabolomics. J Proteome 75(4):1079–1088

Plant Metabolomics and Strategies

Halbay Turumtay, Cemal Sandallı, and Emine Akyüz Turumtay

Contents

Abstract Plant metabolomics is the study of metabolic pathways and processes to understand how plants grow and carry out functions. Over the past years, there has been a switch to a more holistic view of metabolomes. Instead of looking at individual pathways and metabolites, scientists are looking at interactions, broad pictures, and regulations. Combination of spectrometry-based technology database with (un)supervised multivariate statistical methodologies reveals a deep insight into complex metabolite patterns of plant derived samples. The ambitious goal of identifying the structure of all metabolites and hence all metabolic pathways is almost becoming within reach as MS- and NMR-based technologies and data mining approaches are rapidly evolving.

Keywords LC-MS • Metabolomics • Processing of metabolic data • NMR • GC/ MS

H. Turumtay (✉)
Department of Energy System Engineering, Karadeniz Technical University, 61830 Trabzon, Turkey
e-mail: halbay.turumtay@gmail.com

C. Sandallı
Department of Biology, Recep Tayyip Erdogan University, 53100 Rize, Turkey

E.A. Turumtay
Department of Chemistry, Recep Tayyip Erdogan University, 53100 Rize, Turkey

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_17

1 Introduction

Recent development in technologies on DNA sequencing (genomics), gene expression analysis (transcriptomics), and protein analysis (proteomics) were not enough to identify different metabolic pathway in plants. The missing link in functional genomics strategies might fill by metabolomics studies. Metabolomics is the term used for high-throughput analysis of complex metabolite mixtures in plant extracts. Metabolic data outputs reveal a completely different level of information from genomic source, which does not necessarily show collinearity with genome sequencing (Saito and Matsuda 2010). The total number of the metabolites of a single plant is very speculative and is expected to exceed the number of currently known metabolites multiple times. Roessner et al. (2001) and von Roepenack-Lahaye et al. (2004) estimated that Arabidopsis comprise approximately 5000 metabolites of which 1015 have been identified today. For the plant kingdom, almost 100,000 (mainly secondary) metabolites have been identified to date (Oksman-Caldenty et al. 2004). Many important crops have already been targeted for their metabolomics profiling such as rice (Hall et al. 2008), wheat (Graham et al. 2009), tomato (Moco et al. 2006), melon (Moing et al. 2011), coffee (Lindinger et al. 2009), and potato (Beckman et al. 2007).

The metabolome consist of a variety of metabolites with different physic-chemical properties (polarity, acidity, molecular weight, extractability, affinity.). It also extends over an estimated nine magnitudes of concentration (pM-mM) (Dunn and Ellis 2005). This complexity, in combination with the technical limits of analytic instruments, makes it currently impossible to get overview of the entire metabolome in one single or small number of analyses. All current extraction and detection techniques, irrespective of their level of sophistication, have an unavoidable intrinsic bias toward certain metabolite groups (Hall 2006).

Based on objectives of the study on the one hand and technical limits and capacities on the other hand, metabolite analysis is divided into four classes (Fiehn 2001).

1. Metabolome targeted analysis. By means of this type of analysis, quantitative information on one or selected group of metabolites is congregate. Extensive extraction and separation are commonly obliged to avoid interfering metabolites. Mostly, a high sensitive detector is used (e.g., fluorescence and single ion monitoring (SIM-MS)). This techniques is needed when low detection limits are required such as detection of phytohormones (Prinsen et al. 2000)
2. Metabolic profiling. This approach aims at detecting several predefined targets that are typically metabolites of specific pathways or with related chemical structures such as amino acids. The techniques comprise both sensitive and specific instruments. Gas chromatography (GC), liquid chromatography (LC), and capillary electrophoresis (CE) are generally used in combination with common detector such as mass spectrometry (MS) and photodiode array (PDA) (Theodoridis et al. 2012).
3. Metabolomics. This class is rather theoretical since it refers to the identification and quantification of whole metabolome. Because this is practically impossible,

the term "metabolomics" is commonly used when the metabolome is studied in unbiased way, even if only a fraction of it is revealed. Generally same chromatographic tools are used as in metabolic profiling, commonly combined with MS-detection.

4. Metabolic fingerprinting. This is a high-throughput qualitative screening of metabolic composition with primary aim of sample comparison and classification. Although in general no attempt is made to identify the metabolites, these fingerprints can provide information on functional groups or compound classes. Typical tools are Fourier transformation cyclotron resonance mass spectrometry (FT-ICR/MS) (Johnson et al. 2003), NMR (Reo 2002; Ratcliffe and Shachar-Hill 2005), and direct injection mass spectroscopy (DIMS) (Dunn and Ellis 2005; Aharoni et al. 2004)

2 Recent Development on Plant Metabolomics

As stated before, a remarkably broad metabolic profile can be achieved in single analysis (Kikuchi et al. 2004; von Roepenack-Lahaye et al. 2004; Tohge et al. 2005), but not one single technique is able to measure the complete metabolome. Multiple parallel chromatographic techniques are often required to gain desired broad metabolic picture (Hirai et al. 2004). For instance, untargeted (GC-MS) and targeted (HPLC-MS) metabolic profiling analyses were performed on strawberry fruits (Zhang et al. 2011).

The techniques that come closest to the definition of metabolomics are generally hyphenated techniques. Among these are LC/MS, GC/MS, and CE/MS. LC/MS is the most used with reversed phase for detection of secondary metabolites (Murch et al. 2004). However, LC/MS with hydrophilic interaction chromatography (HILIC) was also used for detection of oligosaccharides and sugar nucleotides (Tolstikov and Fiehn 2002). GC/MS is the gold standard to detect sugars, sugar alcohols, amino acids, volatile compounds, small organic acids, and lignin monomers (G type, S type, and H type oligolignols (Roessner et al. 2001)).

In order to perform performing in vivo metabolomics in plants Laser Ablation Electrospray Ionization-Mass Spectrometry Imaging (LAESI-MSI) has been successfully applied (Etalo et al. 2015). These recent finding results indicate LAESI-based imaging approaches have a potential of as a reliable and fast way to perform metabolomics analyses on living tissues (Etalo et al. 2015).

3 LC-MS

LC/MS is the mostly used with reverse phase and gradient elution in metabolite profiling of plants. This makes it possible to separate rather less polar metabolites, usually secondary metabolites. Highly polar compounds (e.g., sugars) do not show

affinity for the hydrophobic stationary phase and will coelute in a cluttered and useless front. Change of the pH and the polarity of the mobile phase determine mostly which metabolites could be separated and thus detected. The most frequently used ionization sources in LC/MS are atmospheric pressure chemical ionization (APCI) and electrospray (ESI) to ionize molecules with acid and/or base functionalities.

A new generation of mass spectrometers such as FT-ICR/MS and Orbitrap show a dramatic improvement of scan rate, sensitivity, resolution, and accuracy (Hirai et al. 2004; Hu et al. 2005; Zubarev and Makarov 2013). These technical improvements might increase the number of the compounds that can be measured in one chromatographic separation and has become a powerful addition to mass spectrometric techniques for increasing selectivity and confidence of routine analyses.

4 GC/MS

GC/MS is valuable tool to detect a variety compounds simultaneously, such as sugars, sugar alcohols, amino acids and small organic acid. However, it is limited in detection of organic diphosphates, cofactors, and metabolites larger than trisaccharides (Weckwerth and Fiehn 2002). The method is claimed to be sensitive, qualitative, reproducible, and relatively fast (one separation takes approximately 1 h). Besides, GC/MS is suitable for automation (Roessner et al. 2000).

Biological variation of plant metabolites measured by GC/MS has been shown to be substantial and to surpass the technical variation severalfold (Roessner et al. 2000; Hall 2006; Kose et al. 2001). These points to the remarkably high flexibility of the metabolome, without affecting the visual phenotype. Moreover, it also illustrates the need to analyze enough biological replicates to estimate the biological variation within the system whereas technical repeats are less important. In case of a typical metabolic profiling, it is advised to work with approximately 12 biological replicates. (Fuzfai et al. 2004; Kaiser and Benner 2012; Pitthard and Finch 2001; Ratsimba et al. 1999)

GC/MS can be applied to detect volatile metabolites (e.g., alcohols and monoterpenes). However, the technology is more broadly applicable to group of nonvolatile, polar (mainly primary) metabolites, such as amino acid, organic acids and sugars, by converting them into volatile and thermostable compounds through chemical derivatization. Derivatization reduces the polarity of the functional groups thereby facilitating their separation by GC. Carbonyl functions are generally converted to the methoxime and acidic protons, which are replaced by silyl-group (such as trimethylsilyl-TMS) (Gulberg et al. 2004; Halket et al. 2005). Despite the big advantages of derivatization, it has some drawback as well. The presence of sterically hindered groups can lead to partial derivatization, especially in case of sterical trimethylsilyl groups. Consequently multiple products are formed from a single metabolite, hereby complicating the chromatogram (Halket et al. 2005). Furthermore, derivatised products are relatively labile, which result in concentration changes over time. However, these drawbacks could be complemented by preparing samples in daily batches, with a common reference throughout the measurement period (Tikunov et al. 2005).

5 NMR

NMR based metabolomics studies is an important tool for biological system and have been applied in various organism including animals, plants, and bacteria. ^{1}H NMR has been predominant profiling method since it fast, simple and could be used for different purposes in plant metabolomics such as quality control (Rasmussen et al. 2006), chemotaxonomy (Roos et al. 2004; Le Gall et al. 2004), and analysis of transgenic plants (Le Gall et al. 2004; Colquhoun 2007). Up to now, a plethora of applications of NMR based metabolomics have been reported. Generally, some 30–150 metabolites are identified (Kim et al. 2011).

The major advantages of high-throughput analysis of NMR analysis are the ease of quantification and simple sample preparation (not at al in certain cases). NMR could also provide information on the absolute quality of metabolites, and thus the ratio and amount of components in a mixture can be determined (Kim and Ralph 2014). Furthermore, NMR also provides a data about molecules stereochemical details (Seger and Sturm 2007). Compared to MS, the weakness point of NMR is its low sensitivity, although recent findings have led to a considerable increase in sensitivity of NMR (Grivet and Delort 2009).

6 Metabolite Identification

One of the challenges in metabolomics is the identification of unknown metabolites. Because of the huge discrepancy between the amount of sample needed to detect unknown peaks with very low abundance and the amount of sample required for structural elucidation of unknown compounds, only a fraction of the measured peaks could be assigned to known compounds. In GC/MS typically could detect metabolites are known and numbers in LC/MS are comparable. Depending on used method, different strategies are possible to unveil the identity of unnamed peak. Interpretation of mass spectral data is often an intricate and time-consuming task. The use of MS-libraries in combination with the retention time or retention indices (RI) is a powerful tool to identify metabolites (Roessner et al. 2001). These MS libraries could be constructed based on chemically synthesized standards. If the molecule is not present in a library, classical analytic chemical tools can be considered (for instance purification and spectroscopic analysis, e.g., NMR, UV, IR).

The current generation of high resolution mass spectrometers such as FT-ICR/MS, LC-PDA-SPE-NMR/MS, and Orbitrap might simplify the identification of unknown peaks based on the MS^{n} spectra. Different techniques were applied will lead to further development of plant metabolomics have been reviewed by Nakabayashi and Saito (2013). Moreover, the high sensitivity of this spectrometer allows us to detect lowly abundant metabolites.

7 Processing of Metabolic Data

Manual processing of small metabolite screening dataset is feasible but its time consuming and only dramatic changes could be observed. In practice, datasets are often too big to look at the metabolites one by one and in addition subtle changes are informative as well. Therefore, thigh-throughput data processing methods are required for metabolomics. One of the elegant ways of this high-throughput system is a candidate substrate–product pair (CSPP) system, which has been developed by Morreel et al. (2014). Morreel et al. (2014) developed an algorithm in which liquid chromatography-mass spectrometry profiles are searched for pairs of peaks that have mass and retention time differences corresponding with those of substrates and products from well-known enzymatic reactions. This method allows the annotation of low-abundance compounds that are otherwise not amenable to isolation and purification. This method will greatly advance the value of metabolomics in systems biology (Morreel et al. 2014).

8 Conclusion and Future Perspective

Metabolomics is a tool to improve our understanding of the metabolic pathways and biochemistry of organisms. It is certain that plant metabolomics is in its infancy and still in a dynamic phase of development. To get a deeper biological meaning of living organisms, metabolic studies need more data about known metabolites. In order to increase the number of known compounds, deeper technologies for identification of unknown compounds are certainly necessary. Recent technological improvements in both metabolite identification and data interpretation revealed that there are still lots of things to do.

References

Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB (2004) Nontargeted metabolome analysis by use of fourier transform ion cyclotron mass spectrometry. OMICS 6:217–234

Beckman M, Enot DP, Overy DP, Draper J (2007) Representation, comparison and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. J Agric Food Chem 55:3444–3451

Colquhoun IJ (2007) Use of NMR for metabolic profiling in plant systems. J Pestic Sci 32:200–212

Dunn WB, Ellis DI (2005) Metabolomics: current analytic platforms and methodologies. Trends Anal Chem 24:285–294

Etalo DW, De Vos CHR, Joosten MHAJ, Hall RD (2015) Spatially resolved plant metabolomics: some potentials and limitations of laser-ablation electrospray ionization mass spectrometry metabolite imaging. Plant Physiol 169:1424–1435

Fiehn O (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp Funct Genomics 2:155–168

Fuzfai ZF, Katona E, Kovacs E, Molnar-Perl I (2004) Simultaneous identification and quantification of the sugar, sugar alcohol and carboxylic acid contents of sour cherry, apple, and ber fruits as their trimethylsilyl derivatives, by gas-chromatography-mass spectrometry. J Agric Food Chem 52:7444–7452

Graham SF, Amigues E, Migaud M, Browne RA (2009) Application of NMR based metabolomics for mapping metabolite variation in European wheat. Metabolomics 5:302–306

Grivet JP, Delort AM (2009) NMR for microbiology: in vivo and in situ applications. Prog Nucl Magn Reson Spectrosc 54:1–53

Gulberg J, Jonsson P, Nordstrom M, Moritz T (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. Anal Biochem 331:283–295

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

Hall D (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol 169:453–468

Hall D, Brouwer ID, Fitzgerald MA (2008) Plant metabolomics and its potential application for human nutrition. Physiol Plant 132:162–175

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. Proc Natl Acad Sci U S A 101:10205–10210

Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443

Johnson HE, Broadhurst D, Goodacre R, Smith AR (2003) Metabolic fingerprinting of salt-stressed tomatoes. Phytochemistry 62:919–928

Kaiser K, Benner R (2012) Characterization of lignin by gas chromatography and mass spectrometry using a simplified CuO oxidation method. Anal Chem 84:459–464

Kikuchi J, Shinozaki K, Hirayama T (2004) Stable isotope labeling of Arabidopsis thaliana for an NMR-based metabolomics approach. Plant Cell Physiol 45:1099–1104

Kim H, Ralph J (2014) A gel-state 2D-NMR method for plant cell wall profiling and analysis: a model study with the amorphous cellulose and xylan from ball-milled cotton linters. RSC Adv 4:7549–7560

Kim HK, Choi YH, Verpoorte R (2011) NMR-based plant metabolomics: where do we stand, where do we go? Trends Biotechnol 29:267–275

Kose F, Weckwerth W, Linke T, Fiehn O (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices. Bioinformatics 17:1198–1208

Le Gall G, Colquhoun IJ, Defernez M (2004) Metabolite profiling using 1H NMR spectroscopy for quality assessment of green tea, Camellia sinensis (L.). J Agric Food Chem 52:692–700

Lindinger C, Pollien P, de Vos RCH, Tikunov Y, Hageman JA, Lambot C, Fumeaux R, Voirol-Baliguet E, Blank I (2009) Identification of ethyl formate as a quality marker of the fermented off-note in coffee by a nontargeted chemometric approach. J Agric Food Chem 57:9972–9978

Moco S, Bino RJ, Vorst O, Vehoeven HA, De Groot J, Van Beek TA, Vervoort J, De Vos CHR (2006) A liquid chromatography-mass spectrometry-based metabolome database for tomato. Plant Physiol 141:1205–1218

Moing A, Aharoni A, Biais B, Rogachev I, Meir S, Brodsky L, Allwood JW, Erban A, Dunn WB, Kay L, de Koning S, de Vos RCH, Jonker H, Mumm R, Deborde C, Maucourt M, Bernillon S, Gibon Y, Hansen TH, Husted SGR, Kopka J, Schjoerring JK, Rolin D, Hall RD (2011) Extensive metabolic cross-talk in melon fruit revealed by spatial and developmental combinatorial metabolomics. New Phytol 190:683–696

Morreel K, Saeys Y, Dima O, Lu F, Van de Peer Y, Vanholme R, Ralph J, Vanholme B, Boerjan W (2014) Systematic structural characterization of metabolites in Arabidopsis via candidate substrate–product pair networks. Plant Cell 26:929–945

Murch SJ, Rupasinghe HP, Goodenowe D, Saxena PK (2004) A metabolomic analysis of medicinal diversity in Huang-qin (Scutellaria baicalensis Georgi) genotypes: discovery of novel compounds. Plant Cell Rep 23:419–425

Nakabayashi R, Saito K (2013) Metabolomics for unknown plant metabolites. Anal Bioanal Chem 405:5005–5011

Oksman-Caldenty KM, Inze D, Oresic M (2004) Connecting genes to metabolites by a systems biology approach. Proc Natl Acad Sci U S A 101:9949–9950

Pitthard V, Finch P (2001) GC-MS analysis of monosaccharides mixtures as their diethyldithioacetal derivatives: application to plant gums used in art works. Chromatographia 53:317–321

Prinsen E, Van Laer S, Oden S, Van Onckelen H (2000) Auxin analysis. Humana Press Inc, Totowa, NJ

Rasmussen B, Cloarec O, Tang H, Staerk D, Jaroszewski JW (2006) Multivariate analysis of integrated and full-resolution 1H-NMR spectral data from complex pharmaceutical preparations: St John's Wort. Planta Med 72:556–563

Ratcliffe RG, Shachar-Hill Y (2005) Revealing metabolic phenotypes in plants: inputs from NMR analysis. Biol Rev 80:27–43

Ratsimba V, Garcia Fernandez JM, Defaye J, Nigay H, Voilley A (1999) Qualitative and quantitative evaluation of mono- and disaccharides in D-fructose, Dglucose and sucrose caramel by gas-liquid chromatography-mass spectrometry Di-D-fructose dianhydrides as tracers of caramel authenticity. J Chromatogr A 844:283–293

Reo N (2002) NMR-based metabolomics. Drug Chem Toxicol 25:375–382

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23:131–142

Roessner U, Willmitzer L, Fernie A (2001) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems Identification of phenocopies. Plant Physiol 127:749–764

Roos G, Roseler C, Buter KB, Simmen U (2004) Classification and correction of St John's wort extracts by nuclear magnetic resonance spectroscopy, multivariate data analysis and pharmacological activity. Planta Med 70:771–777

Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489

Seger C, Sturm S (2007) Analytical aspects of plant metabolite profiling platforms: current standings and future aims. J Proteome Res 6:480–497

Theodoridis GA, Gika HG, Want EJ, Wilson ID (2012) Liquid chromatography-mass spectrometry based global metabolite profiling: a review. Anal Chim Acta 711:7–16

Tikunov Y, Lommen A, De Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics large-scale profiling of tomato fruit volatiles. Plant Physiol 139:1125–1137

Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima JI, Awazuhara M, Inoue E, Takahashi H, Goodenowe DB, Kitayama M, Noji M, Yamazaki M, Saito K (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. Plant J 42:218–235

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

von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S (2004) Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol 134:548–559

Weckwerth W, Fiehn O (2002) Can we discover novel pathways using metabolomic analysis? Curr Opin Biotechnol 13:156–160

Zhang J, Wang X, Yu O, Tang J, Wan X, Fang C (2011) Metabolic profiling of strawberry (Fragaria × ananassa Duch) during fruit development and maturation. J Exp Bot 62:1103–1118

Zubarev RA, Makarov A (2013) Orbitrap mass spectrometry. Anal Chem 85:5288–5296

Noninvasive Methods to Support Metabolomic Studies Targeted at Plant Phenolics for Food and Medicinal Use

Oksana Sytar, Marek Zivcak, and Marian Brestic

Contents

Abstract Metabolomics has emerged as an important tool in many disciplines, including research of plant resources for food and pharmaceutical use. Despite the development of modern, high-throughput methods, the analyses are still relatively costly and laborious. In this chapter, we present the noninvasive fluorescence-based methods, typically used in plant phenomics, which may serve as early steps in metabolomic screening targeted at nutritionally and pharmaceutically important phenolic compounds. The presented results of in situ measurements in a high number of plant species indicate a high interspecific variability, which seems to be promising for further studies. The principle of the methods, previous applications as well as future possibilities are dealt with.

Keywords Chlorophyll fluorescence • Flavonoids • Anthocyanins • Noninvasive methods

O. Sytar • M. Zivcak • M. Brestic (✉)
Department of Plant Physiology, Slovak Agricultural University,
Tr. A. Hlinku 2, 949 76 Nitra, Slovak Republic
e-mail: marian.brestic@uniag.sk; marian.brestic@gmail.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_18

1 Introduction

Metabolomics, one of the "omic" sciences in systems biology, is the global assessment and validation of endogenous small-molecule metabolites within a biologic system. It represents a comprehensive method for metabolite assessment that involves measuring the overall metabolites of biological samples (Nicholson and Lindon 2008). Metabolomics was derived from the field of biochemistry and involves the analysis (usually high throughput or broad scale) of small-molecule metabolites and polymers. The foundations of metabolomics are descriptions of biological pathways and current metabolomic databases (Kanehisa et al. 2002), and are frequently based on well-characterized biochemical pathways. On a more applied level, the bioinformatics of metabolomics involves the identification and characterization of a broad range of metabolites through reference to quantitative biochemical analysis. Although this field is relatively new, there have been significant recent advances (Fernie 2003), and there is scope for many direct applications in plant biotechnology (Fiehn et al. 2000; Roessner et al. 2002).

The metabolomics analyses, in a strict sense, cover mostly invasive biochemical analytical methods, based mostly on mass spectrometry (Lei et al. 2011), nuclear magnetic resonance (NMR) spectroscopy (Kim et al. 2011), or many other chromatographic techniques (Hall 2011). Although the expenses for metabolic studies are decreasing, the metabolomics analyses remain laborious and relatively expensive (Hall 2006, 2011). In general, metabolomics covers two basic approaches: (1) the non-targeted approach is aimed at determination of as many compounds as possible in the samples. This approach may lead to discoveries of new active molecules, but it is slower and more expensive. In contrast, (2) the targeted approach is aimed at a single or relatively narrow, well-defined group of compounds (e.g., amino acids and phenolics), and this approach is research aimed at practical applications, e.g., for food or pharmaceutical purposes (Verpoorte et al. 2005).

In targeted plant metabolomics, the main goal is to achieve a high throughput. Therefore, there is often an initial desire for a rapid prescreening of the samples. This is especially the case when dealing with large sample numbers, where only a limited number of individuals might be expected to be different (Verhoeven et al. 2006). This is the case when searching for valuable genetic resources (e.g., those having a high content of desired compounds) within natural populations or population obtained by crosses or mutagenesis. For that reason, there is a need for high-throughput, noninvasive methods, which can screen rapidly a high number of plants in vivo, resulting in a low number of the most promising accessions, which are further examined by the standard metabolomic techniques.

Depending on the target metabolites, there are several possible noninvasive techniques, which may be used in early steps of metabolomics research. In this chapter, we aim at a technique based on simultaneous measurements of multispectrally induced chlorophyll fluorescence (hereinafter denoted as multiplex measurements). This technique, though not yet widely used, has become more popular due to introduction of commercially available devices in the last decade. Although the technique

has several alternative applications, we demonstrate the usefulness of the techniques in research targeted at phenolic compounds in the aboveground parts of plants, mostly flavonoids and anthocyanins, in medicinal herbs and plants used in human nutrition.

2 Plants as a Source of Phenolic Compounds for Medicinal and Food Use

Herbs are used in many domains, including medicine, nutrition, flavoring, beverages, dyeing, repellents, fragrances, and cosmetics (Djeridane et al. 2006). Many species have been recognized to have medicinal properties and beneficial impact on health, e.g., antioxidant activity, digestive stimulation action, antiinflammatory, antimicrobial, hypolipidemic, antimutagenic effects and anticarcinogenic potential (Wojdyło et al. 2007). Crude extracts of herbs and spices, and other plant materials rich in phenolics are of increasing interest in the food industry because they retard oxidative degradation of lipids and thereby improve the quality and nutritional value of food.

Plants need phenolic compounds for pigmentation, growth, reproduction, resistance to pathogens and for many other functions. These compounds form one of the main classes of secondary metabolites and several thousand (among them over 8150 flavonoids) different compounds have been identified with a large range of structures: monomeric, dimeric, and polymeric phenols.

Phenolic substances are mainly deposited in leaves or bark (in case of trees or bushes), together with other waste products (Yanishlieva 2001). Phenolic substances also serve as protectants against bacterial pathogens (*Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Bacillus cereus*, and *Escherichia coli*) (Haq et al. 2011; Ghasemi et al. 2011). Phenolic compounds are potential antioxidants because there is a relation between antioxidant activity and presence of phenols in common vegetables and fruits (Cai et al. 2004; Fu et al. 2011). A positive linear correlation between antioxidant capacities and total phenolic contents implied that phenolic compounds in tested 50 medicinal plants could be the main components contributing to the observed activities. The results showed that *Geranium wilfordii*, *Loranthus parasiticus*, *Polygonum aviculare*, *Pyrrosia sheaeri*, *Sinomenium acutum*, and *Tripterygium wilfordii* possess the highest antioxidant capacities and total phenolic content among 50 plants tested, and could be rich potential sources of natural antioxidants (Gan et al. 2010).

All plant leaves contain phenolic substances, sometimes in high amounts. e.g., in tea leaves. Most phenolic antioxidants are flavonoids, such as catechins, of different structures and antioxidant activities (Pokorný 2000). Natural antioxidants from herbs and spices were recently reviewed (Yanishlieva et al. 2006). They are well known because of their phenolic content. On the other hand, less familiar materials, such as sweetgrass (*Hierochloe odorata* Wahl.), used for flavoring certain types of alcoholic beverages, were also found very effective as an antioxidant (Zainuddin et al. 2002). Some plant leaves are applied to food, and are widely used. Perhaps the

best known antioxidants from plant leaves are leaves from green and black (fermented) tea. The major polyphenolic constituents present in green tea are epicatechin, epigallocatechin, epicatechin-3-gallate, and epigallocatechin-3-gallate. In addition to the small amount of catechins, black tea contains thearubigins and theaflavins, which are the polymerized forms of catechin monomers and are the major components formed during enzymatic oxidation and the fermentation process (Kaushik et al. 2010).

Catechin content is particularly high in green tea, but its infusions are less concentrated. During the tea fermentation, about 50 % phenolic substances are converted into black tea pigments—theaflavins and thearubigins—nevertheless, they have some residual activity, and black teas, prepared at higher concentrations, are a very good source of antioxidants in Europe (Pokorný 2007).

The main anthocyanins in fruits are glycosides of different anthocyanidins, mainly cyanidin, that are widespread and commonly contribute to the pigmentation of fruits. Citrus fruits differ in their flavonoid profiles from other fruit species, containing flavanones and flavones (hesperidin and naringenin) that are not common in other fruits (Robards and Antolovich 1997).

Flavonoids have been reported to possess a wide range of activities in the prevention of common diseases, including CHD, cancer, neurodegenerative diseases, gastrointestinal disorders, and others (González-Gallego et al. 2007). Flavonols are found in high concentrations in onions, apples, red wine, broccoli, tea, and *Ginkgo biloba* (Tulyathan et al. 2005). The most common in the American diet are quercetin (70 %), kaempferol (16 %), and myricetin (6 %) (AAFC, Canada's Functional Food and Natural Health Products Industry 2007).

Among cereal plants, the major source of polyphenols is buckwheat. Particularly buckwheat has gained its fame due to its broad spectrum of flavonoids characterized by health benefits, i.e., cholesterol reduction (Kayashita et al. 1997), tumor inhibition (Chan 2003), hypertension regulation (Ma et al. 2006), and control of inflammation, carcinogenesis (Ishii et al. 2008), and diabetes (Kawa et al. 2003). Buckwheat-based products such as noodles, pancakes, and buckwheat corn muffins are consumed in many countries most especially in China, Japan, Korea, Nepal, and European countries.

Epidemiological studies have repeatedly shown an inverse association between the risk of chronic human diseases and the consumption of polyphenol-rich diet (Scalbert et al. 2005; Arts and Hollman 2005). The phenolic groups in polyphenols can accept an electron to form relatively stable phenoxyl radicals, thereby disrupting chain oxidation reactions in cellular components (Clifford 2000). It is well established that polyphenol-rich foods and beverages may increase plasma antioxidant capacity. This increase in the antioxidative capacity of plasma following the consumption of polyphenol-rich food may be explained by the presence of reducing polyphenols and their metabolites in plasma, by their effects upon concentrations of other reducing agents (sparing effects of polyphenols on other endogenous antioxidants), or by their effect on the absorption of pro-oxidative food components, such as iron. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2835915/-R1 (Scalbert et al. 2005). Consumption of antioxidants has been associated with reduced levels

of oxidative damage to lymphocytic DNA. Similar observations have been made with polyphenol-rich food and beverages, indicating the protective effects of polyphenols (Vitrac et al. 2002). There are increasing evidences that as antioxidants, polyphenols may protect cell constituents against oxidative damage and, therefore, limit the risk of various degenerative diseases associated with oxidative stress (Luqman and Rizvi 2006; Pandey et al. 2009). The role of polyphenols in human health is still a fertile area of research. Based on our current scientific understanding, polyphenols offer great hope for the prevention of chronic human diseases (Pandey and Rizvi 2009).

The content of some important secondary metabolites in the some medicinal herbs of the family Asteraceae and the plants with the highest content of any nutrient are known. *Echinacea purpurea* L. is one of the most important medicinal herbs and is a native perennial of *Asteraceae* grown in North America, which is used pharmacologically and for aesthetic enjoyment. In 2005, *Echinacea* products ranked among the top botanical supplements sold in the USA. Varieties of *Echinacea purpurea* all contain similar main ingredients including caffeic acid derivatives, alkamides, flavonoids, essential oils, and polyacetylenes, the medicinal activities of which are yet to be exactly identified with the corresponding diseases they are effective against (Tzu-Tai Lee et al. 2010). We confirm that this is a well-known medicinal plant, but we would recommend to screen different representatives of the *Asteraceae* for their flavonoid content and compare it to the flavonoid content of *Echinacea purpurea*. For example, *Tagetes patula* Linn. (*Asteraceae*) is a source of essential oil, limonene, and caryophyllene from leaves and roots (Patel 2012). The leaves or oil from seeds of representatives of *Helianthus* sp. (*Asteraceae*) is known as a source of provitamin A (Leporatti and Ivancheva 2003), but the flavonoid content in the leaves of different representatives is still unknown. The leaves and flowering top of the medicinal herb *Calendula officinalis* has been used for their spasmolytic and analgesic effects and for treating inflamed wounds (Leporatti and Ivancheva 2003). At present information about antioxidant activity, phenolic and flavonoid content of *Calendula officinalis* flower extracts is available, but information regarding flavonoid content of the leaves is still missing. So the flavonoid content of the leaves and herbals of known medicinal plants is still unknown and needs screening. Nowadays, in the scientific literature, mostly information about flavonoid and anthocyanin contents of grape berries is presented, but the same information for other kinds of berries is still missing.

3 The Noninvasive Fluorescence-Based Phenomic Method for Determination of Plant Phenolics

Chlorophyll *a* fluorescence represents the reemission of light absorbed by photosynthetic pigments with emission spectra in red to far-red color (600–800 nm), with peaks at ~680 and ~730 nm. The records of chlorophyll fluorescence signals emitted by plants have been used in a broad range of applications for decades

(Kalaji et al. 2014). The techniques based on chlorophyll fluorescence have been widely used in numerous applications, including fast screening of genetic resources of crops (Zivcak et al. 2008a, b, c; Brestic et al. 2012; Brestic and Zivcak 2013), ecophysiological studies (Repkova et al. 2008; Zivcak et al. 2013, 2014a, b), studies of nutrient deficiencies (Kalaji et al. 2014; Zivcak et al. 2014c, d; Galambošová et al. 2014), or mechanistic studies in mutants (Datko et al. 2008; Brestic et al. 2008, 2014, 2015).

Although the emission of chlorophyll fluorescence is directly related to the photochemical activity running on the thylakoid membranes in the chloroplast, the fluorescence signal is strongly influenced also by optical properties of plant tissues not directly related to photochemical processes. It was shown, however, that adjustment of leaf optical properties is not purposeless, but usually serves as a protection for photosynthetic structures. Thus, in addition to others, an important defense mechanism against the deleterious effects of solar radiation involves synthesis of relatively stable compounds that serve as light screens and/or internal traps (Day et al. 1994; Bilger et al. 1997, 2001; Smith and Markham 1998; Cockell and Knowland 1999; Barnes et al. 2000; Merzlyak and Chivkunova 2000; Cerovic et al. 2002; Steyn et al. 2002; Merzlyak and Solovchenko 2002; Pfündel et al. 2006). Depending on concentration in cells and tissues, the protective compounds reduce the fraction of radiation absorbed by light-sensitive cell components, and thereby diminish light-induced damage. Probably the most important position among compounds providing the passive photoprotection (screen) in plants are vascular flavonoids (Bilger et al. 1997; Cockell and Knowland 1999; Barnes et al. 2000; Cerovic et al. 2002; Pfündel et al. 2006) and anthocyanins (Merzlyak and Chivkunova 2000; Steyn et al. 2002; Pfündel et al. 2006), as well as extrathylakoid carotenoids (Merzlyak and Solovchenko 2002; Han et al. 2003; Merzlyak et al. 2005) which, however, have different spectral ranges of action.

The phenolic compounds (flavonoids, hydroxycinnamic acids) have absorption maxima in the UV part of the spectrum. The flavonoids are located either in the epidermal vacuoles, cell walls, or dissolved in epicuticular wax (Shimazaki et al. 1988; Koostra 1994; Strid et al. 1994; Wollenweber and Dietz 1981). They have absorption maxima around 260 nm (isoflavones, flavanones), 320 nm (hydroxycinnamic acids), 260 and 340 nm (flavones), or 360 nm (flavonols) (Jurd 1957; Harborne 1989; Mabry et al. 1970), although the relative importance of the different phenolic compounds as the UV-screen remains an open question (Cerovic et al. 2002).

Based on the strictly UV-absorbing properties, the effects of phenolic compounds on visible light-induced chlorophyll fluorescence is negligible, whereas their presence strongly suppresses the chlorophyll fluorescence emission under UV excitation (Bilger et al. 1997). This phenomenon has been successfully applied for estimation of transmittance of UV radiation by chlorophyll fluorescence (Sheahan 1996; Bilger et al. 1997; Barnes et al. 2000; Burchard et al. 2000; Ounis et al. 2001a). As the experiments confirmed that the phenolic compounds in the epidermis are responsible for most of the UV-absorption of the leaf (Day et al. 1994), the ratio of visible light-excited to UV-excited chlorophyll fluorescence can serve as an

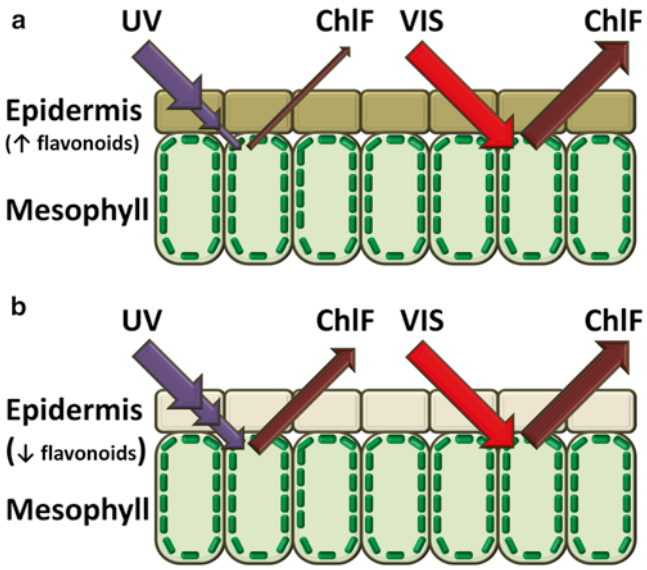


Fig. 1 A schematic drawing of the adaxial part of a leaf cross-section illustrating the principle of the chlorophyll fluorescence (ChlF) method for assessment of UV-absorbing compound content in the sample with high (**a**) and low (**b**) flavonoid content. The thickness of the beams indicates relative intensity

indirect measure of content of the UV-absorbing phenolic compounds in leaves (Cerovic et al. 2002), as shown by the model (Fig. 1).

In parallel, anthocyanins are water-soluble vacuolar pigments of higher plants. They are responsible for red coloration of plant tissues, especially in fruits (Saure 1990; Merzlyak and Chivkunova 2000; Ma and Cheng 2004). Anyway, they can occur also in plant leaves. In many cases, significant accumulation of anthocyanins is induced as a result of environmental stresses such as low temperature, nitrogen and phosphorus deficiencies, UV-B stress, drought, pathogen infections, or due to toxic effects (Harborne 1976; Saure 1990; Chalker-Scott 1999). Anthocyanins absorb strongly in the green region of the spectrum. Gitelson et al. (2001) found that the spectral band around 550 nm (green) was sensitive to anthocyanin content. Thus, similarly to flavonoids, the ratio of red (or blue) light-excited to green light-excited chlorophyll fluorescence can serve as an indirect measure of anthocyanin content in plant samples, as shown in the model (Fig. 2).

In the previous decades, numerous studies examined and confirmed possibility to use the chlorophyll fluorescence signal in the estimation of phenolics and anthocyanins. In addition to self-constructed devices or standard fluorometers combined with external light sources and filters, which were used in the majority of studies, the factory-made special devices for this purpose were also introduced.

A portable UV-A PAM fluorometer (Walz, Germany) device has also been used to measure epidermal UV transmittance of leaves in situ (Bilger et al. 2001). Light

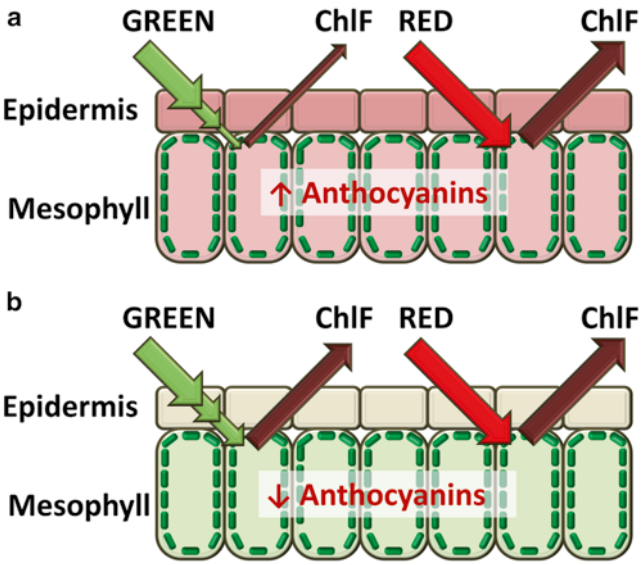


Fig. 2 A schematic drawing of the adaxial part of a plant sample cross-section illustrating the principle of the chlorophyll fluorescence (ChlF) method for assessment of anthocyanin content using simultaneous green and red excitation, in samples with high (**a**) and low (**b**) anthocyanin content. The thickness of the beams indicates relative intensity

emitting diodes (LEDs) produce excitation radiation in the UVA (peak 375 nm) and blue (peak 400 nm) wavelength regions. Chlorophyll fluorescence (ChlF) is measured at wavelengths above 650 nm. The instrument allows screening assessment of compounds only in the UV-A band (at 375 nm), although many flavonoids have absorption peaks at shorter wavelengths, typically in the UV-B region. Chlorophyll fluorescence spectroscopy studies have shown that detection of flavonoids using an excitation wavelength centered at their absorption maximum can be misleading at high concentrations of the compounds (Bidel et al. 2007; Agati et al. 2011). Then by using 375 nm radiation (i.e., the tail of flavonoid absorption) for excitation, detection over a larger range of concentrations is possible.

Epidermal absorbance measured using a UV-A PAM was found to correlate well with the quercetin content of *Brassica oleracea* leaves, determined by HPLC (Hagen et al. 2009). Kolb et al. (2005) compared UV-B transmittance, measured using a Xe-PAM system, with UV-A transmittance measured using a UV-A PAM fluorometer, for *Vitis* spp. and *H. vulgare* leaves. The authors concluded that a good assessment of compounds absorbing in the UV-B band is possible with the UV-A PAM fluorometer once the relationship between the absorbance in the two UV bands has been defined.

The research group of Z. Cerovic (France) developed several devices using the principle of multispectrally induced chlorophyll fluorescence described above (Cerovic et al. 2012). In principle, they introduced two types of devices: leaf clip-based

instrument (commercially available under trademark Dualex, Force-A, France) as well as the non-contact type of instrument (under trademark Multiplex, Force-A, France). While the Dualex system measures only two signals (e.g., UV and VIS-light induced chlorophyll fluorescence), several kinds of Dualex are produced, specialized for estimation of UV-absorbing compounds (flavonoids), anthocyanins, or chlorophylls. In contrast, the Multiplex system measures simultaneously different fluorescence signals after excitation under several spectral regions of light (UV, blue, green, red excitation); thus, this system enables estimating flavonoids, anthocyanins, chlorophylls and any other information from a single measurement. Thanks to the fact that the Multiplex system does not need any leaf clip, it can be used for measurements even in objects other than flat leaves, e.g., fruits, stems, and flowers. This makes this system especially useful for special applications, potentially also in automated systems (as it needs no direct contact with plants).

The use of blue-excited chlorophyll fluorescence (ChlF) as a reference signal by the Xe-PAM and UV-A PAM renders them inapplicable for use with anthocyanin-containing leaves (Pfundel et al. 2007), *Malus* spp. (red Aroma variety) (Hagen et al. 2006), or any other tissue containing blue-absorbing compounds. This problem is circumvented with the Dualex portable instrument (Force-A, Orsay, France: Goulas et al. 2004), which uses excitation bands in the UV-A (375 nm) and in the red (657 nm) rather than in the blue band. ChlF is detected in the near-infrared, at about 730 nm. This instrument facilitates use of leaf disks and small leaves, which are positioned between the excitation head and the detection head. Positive correlations have been found between Dualex epidermal absorbance (the sum of absorbance of both leaf sides) and total flavonoid content obtained using HPLC in *Vitis* spp. leaves but there is no correlation with hydroxycinnamic acid contents (Agati et al. 2008). In a further methodological study, epidermal UV attenuation of *Fraxinus excelsior* (L.) and *Acer platanoides* (L.) was measured using the Dualex fluorometer, spectrophotometry of extracts, and the Folin–Ciocalteu colorimetric method. Close correlations were found between Dualex and spectrophotometric readings, but all correlations were highly species-specific (Barthod et al. 2007).

The Multiplex (MPX) fluorimetric sensor (Force-A, Orsay, France) is described in detail elsewhere (Tavarini et al. 2008). The development of this sensor has allowed for detection of flavonoids over a large sample area, up to 8 cm in diameter. The sensor uses one UV source (375 nm) and three LED matrices emitting at 470, 516, and 635 nm (Ben Ghozlen et al. 2010). The Multiplex has three detection channels in the blue-green, red, and far-red spectral regions: the latter two detecting 680–690 and 730–780 nm fluorescence, respectively, corresponding to the two emission peaks of chlorophyll. It measures fluorescence emitted by chlorophyll, in the red (RF) and far-red (FRF) spectral regions, under excitation with different light-emitting diode (LED) sources in the UV (375 nm) and visible (blue at 450 nm, green at 515 nm and red at 630 nm). Three synchronized photodiode detectors recorded fluorescent yellow, red and infrared fluorescence (Bursal et al. 2013). Transferring plants, acclimated for 3 weeks to 30 % sunlight, to 85 % sunlight, caused an exponential increase in the flavonoid index, which reached a maximum within 10 days (Agati et al. 2011). These data demonstrate the value of multispectrally induced

chlorophyll fluorescence methods in noninvasive monitoring of acclimation dynamics. This method cannot substitute for the analytical chemistry of purified individual flavonoids. Yet the flexibility and potential to perform a large number of parallel measurements with minimal sample preparation repeated on the same target over a period of time makes multispectrally induced chlorophyll fluorescence an important tool for studies of UV response of flavonoids.

Flavonoid content is determined as the FLAV index, which is derived from UV absorption properties of flavonoids. The intensity of the chlorophyll fluorescence (ChlF) emitted by a sample depends on the amount of excitation light able to reach the Chl pigment, that is on the transmittance of the epidermis at the excitation wavelength (Ghozlen et al. 2010). Differences in chlorophyll fluorescence (ChlF) caused by excitation in the UV (ChlF-UV) and the visible bands (ChlF-VIS) are used to assess the amount of UV-absorbing compounds located in the epidermis. The ChlFUV/ChlF-VIS ratio represents the UV transmittance (i.e., lack of absorbance) of the epidermis. Flavonols in the epidermis can attenuate part of the incident radiation in the UV-A region before this can reach the Chl molecules. Since the long wavelengths of visible light are not absorbed by flavonols, ChlF excited by green or red wavelengths can be considered as a reference signal. To calculate the relative amount of UV-absorbing compounds, two ChlF signals under UV (ChlF-UV) and red excitation (ChlF-RED) can be used to obtain an index proportional to the flavonoid content in the leaf epidermis (Mendes Novo et al. 2012):

$$\mathrm{FLAV} = \log\left(\mathrm{ChlF\ \ RED} / \mathrm{ChlF\ \ UV}\right).$$

Measured levels of flavonoids in different plant species are expressed in relative units (RU), because the ratio of the optical densities of the two gives a dimensionless number. The multispectrally induced chlorophyll fluorescence technique allows continuous assessment of changes in UV-absorbing compound accumulation over time in one particular sample. Therefore, this is a powerful tool for comparative studies and high throughput screening. The limitations of this method are that only epidermal compounds are detected and that it cannot give information on the flavonoid composition in deeper layers. However, due to the dominant function of flavonoids in UV-protection, in most of the plants, the dominant contribution of flavonoids located in the epidermis, in which the method will be useful. This creates the scope for further research.

The same principle can be applied for the estimation of anthocyanin content using the red-to-green excitation ratio of chlorophyll fluorescence. This ratio was confirmed by the screening of leaves (Pfündel et al. 2007) and apples (Merzlyak et al. 2008) by anthocyanins. Good linear correlation was found when the decadic logarithm of the red-to-green excitation ratio of chlorophyll fluorescence was used for estimation of the anthocyanin content of the leaf epidermis or fruit skin, like olive (Agati et al. 2005) and grape (Agati et al. 2007). Thus, the relative anthocyanin ratio is calculated as follows (Cerovic et al. 2007):

$$\mathrm{ANTH} = \log \mathrm{FER\_RG} = \log\left(\mathrm{ChlF\ \ RED} / \mathrm{ChlF\ \ GREEN}\right).$$

However, in specific conditions of a very dark fruits (e.g., blue grapes) it was shown that in mature grapes the ANTH index, as defined above, decreases with increasing anthocyanin content due to a larger decrease (shielding) of the red-induced signal compared to green-induced signal (Cerovic et al. 2007). Therefore, for viticulture, the additional index was proposed, which increases with anthocyanin content:

$$\mathrm{ANTH\_GR} = \log \mathrm{FER\_GR} = \log\left(\mathrm{ChlF\ \ GREEN} / \mathrm{ChlF\ \ RED}\right) - \mathrm{constant} + 1.$$

It is simply the ANTH index mentioned above, inverted and normalized by subtracting a constant corresponding to the green berry (without anthocyanin).

In addition, another index denoted as FERARI (i.e., fluorescence excitation ratio anthocyanin relative index) was proposed that has a good positive correlation with anthocyanin content of red grape berries (Ben Ghozlen et al. 2010):

$$\mathrm{FERARI} = \log\left(5000 / \mathrm{FRF\_R}\right).$$

As a single fluorescence signal is used, this index should be used only when the measurements are done using a constant-distance. This is not useful in situ (in field conditions), but it can be used in an automated system, like for berries on trays.

The multispectrally induced chlorophyll fluorescence technique requires the use of equivalent, and representative plant tissues. Many ecophysiological studies compare measurements of mature leaves, sampled from two-thirds up from the base of the plant. Thus, the first (i.e., oldest or proximal) and youngest leaves (i.e., terminal or distal) usually are not sampled to avoid any potential effect of development on the flavonoid profile (Laitinen et al. 2002). Besides foliage age, a plant's age is also important when assessing the flavonoid profile of leaves, stems, or roots (Laitinen et al. 2002; Keski-Saari and Julkunen-Tiitto 2003). Moreover, the leaves need to be consistently sampled from the same side of each plant to assure flavonoid content also varies within individual leaves, stems, or roots. In grasses, flavonoids accumulate during leaf expansion. However, this is not a homogenous process. Higher levels of flavonoids accumulate at the leaf tip, which receives more radiation than the base (Cartelat et al. 2005). Changes in flavonoid profile during leaf expansion are genotype-specific. Laitinen et al. (2002) reported that concentrations of flavonoid aglycones decreased, but flavonoid-glycoside content increased with the phenological advance from mature buds to young leaves in 30-year-old *Betula pendula* trees. The many environmental stressors also impact on plant development (Potters et al. 2007), potentially leading changes in flavonoid profile.

UV screening caused by UV-absorbing substances, fruit quality, leaf tissue structure, and disease symptoms can be detected by multispectral fluorescence methods. At the same time it is good to evaluate results regarding phases of development, the type of leaves (thickness, water content, and other morphological parameters), and possible influence of stress factors. Given these complexities, the development of a sampling strategy is a critical component of comparative flavonoid analysis.

4 Experimental Chlorophyll Fluorescence-Based Approaches to Determine Phenolics in Plants

Indirect measurements of UV-absorbing capacity using dual-wavelength induced chlorophyll fluorescence has been used in numerous studies (Table 1). While the early works were aimed mostly at theoretical aspects, in the last decade, we can observe an increase of the works aimed at practical application of the methods in screening of plant material. Anyway, the most of works dealing with UV-induced chlorophyll fluorescence were aimed at the leaves of field crops (mostly wheat and barley), woody plants (mostly leaves of *Vitis*), much less on vegetables and leaves of medicinal herbs. Even less numerous were the papers dealing with fruits, dealing mostly with grapes and apples (Table 1).

The works listed above (Table 1) were differing both in techniques used for non-invasive analysis and in the aims of the research. Fr example, UV-induced chlorophyll fluorescence and visible light-induced fluorescence emitted by leaves have been proposed as useful indicators of plant physiological status under stress conditions for barley plants. Total fresh biomass decreased with decreasing N supply, whereas the leaf content of soluble phenolic compounds increased. This increase in leaf phenolic compounds observed with limiting N supply was accompanied by large increases of the visible light-induced fluorescence intensity and VIS-ChlF/UV-ChlF ratio (i.e., FLAV ratio) of leaf sections and, to a lower extent, by a decrease of the leaf epidermal transmittance of UV radiation (as estimated by the ratio of ChlF intensities induced by UV and blue excitation) (Mercure et al. 2004). Thanks to possibility of simultaneous assessment of chlorophyll content (using red to far-red fluorescence ratio) and UV-absorbing capacity (FLAV index) on the same leaf spot, the nitrogen balance index (NBI) was proposed, which correlated well with leaf optical properties strongly influenced by nitrogen supply, thus making this parameter useful in nitrogen nutrition management (Cerovic et al. 2012; Galambošová et al. 2014). Of course, this method is not related to the screening of plants for food and medicinal use, therefore, these results, even of a high practical importance, are not discussed further in this chapter.

Another possible application is in the detection of effects of infections or in screening for plant material resistant to pathogens due to high phenolic content in the epidermis. Belasque et al. (2008) employed fluorescence spectroscopy to detect stress caused by citrus canker (bacterial disease caused by *Xanthomonas citri–X. axonopodis* pv. *citri*) and mechanical injury. A portable fluorescence spectroscopy system was taken to the greenhouse and the measurement probe was placed 2 mm above the leaf (attached to greenhouse plants) for collecting data from different samples during the period of study (60 days). The samples of leaves collected from the field (detached leaves) as well as leaves from greenhouse plants (attached leaves) were analyzed for 60 days under four different conditions: leaves with no stress, leaves with mechanical stress, leaves with disease, and leaves with disease and mechanical stress. The three ratios used were: (1) ratio between fluorescence intensity at 452 and 685 nm, (2) ratio between fluorescence intensity at 452 and 735 nm,

Table 1 List of papers reporting noninvasive analysis of UV-absorbing compounds (flavonoids, phenolics) in different plant species

Type of plants	Species	References
Field crops and vegetables (leaves)	*Brassica oleracea* L. var. *capitata* (cabbage)	Pfündel et al. 2007
	Brassica oleracea L. var. *italica (broccoli)*	Bengtsson et al. 2006
	Hordeum vulgare L. (barley)	Ounis et al. 2001a; Cerovic et al. 2002; Wagner et al. 2003; Mercure et al. 2004; Kolb et al. 2005; Pfündel et al. 2007; Shaw et al. 2014
	Lactuca sativa L. (lettuce)	Pfündel et al. 2007
	Nicotiana spp. (tobacco)	Ounis et al. 2001a, b; Cerovic et al. 2002; Demkura et al. 2010;
	Phaseolus vulgaris L. (bean)	Cerovic et al. 2002; Louis et al. 2006;
	Pisum sativum L. (pea)	Ounis et al. 2001a, b;
	Secale cereale L. (rye)	Burchard et al. 2000
	Solanum tuberosum L. (potato)	Bélanger et al. (2006)
	Spinacia oleracea L. (spinach)	Cerovic et al. 2002;
	Solanum tuberosum L. (potato)	Bélanger et al. (2006)
	Triticum aestivum L. (wheat)	Ounis et al. 2001a, b; Cartelat et al. 2005; Cerovic et al. 2005; Bürling et al. 2013
	Vicia faba L. (faba bean)	Bilger et al. 2007
Wild grown herbs (leaves, green parts of plants)	*Aeonium haworthii* Webb & Berth	Pfündel et al. 2007
	African forage plants	Scogings et al. 2014
	Arabidopsis thaliana Heyn.	Cerovic et al. 2002; Bilger et al. 2007;
	Centella asiatica (L.) *Urban*	Muller et al. 2013
	Crassula ovata (Mill.) Druce	Pfündel et al. 2007
	Ligustrum vulgare L	Agati et al. 2011
	Myrtus communis L.	Agati et al. 2011
	Oenothera stricta Ledeb.	Barnes et al. 2008
	Oxyria digyna (L.) Hill	Bilger et al. 2007
	Phyllirea latifolia L	Agati et al. 2011
	Regnellidium diphyllum L.	Pfündel et al. 2007
	Rumex longifolius DC.	Bilger et al. 2007
	Sedum telephium L.	Pfündel et al. 2007
	Verbascum thapsus L.	Barnes et al. 2008

(continued)

Table 1 (continued)

Type of plants	Species	References
Woody plants (leaves)	*Acer platanoides* L	Barthod et al. 2007
	Betula pendula L.	Morales et al. 2011
	Callicarpa bodinieri H. Léveillé	Demotes-Mainard et al. 2008
	Fagus sylvatica L.	Lenk and Buschmann 2006
	Fraxinus excelsior L.	Barthod et al. 2007
	Kolkwitzia amabilis Graebner	Pfündel et al. 2007
	Lagerstroemia indica L.	Demotes-Mainard et al. 2008
	Parthenocissus tricuspidata Planc.	Pfündel et al. 2007
	Quercus petraea L.	Louis et al. 2009; Meyer et al. 2009
	Viburnum sp.	Pfündel et al. 2007; Demotes-Mainard et al. 2008
	Vitis spp.	Kolb et al. 2001; Pfündel 2003; Kolb et al. 2005; Pfündel et al. 2007; Agati et al. 2008; Latouche et al. 2013
Fruit species (fruits)	*Actinidia chinensis* Planch.	Pinelli et al. 2013
	Malus domestica L. (apple)	Hagen et al. 2006; Merzlyak et al. 2008; Betemps et al. 2012
	Vitis spp. (grape)	Kolb et al. 2003; Lenk et al. 2007; Bélanger et al. 2008; Ghozlen et al. 2010

and (3) ratio between fluorescence intensity at 685 and 735 nm. The studies reported the potential of fluorescence spectroscopy for disease detection and discrimination between the mechanical and diseased stress. Lins et al. (2009) conducted field experiments to discriminate citrus canker-stressed leaves from chlorotic (caused by *Xylella fastidiosa* bacteria) and healthy leaves. In addition, they conducted leaf detachment experiments to monitor effect of time (up to 12 h) on the fluorescence of detached leaves using fluorescence spectroscopy. In their study, two indices/figures of merit were used to assess the difference between healthy and citrus canker-infected leaves.

It can be expected also increase of applications of dual or multiple wavelength-induced chlorophyll fluorescence in assessment of the effects of abiotic stresses. Chlorophyll fluorescence (ChlF) excitation spectra were measured to assess the UV-sunscreen compounds accumulated in fully expanded leaves of three woody species belonging to different chemotaxons (i.e., *Morus nigra* L., *Prunus mahaleb* L., and *Lagerstroemia indica* L.), grown in different light microclimates. The logarithm of the ratio of ChlF excitation spectra (logFER) between two leaves acclimated to different light microclimates was used to assess the difference in epidermal absorbance. It increased with increasing solar irradiance intercepted for the three species. This epidermal localization of UV-absorbers was confirmed by the removal

of the epidermis. It was possible to simulate epidermal absorbance as a linear combination of major phenolic compounds identified in leaf methanol extracts by HPLC-DAD. Under UV-free radiation conditions, shaded leaves of *M. nigra* accumulated chlorogenic acid. Hydroxybenzoic acid derivatives and hydroxycinnamic acid derivatives greatly increased with increasing PAR irradiance under the low UV-B conditions found in the greenhouse. These traits were also observed for the hydroxycinnamic acid of the two other species. Flavonoid (FLAV) accumulation started under low UV-A irradiance, and became maximal in the adaxial epidermis of sun-exposed leaves outdoors (Agati et al. 2007).

Anyway, there is relatively poor information on the possible application of the proposed methods in screening of plants for medicinal and food use. Therefore, we made a large initial screening in collection plants grown in the same, sun-exposed conditions within open-air expositions of botanical garden (Botanical Garden Slovak Agricultural University in Nitra, Slovakia, central Europe). The results of noninvasive measurements using Multiplex-3 sensor (Force-A, France) show relatively high interspecific variation in epidermal UV-shielding, associated with flavonoid content (UV-absorbing compounds) in plant epidermis, expressed in relative units of *Flav* ratio (Table 2).

Assuming that FLAV index correlates well with leaf flavonoid content (Cerovic et al. 2012), observed differences in *Flav* indicate more than 5 times higher flavonoid content in a group with very high *Flav* values compared to plants belonging to the group with the lowest observed values.

To analyze the distribution of values within relative species, we compared in detail the medicinal plants (herbs) belonging to three families: *Rosaceae*, *Asteraceae*, and *Lamiaceae* with numerous observed plant species. The light-exposed side of leaves (20–25 records in each species) was measured by the fluorimetric sensor. The flavonoid content (value of *Flav* parameter) has been evaluated in 13 plant species of the family *Asteraceae* (Fig. 3).

Among the monitored plants of the family *Asteraceae* the maximum value of flavonoids has been found in the leaves of sunflower (*Helianthus multiflorus*, 1.65 RU), which was significantly higher compared to all other representatives of the family *Asteraceae*. Lowest *Flav* value has been observed in the leaves of marigold (*Calendula officinalis*, 0.14 RU). Raal and Kirsipuu (2011) addressed the spectrophotometric determination of the amount of flavonoids in different varieties of *Calendula* species, finding relatively high intraspecific variability, which was not associated with different colors of inflorescence.

For another two representatives of *Helianthus* sp.—*Helianthus annus* and *Helianthus tuberosus* flavonoid content reached only 65 % and 62 % compared to the flavonoid content in the leaves of *Helianthus multiflorus*. These results indicate that flavonoid content can be significantly different, even in the representatives of one genus. At the same time it has been found that leaves contain the most allelochemicals because those in the roots are lost by leaching and those from stems are translocated (Kamal 2011). The highest flavonoid content has been found in the leaves of *Echinops ritro*, too (Fig. 1). But the flavonoid content in the leaves of *Echinops ritro* was 25 % lower compared to the flavonoid content in the leaves of

Table 2 List of plants ranked according to relative values of UV-induced chlorophyll fluorescence measured in situ

Level of UV-absorbance	Species
Very low (Flav < 0.5)	*Malva sylvestris*, *Calendula officinalis*, *Trifolium repens*, *Reseda lutea*, *Lactuca serriola*, *Amaranthus retroflexus*, *Medicago sativa*, *Oxalis purpurea*, *Portulaca oleracea*, *Cerasus avium*, *Lavandula angustifolia*, *Stachys byzantine*, *Ambrosia artemisiifolia*, *Convolvulus arvensis*, *Rosmarinus officinalis*, *Nicotiana alata*, *Borago officinalis*, *Plantago lanceolata*, *Echinochloa crus-galli*
Low (Flav 0.5–0.75)	*Crocosmia massoniorum*, *Amaranthus cruentus*, *Mentha spicata*, *Robinia pseudoacacia*, *Althaea cannabina*, *Parthenocissus quinquefolia*, *Ficus carica*, *Lilium hemerocallis*, *Helianthus annuus*, *Zinnia elegans*, *Prunella grandiflora*, *Acer tataricum*, *Potentilla recta*, *Ipomoea batatas*, *Hibiscus syriacus*, *Mentha spicata*, *Echinacea purpurea*, *Plantago media*, *Acer pseudoplatanus*, *Eryngium planum*, *Laurocerasus officinalis*, *Carpinus betulus*, *Nigella damascene*, *Colutea arborescens*, *Solidago canadensis*, *Sedum aizoon*, *Fallopia dumetorum*, *Verbena hybrid*, *Plantago arenaria*, *Clematis vitalba*, *Scaevola aemula*, *Melissa officinalis*, *Helianthus tuberosus*, *Wisteria sinensis*
Medium (Flav 0.75–1)	*Hibiscus rosa-sinensis*, *Aristolochia clematitis*, *Dictamnus albus*, *Platanus orientalis*, *Ginkgo biloba*, *Saponaria officinalis*, *Impatiens balsamina*, *Yucca gloriosa*, *Sophora japonica*, *Celosia argentea*, *Phytolacca Americana*, *Liatris spicata*, *Salvia officinalis*, *Negundo aceroides*, *Calluna vulgaris*, *Armoracia rusticana*, *Betula pendula*, *Larix decidua*, *Tagetes patula*, *Dahlia pinnata*, *Hippophaë rhamnoides*, *Anthericum ramosum*, *Berberis vulgaris*, *Asparagus officinalis*, *Gleditschia japonica*, *Imperata cylindrical*, *Acanthus mollis*, *Anchusa officinalis*, *Koelreuteria paniculata*, *Coleus blumei*, *Gaillardia grandiflora*, *Robinia neomexicana*, *Juniperus virginiana*, *Convalaria majalis*, *Thuja occidentalis*, *Quercus robur*, *Arundo donax*, *Laburnum anagyroides*, *Tilia cordata*, *Aquilegia vulgaris*, *Alchemilla mollis*, *Althaea armeniaca*, *Olea europaea*
High (Flav 1–1.25)	*Eriobotrya japonica*, *Cleome spinosa*, *Rosa rubiginosa*, *Datura stramonium*, *Bergenia cordifolia*, *Citrus* sp., *Acer palmatum*, *Digitalis grandiflora*, *Agrimonia eupatoria*, *Viburnum opulus*, *Cotoneaster horizontalis*, *Cotinus coggygria*, *Castanea sativa*, *Fagus sylvatica*, *Limonium gmelinii*, *Rheum rhabarbarum*, *Alnus glutinosa*, *Clematis integrifolia*, *Corylus avellana*, *Aesculus hippocastanum*, *Lysimachia clethroides*, *Securigera varia*, *Fraxinus excelsior*, *Corylus avellana*, *Rudbeckia fulgida*, *Fraxinus excelsior*, *Hedera helix*, *Acer pseudoplatanus*, *Scabiosa stellate*, *Viburnum rhytidophyllum*, *Weigela florida*, *Rosa canina*, *Fraxinus pennsylvanica*, *Pinus strobus*, *Swida alba*, *Pinus cembra*, *Fraxinus ornus*, *Ricinus communis*
Very high (Flav > 1.25)	*Echinops ritro*, *Mahonia aquifolium*, *Ailanthus altissima*, *Sorbus domestica*, *Yucca aloifolia*, *Nerium oleander*, *Pinus armandii*, *Hypericum calycinum*, *Calycanthus floridus*, *Helianthus multiflorus*, *Catalpa bignonioides*

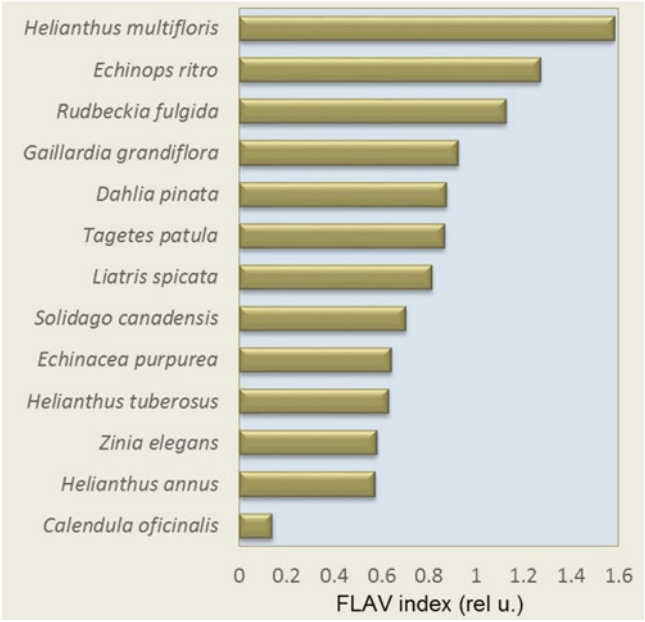


Fig. 3 Content of flavonoids in the leaves investigated plant species of the family *Asteraceae* (modified according to Sytar et al. 2015)

Helianthus multiflorus. Echinops species *(Echinops echinatus*, *Echinops niveus*, and *Echinops integrifolius*) are known to elaborate flavonoids (Singh and Pandey 1994; Singh et al. 2006; Senejoux et al. 2013).

The high flavonoid content (1.13 RU) has been found in the leaves *Rudbeckia fulgida* (orange coneflower), a species of flowering plant of the family *Asteraceae*, native to eastern North America. The data regarding flavonoid or phenolic content has been mostly presented for *Echinacea purpurea*, but not for *Rudbeckia fulgida* (orange coneflower). However, in the leaves of orange coneflower, the flavonoid content was two times higher compared to the leaves of purple coneflower (*Echinacea purpurea)*, where the flavonoid content in the leaves reached the value of 0.64 RU. Previous studies have shown that *Echinacea purpurea* extracted with a 55 % ethanol at 55 °C contained 86.0 ± 4.6 mg quercetin equivalent g^{-1} of flavonoid content (Lee et al. 2010).

It is known that the activity of antioxidants and their content (phenolics, flavonoids) in different plants are dependent on extracting solvents such as absolute methanol, ethanol, acetone, and ethyl acetate which also makes a difference during estimation of plant antioxidants (Anokwuru et al. 2011). In the leaves of French marigold (*Tagetes patula*) a flavonoid content of 0.87 RU has been estimated which represents the middle level of flavonoid content among the assessed plant species of the family *Asteraceae*. Rop et al. 2012 investigated the flavonoid content in the flowers of French marigold *Tagetes patula* and found 1.90 kg rutin g^{-1} of fresh

weight. The authors concluded that flavonoid synthesis may be conditional on the genetic origins of various kinds of flowers.

At the leaves of *Gaillardia grandiflora*, the blanket flowers, it has been estimated the flavonoid content 0.93 RU which was less on 44 % compared to the highest flavonoid content in the leaves of *Helianthus multiflorus.* Cytotoxic compounds with flavonoid nature from the leaves of *Gaillardia aristata* Pursh. have been established. Ten phenolic compounds were isolated from the leaves of *Gaillardia aristata* by applying bioassay guided fractionation (Maha and Zeinab 2012). The flavonoid content in the leaves of *Dahlia pinnata* (0.88 RU), *Liatris spicata*, dense blazing star (0.82 RU), *Solidago canadensis*, Canada golden-rod (0.70 RU) was on the average level among the investigated plant representatives of the family *Asteraceae*. The flavonoid pigments of *Liatris spicata* were isolated and identified as 3-glucoside, 3-rutinoside, and 3-glucoside-7-rhamnoside of quercetin (Kagan 1968). *Solidago canadensis* is typical of a flavonoid-rich herb and flavonol quercetin and its glycosides quercitrin and rutin have been found as major constituents of ethanolic extracts (Apáti et al. 2006). Air dried herbs of *Solidago canadensis* were extracted with methanol and HPLC analysis revealed phenolics (chlorogenic acid, caffeic acid, kaempferol-3-O-α-L-rutinoside (nicotiflorin), quercetin-3-O-β-D-rutinoside (rutin), quercetin-3-O-β-D-galactoside (hyperoside), quercetin-3-O-β-D-glucoside (isoquercitrin), quercetin-3-O-β-D-rhamnoside (quercitrin), kaempferol-3-O-α-L--rhamnoside (afzelin), and quercetin from solidaginis herba (Apati et al. 2002).

In the family *Lamiaceae* flavonoid content in eight medicinal plant species has been investigated (Fig. 4). The range of individual flavonoids in the monitored species of the family *Lamiaceae* is 0.40–0.90 RU. The lowest levels of flavonoids (0.40 RU) have been detected in the leaves of *Lavandula angustifolia* and in the known herb rosemary (*Rosmarinus officinalis*) lower levels of flavonoids (0.42 RU) have been observed. Many investigations of flavonoid content were done with air-dried herb extracts of different medicinal plants, but presently, there is no clear information about total flavonoid content in the leaves or other parts of herbs. For example Yoo et al. (2008) with colorimetric determination investigated a number of flavonoids in the leaves of 17 selected herbs. Chamomile (*Chamaemelum nobilis* L.), rosehip (*Rosa rubiginosa*), hawthorn (*Crataegus pinnatifida*), lemon verbena (*Aloysia triphylla*), green tea (*Camellia sinensis* L.), and black tea (*Camellia sinensis* L.) have the highest flavonoid content in the herb extracts among the investigated species. Among all the 17 herbs evaluated rosemary (*Rosmarinus officinalis*) herb extract got the second highest flavonoid content (448.4 mg catechin 100 g^{-1} fresh weight). In the lavender (*Lavandula angustifolia* Mill.) mean flavonoid content has been found (390.4 mg catechin 100g^{-1} fresh weight) (Yoo et al. 2008).

An average value of flavonoids in the leaves of lemon balm (*Melissa officinalis*) (0.72 RU). Atanassova et al. (2011) identified the highest flavonoid content in lemon balm (45.06 mg catechin 100 g^{-1} dry weight) and in the sage (*Salvia officinalis*) found average amounts of flavonoids (27.54 mg catechin 100 g^{-1} dry weight) among the evaluated herbs. The result of Multiplex measurements recorded in the sage leaves showed the second highest flavonoid content (0.84 RU) after coleus (*Coleus blumei* = syn. *Solenostemon scutellarioides*), where the flavonoid content of leaves

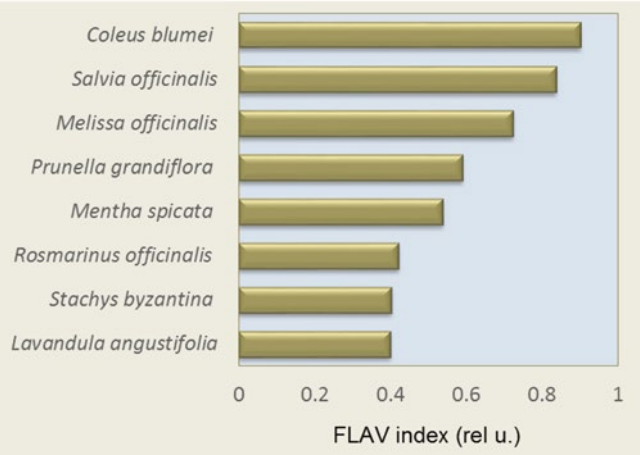


Fig. 4 Content of flavonoids in the leaves investigated plant species of the family *Lamiaceae* (modified according to Sytar et al. 2015)

was 0.90 RU. Leaves of sage (*Salvia officinalis*) and *Coleus blumei* with purple leaves were shown to be significantly higher in flavonoid content compared to all other experimental species of *Lamiaceae.*

Coleus blumei (*Lamiaceae*) is an ornamental plant, growing all over the world with an enormous number of different cultivars that vary in color; the incredible foliage is interesting, with arrays of color combinations unmatched by other species and shape of the leaves (Karen 1999). *Coleus blumei* has an interesting ability to change its leaf color depending on the intensity of the sunlight (Garcia et al. 1973). The flavonoid contents determined in the dried *Lamiaceae* leaf extracts are presented in Table 5. The amount of flavonoid in dried leaves ranged from 0.18 to 15.21 mg QE/g dried samples. There were significant differences in flavonoid contents among the six *Lamiaceae* leaf extracts. *Coleus blumei*—purple leaf extract contained significantly higher amount of flavonoid compared to other leaf extracts, while the extracts of *Coleus amboinicus* had significantly the lowest flavonoid content. Briefly amount of flavonoids was the highest in *Coleus blumei*—purple leaves, followed by *Coleus blumei*—red leaves, *Coleus amboinicus*, *Coleus aromaticus*, and *Pogostemon cablin* (Khattak and Taher 2011).

Family *Rosaceae* is a well-known family for the presence of anticancer, antioxidant compounds. Previously other constituents as flavonoids, phenolic acids (Palme et al. 1996; El-Mousallamy et al. 2000), tannins (Wang and Ji 2008) have been found in this family.

In Fig. 5 the flavonoid content in ten monitored species of the family *Rosaceae* is shown. The highest value of flavonoids was determined in the leaves of dog rose (*Rosa canina*) (1.18 RU). The flavonoid content in the rose family (*Rosaceae*) ranged from 0.38 to 1.18 RU and decreased in the following order: dog rose (*Rosa canina*) 1.18 RU > cotoneaster (*Cotoneaster horizontalis*) 1.06 RU > agrimony

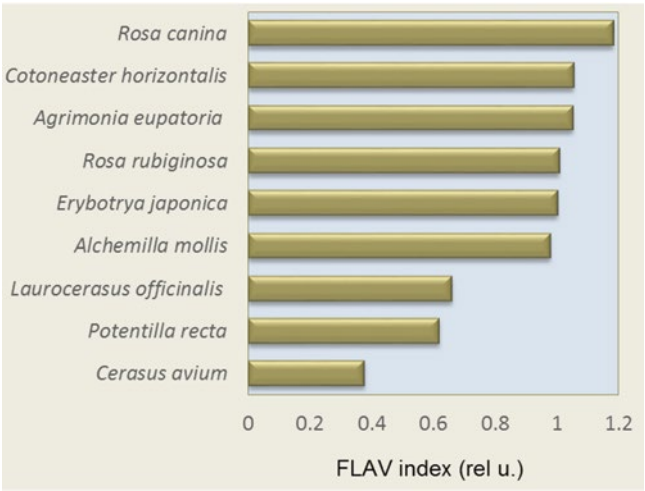


Fig. 5 Content of flavonoids in the leaves investigated plant species of the family *Rosaceae* (modified according to Sytar et al. 2015)

(*Agrimonia eupatoria*) 1.05 RU > rusty rose (*Rosa rubiginosa*) 1.01 RU > Japanese loquats (*Eriobotrya japonica*) 1.00 RU > soft lady's mantle (*Alchemilla mollis*) 0.98 RU > laurel medicinal (*Laurocerasus officinalis*) 0.66 RU > direct cinquefoil (*Potentilla recta*) 0.62 RU > bird cherry (*Cerasus avium*) 0.38 RU.

The second highest flavonoid content has been observed in the leaves of dog rose (*Rosa canina*)—1.18 RU (Fig. 3). Nowak and Gawlik-Dziki (2007) estimated amount of flavonols by HPLC (quercetin and myricetin) in the extracts of the leaves of several species of the genus *Rosa. Rosa canina* showed lower flavonol content of 8.53 mg g^{-1} dry matter, while *Rosa rubiginosa* showed the second highest value for flavonols, 18.27 mg g^{-1} dry matter. The authors suggest that extracts of Rosa spp. could be used as natural antioxidants and as part of functional food (Nowak and Gawlik-Dziki 2007).

The content of flavonoids in the leaves of *Cotoneaster horizontalis* has been shown second highest position among investigated representatives of family *Rosacea.* Quantitative determination of the total polyphenols and flavonoids of aerial parts of *Cotoneaster horizontalis* Decne. (family *Rosaceae*) was performed colorimetrically using Folin–Ciocalteu and aluminum trichloride methods has been found the concentrations of flavonoids and flavonol contents expressed as rutin equivalent were 6.8 ± 0.76 and 2.2 ± 00 mg g^{-1} plant extract rutin equivalent respectively. HPLC analysis of total flavonoids showed the presence of three flavonoids (quercetin, naringenin, and luteolin), and luteolin was the major compound (9.20 mg/100 g dried plant) (Shaza et al. 2012). The phytochemical analysis of the ethanolic extract of the branches of *Cotoneaster horizontalis* revealed the presence of β-carotene, ascorbic acid, less amounts of α-tocopherol and amygdalin (vitamin B17). Information about content and composition of flavonoids in the leaves of *Cotoneaster horizontalis* in the literature data is missing nowadays.

Average values of flavonoids have been recorded in the herb rusty rose (*Rosa rubiginosa*) (1.01 RU) and in the leaves of agrimony (*Agrimonia eupatoria* 1.05 RU) (Fig. 3). Yoo et al. (2008) determined a flavonoid content of 400.5 mg catechin 100 g^{-1} of fresh matter by a colorimetric method in the leaves of rusty rose, which belonged to the group of evaluated herbs with a high total flavonoid content. Kubínová et al. (2012) have estimated the highest flavonoid content (3.5 mg quercetin g^{-1} dry weight) in a methanolic extract of the flowering aerial parts of agrimony among the five species of the genus *Agrimonia* evaluated (Kubínová et al. (2012).

A high flavonoid content has been found in the leaves of *Agrimonia eupatoria.* This species is rich in chemical constituents (flavonoids, tannins, aromatic acids, triterpenes, coumarins, terpenoids, glycosides, and vitamins B and K) that can mediate antioxidant, antibacterial, and anti-inflammatory effects (Xu et al. 2005). Dried aerial parts of *A. eupatoria* (leaves, stems, and flowers) were used for preparation of aqueous and methanol extracts and study antitumor potential of these extracts. Chemical analysis of *A. eupatoria* extracts (aqueous and methanol) revealed that the plant was positive for several secondary metabolites. Both aqueous and methanol extracts were positive for flavonoids, alkaloids, tannins, and glycosides and negative for saponins. Flavonoids were further identified by thin layer chromatography (TLC), and as suggested by RF values of the separated extracts, the aqueous extract contained myricetin, azoleatin, vitexin, and isoorientin, while the methanol extract contained kaempferol, quercetin, isorhamnetin, and myricetin (Ad'hiah et al. 2013).

Loquat (*Eriobotrya japonica* Lindl.) is a perennial subtropical fruit tree and application of multiplex fluorimetric sensor for analysis flavonoid content revealed a high flavonoid content in the leaves of loquat. Many studies demonstrated that large amounts of flavonoids and phenolics were found in the fruit and leaf of loquat (Jung et al. 1999; Hong et al. 2008; Louati et al. 2003), and both the methanol extract of loquat leaf and its individual fraction exhibited a strong antioxidant capacity (Jung et al. 1999). Methanol had the highest extraction efficiency among five solvents, followed by ethanol. Considering safety and residue, ethanol is better as an extraction solvent. The average flavonoid and phenolic content of loquat flower of five cultivars were 1.59 ± 0.24 and 7.86 ± 0.87 mg g^{-1} DW, respectively, when using ethanol as an extraction solvent. The contents of both bioactive components in flowers at different developmental stages and in various floral tissues clearly differed, with the highest flavonoid and phenolic content in flowers of stage 3 (flower fully open) and petals, respectively (Zhou et al. 2011). So these data again confirm our suggestion regarding the important role of prescreening with application multiplex fluorimetric sensor with an aim to estimate flavonoid content in different plant parts which can make the choice of plant parts and medicinal herb species with high flavonoid content easier for extraction and further identification of flavonoids.

Among investigated representatives of family *Rosaceae* many of them got higher flavonoid content in the leaves compared to the representatives of family *Lamiaceae*. For example in the *Alchemilla mollis* leaves has been found higher flavonoid content (0.98 RU) than in *Coleus blumei* and *Salvia officinalis* (0.84 RU). Different studies

showed that the flavonoids compounds present in the plant are responsible for the pharmacological activity of Lady's mantle (Jonadet et al. 1986). The aerial flowering parts of *Alchemilla mollis* were collected within phenophase—full blossoming, air-dried and used for preparation of ethanolic extracts. Further purification by RP-18 CC led to the isolation of eight flavonoid glycosides: *cis* and *trans*-tiliroside, rhodiolgin, hyperoside, isoquercitrin, miquelianin, sinocrassoside D2, and gossypetin-3-*O*-b-D-galactopyranosyl-7-*O*-α-L-rhamnopyranoside (Trendafilovaa et al. 2011).

This is fact for phytotherapy needs and research investigation mostly were used aerial flowering parts or aerial parts of plants, but for better knowledge's what exist in each plant parts (leaves, stems, flowers) application multiplex fluorimetric sensor with aim to estimate flavonoid content is new stage in the area of modern plant physiology and phytotherapy.

It was indicated that both the water extract of cherry stem (WECS) and ethanol extract of cherry stem (EECS) have both antioxidant and antiradical properties, and there is a correlation between these properties and the phenolic and flavonoid contents. Quantities of quercetin, α-tocopherol, pyrogallol, ascorbic acid, and other phenolic acids were detected by high performance liquid chromatography and tandem mass spectrometry (LC-MS/MS) (Bursal et al. 2013). In this experimental work with application Multiplex fluorimetric sensor for screening flavonoid content in the leaves of different representatives of family *Rosaceae*, the leaves of bird cherry (*Cerasus avium* L.) have a low flavonoid content. An average flavonoid content has been found in the leaves of laurel medicinal (*Laurocerasus officinalis*) and direct cinquefoil (*Potentilla recta*), and the flavonoid content in their leaves was higher than that in the leaves of *Cerasus avium* L. by more than 25 %.

5 Applications of Noninvasive Approaches to Determine Anthocyanins

The anthocyanins represent the largest group of water soluble pigments in the plant; they are important antioxidants in human foods, with broad applications for food, pharmaceutical, and cosmetic industries. Therefore, the noninvasive methods can be very useful in practical applications. Therefore, the chlorophyll fluorescence-based assessment of anthocyanins in plants and fruits (described above) was examined in numerous studies (Table 3).

It is, however, evident that the number of papers dealing with anthocyanins is lower than in flavonoids (Table 1). Moreover, the most of studies is aimed at evaluation and improvement of noninvasive measurements of anthocyanins in grapes for producing wine. Quite well is also covered the estimation in apples, but there is a very low information on other fruit species.

Anyway, the use of non-destructive methods to assess fruit quality is of wide interest (Butz et al. 2005), since they represent a rapid tool for fruit sorting, evaluating storage conditions and monitoring fruit ripening. Portable sensors for fruit ripening

Table 3 List of papers reporting noninvasive analysis of green light-absorbing compounds (anthocyanins) in different plant species

Type of plants	Species	References
Field crops and vegetables (leaves)	*Brassica oleracea* L. var. *capitata* (cabbage)	Pfündel et al. 2007
	Hordeum vulgare L. (barley)	Pfündel et al. 2007; Shaw et al. 2014
	Lactuca sativa L. (lettuce)	Pfündel et al. 2007
	Phaseolus vulgaris L. (bean)	Louis et al. 2006
Wild grown herbs (leaves, green parts of plants)	*Aeonium haworthii* Webb & Berth	Pfündel et al. 2007
	Cornus alba L.	Gitelson et al. 2001
	Cotoneaster alaunica Golite	Gitelson et al. 2001
	Crassula ovata (Mill.) Druce	Pfündel et al. 2007
	Regnellidium diphyllum L.	Pfündel et al. 2007
	Sedum telephium L.	Pfündel et al. 2007
	Sedum telephium L.	Pfündel et al. 2007
Woody plants (leaves)	*Acer platanoides* L	Gitelson et al. 2001
	Rosa spp. (flowers)	Terfa et al. 2013
	Fraxinus excelsior L.	Barthod et al. 2007
	Kolkwitzia amabilis Graebner	Pfündel et al. 2007
	Parthenocissus tricuspidata Planc.	Pfündel et al. 2007
	Viburnum sp.	Pfündel et al. 2007
	Vitis spp.	Pfündel et al. 2007; Matese et al. 2013
Fruit species (fruits)	*Elaeis guineensis* Jacq. (oil palm)	Hazir et al. 2012a, b
	Fragaria spp. (strawberry)	Wulf et al. 2008; Fan et al. 2011
	Malus domestica L. (apple)	Hagen et al. 2006; Merzlyak et al. 2008; Kuckenberg et al. 2008; Betemps et al. 2012
	Vitis spp. (grape)	Kolb et al. 2003; Lenk et al. 2007; Bélanger et al. 2008; Ghozlen et al. 2010; Cerovic et al. 2008; Le Moigne et al. 2010; Bramley et al. 2011; Baluja et al. 2012a, b; Maoz et al. 2014

directly applied in the field to predict the best harvest period would be particularly appealing. One possible technique is based on chlorophyll fluorescence detection and the anthocyanins screening effect on the excitation light used for the measurement: the larger the anthocyanins concentration in the berry skin, the lower the chlorophyll fluorescence signal; this is the parallel of the parameter FERARI, mentioned before.

The comparison of the spectral characterisation of the tomato fruit surface pigments from the immature to over-ripe stage, using spectroscopy techniques based on visible fluorescence emission upon excitation in the same or ultraviolet spectral regions has been verified the spectral band for optimal conditions for fruit harvesting using non-destructive techniques. The pattern of pigment composition changed markedly during ripening and showed progressive disappearance of chlorophyll with a concomitant increase in carotenoids until the fully ripe stage. The main fluorescence spectral features belonging to anthocyanins, flavonoids, carotenoids and chlorophyll a after excitation of skin tomato pigments at different laser wavelengths was identified. In comparing, the fluorescence spectral ratios at the excitation wavelength 266 nm, significant differences were obtained for the spectral ratios of chlorophyll/flavonoids and carotenoids/chlorophyll (Lai et al. 2007).

Anthocyanins in olive (*Olea europaea* L.) fruits at different degrees of pigmentation were assessed nondestructively by measuring chlorophyll fluorescence (ChlF). As expected, the in vivo anthocyanins absorption maximum increased in intensity going from less to more mature olives and was higher in the sun-exposed olive side with respect to the sun-shaded side (Agati et al. 2005).

Grape phenolic maturity is usually assessed by destructive wet chemistry in the laboratory. Yet, for precision agriculture or continuous monitoring of maturation, more rapid and non-destructive methods are needed. The chlorophyll fluorescence measurement results are present mostly for wine grape (Cerovic et al. 2008). Bramley et al. 2011 has demonstrated the utility of chlorophyll fluorescence screening for characterizing within-vineyard variation in grape berry anthocyanins. Anthocyanins are formed via the bonding of glycosides and anthocyanidins such as pelargonin, chrysanthemin, and delphinidin. They are polyphenol flavonoid compounds synthesized by the shikimic acid pathway, and mainly present in vacuoles in plant cells (Moyer et al. 2002). Recently, the topic of estimation of anthocyanins in grapes has been well developed, thanks to a French research group of Z. Cerovic (Ghozlen et al. 2010; Cerovic et al. 2008, 2012, etc.). On the other hand, there is a lack of information on the estimation of anthocyanins in other kinds of fruits and berries.

Therefore, we have done also preliminary examination of different kinds of berries and their varieties in situ (Tables 4–6). We present the relative values of anthocyanins (ANTH) together with simultaneously recorded values of flavonoid units (Flav). Based on the chlorophyll fluorescence measurements, it was found that the highest flavonoid content in the studied species of fruit was recorded in the blackberry (*Rubus plicatus*) 1.662 RU and the lowest (but still high) level of total flavonoids were observed in black currant (*Ribes nigrum*) with a value of 1.032 RU. Red raspberry (*Rubus idaeus*) and strawberry (*Fragaria*×*ananassa* D.) got high flavonoid and anthocyanin contents. At the same time the highest anthocyanin index was reported in the fruits of *Grossularia albus* (1.135 RU).

Fruits and leaves from different cultivars of thornless blackberry (*Rubus* sp.), red raspberry (*Rubus idaeus* L.), black raspberry (*Rubus occidentalis* L.), and strawberry (*Fragaria*×*ananassa* D.) were analyzed for total antioxidant capacity, total phenolic content, and total anthocyanin content. Blackberries and strawberries had

Table 4 Flavonoid and anthocyanin contents in the berries of families *Grossulariaceae* and *Rosaceae*

Type of plant	FLAV index	ANTH index
Blackberry (*Rubus plicatus*)	1.662 ± 0.025	0.140 ± 0.003
Red raspberry (*Rubus idaeus*)	1.548 ± 0.014	0.130 ± 0.003
Strawberry *(Fragaria* sp.)	1.484 ± 0.014	0.129 ± 0.003
Red currant *(Ribes sativum)*	1.259 ± 0.019	0.156 ± 0.006
Jostaberries (*R. nidigrolaria*)	1.222 ± 0.038	0.125 ± 0.004
White currant (*Ribes rubrum*)	1.219 ± 0.014	0.125 ± 0.002
Grossularia albus	1.135 ± 0.016	0.156 ± 0.004
Black currant (*Ribes nigrum*)	1.032 ± 0.034	0.117 ± 0.003

Table 5 Flavonoid and anthocyanin contents in the fruits of different white currant cultivars

Cultivar	FLAV index	Anthocyanin index
Viktoria	1.267 ± 0.066	0.132 ± 0.017
Primus	1.275 ± 0.052	0.125 ± 0.009
Blanka	1.188 ± 0.033	0.124 ± 0.006

Table 6 Flavonoid and anthocyanin contents in the fruits of different *red currant* cultivars

Cultivar	FLAV index	Anthocyanin index
Trent	1.338 ± 0.045	0.168 ± 0.018
Heineman	1.225 ± 0.080	0.151 ± 0.023
Red lake	1.215 ± 0.063	0.149 ± 0.017

the highest antioxidant activity values during the green stages, whereas red raspberries had the highest antioxidant activity at the ripe stage. Total anthocyanin content increased with maturity for all the three species of fruits. The results also showed a linear correlation between total phenolic content and antioxidant activity for fruits and leaves. Of the ripe fruits tested, on the basis of wet weight of fruit, cv. Jewel black raspberry and blackberries may be the richest source for antioxidants (Wang and Lin 2000). These results have been confirmed with our prescreening results with chlorophyll fluorescence measurements of flavonoids and anthocyanins which also showed high flavonoid and anthocyanin content in blackberry (*Rubus plicatus*) berries.

The antioxidant capacity of black currant, blueberry, raspberry, red currant, and cranberry extracts was determined using the FRAP assay which shown that complex spectrum of anthocyanins was the major contributor to the antioxidant capacity of black currants and blueberries, whereas the lower antioxidant capacity of red currants and cranberries was due mainly to a reduced anthocyanin content. Raspberries also had lower anthocyanin content than black currants and blueberries, but there was only a slight decline in the antioxidant capacity (Borges et al. 2010).

Anthocyanins and proanthocyanidins were characterized by HPLC-ESI-MS/MS coupled with a diode array and/or fluorescent detector in seven cultivars of *Ribes nigrum* (black currant) and *Ribes rubrum* (red currant, Red Lake), six cultivars of *Ribes grossularia* (gooseberries), *Aronia melanocarpa* (chokeberry), and Sambucus nigra (elderberry). All the seven cultivars of *Ribes nigrum* (black currant) and *Ribes rubrum* (red currant, Red Lake) had similar anthocyanin profiles of 14 detected anthocyanins, but individual anthocyanin concentrations varied slightly (Wu et al. 2004).

The study of antioxidants content in the different strawberry cultivars showed the important role played by the genetic background on the chemical and antioxidant profiles of strawberry fruits. Significant differences were found between genotypes for the total antioxidant capacity and for all tested classes of compounds (Tulipani et al. 2008).

Total flavonoid content in the experimental cultivars of white currants (*Ribes sativa*) ranged from 1.19 to 1.28 RU. On the basis of total flavonoids the cultivars can be listed in the following order: Primus > Victoria > Blanka. The total anthocyanin content in the experimental cultivars of white currants (*Ribes sativa*) ranged from 0.12 to 0.13 RU. On the basis of total anthocyanins the cultivars can be listed in the following order: Viktoria > Primus > Blanka (Table 5).

The total flavonoid content in the experimental cultivars of red currants (*Ribes rubrum*) ranged from 1.22 to 1.34 RU. On the basis of total flavonoids the cultivars can be listed in the following order: Trent > Heineman > Red Lake. Total anthocyanin content in the experimental cultivars of red currants (*Ribes rubrum*) ranged from 1.14 to 1.17 RU (Table 6). On the basis of total flavonoids the cultivars can be listed in following order: Red Lake > Heineman > Trent. The difference in the anthocyanin and flavonoid content can be explained by presence of another compounds of phenolic nature in the berry pericarp (skin) (Braidot et al. 2008).

The presence of antioxidant compounds can be considered as a quality parameter for edible fruits. In the genus Ribes, quercetin is the main compound as in gooseberry, red currant, and black currant. Ellagic acid is the main phenolic compound in the berries of the genus Rubus (red raspberry, Arctic bramble, and cloudberry) and genus Fragaria (strawberry). Our data suggest that berries have a potential as good dietary sources of quercetin or ellagic acid (Häkkinen et al. 1999).

Phenylpropanoid and flavonoid compounds scanning microscopy in ecological biochemistry of usually accumulate in the central vacuoles of guard cells phenolic plant metabolites, and epidermal cells as well as subepidermal cells of leaves (Moskowitz and Hradzina 1981; Schnabl et al. 1986, 1989) and shoots (Ozimina 1979). Furthermore, some compounds were found to be from rhizomes of horsetail covalently linked to plant cell walls (Strack et al. 1988; Schnitzler et al. 1996), others occur in waxes (Schmutz et al. 1994) or on the external surfaces of plant organs (Cuadra and Harborne 1996). Direct microscopic observation of phenolic compounds is restricted to anthocyanin-containing tissues, where the target compounds are colored red, purple, or blue. Several classes of phenolic compounds, for example, hydroxycinnamic acids, coumarins, stilbenes, and styrylpyrones (Veit et al. 1995; Gorham 1995), are strongly autofluorescent

when irradiated with UV or blue light. Therefore, fluorescence microscopy is a powerful tool for studying tissue localization of these metabolites. To investigate non-colored and non-fluorescent phenolic compounds, other techniques such as immunocytochemical detection using specific antibodies (Ibrahim 1992; Grandmaison and Ibrahim 1996), histochemical staining with chromogenic reagents (e.g., lignin with phloroglucinol/HCl, pro-anthocyanins with dimethylamino-cinnamaldehyde (Treutter 1989), or induction of secondary fluorescence (e.g., flavonoid-staining with Naturstoffreagenz A (Vogt et al. 1994; Reinold and Hahlbrock 1997) have also been applied.

In addition to the noninvasive method based on a simple excitation of plant sample with light pulses of different wavelengths, several modifications of the method have also been developed and they are tested at present. First is the nondestructive application of laser-induced fluorescence spectroscopy, which has been used for quantitative analyses of phenolic compounds in small samples, such as fruits of strawberry (Wulf et al. 2008) or citrus (Lins et al. 2009), or, alternatively, in large scale, for remote sensing assessment of canopies or plants (Ounis et al. 2001a, b). Another promising application of the dual (multiple) wavelength induced chlorophyll fluorescence is the fluorescence imaging (Lenk and Buschmann 2006; Lenk et al. 2007), which can be used either to identify the spatial heterogeneity in distribution of flavonoids/anthocyanins on the plant, or, importantly, can be applied within automated phenomic facilities, which can be used in high-throughput screening of plant samples in laboratory or in field.

6 Conclusions

The phenolic compound-targeted metabolomic research can be streamlined by prescreening of plant resources using noninvasive methods based on multiwavelength induced chlorophyll fluorescence records. It has previously been shown that this approach enables to estimate flavonoid and anthocyanin content, which has been successfully applied in several applications, especially in the assessment of qualitative traits in grapes for wine production. However, the method seems to be promising even in the assessment of natural plant sources of phenolic antioxidants; however, there is not enough published data in this area. Our findings on interspecific variability in content of UV-absorbing compounds (flavonoids and anthocyanins) in leaves of numerous plants, especially representatives of families *Asteraceae*, *Lamiaceae*, and *Rosaceae* and berries of family *Grossulariaceae* and *Rosaceae*, confirmed and suggested an important role of prescreening of plants using commercially available fluorimetric sensors. Estimation of flavonoid content in different plants and parts of plants may facilitate selection of species as well as parts of medicinal and food plants with high antioxidant (flavonoids, anthocyanins) content. Thus, for laborious and expensive procedures, such as extraction and further spectrophotometric, TLC, HPLC, LC-MS/MS identification flavonoids composition can be used the most promising candidates only, which may substantially increase the success of selection.

Fluorescence-based sensors can be used directly in the nature or field conditions, with the advantage of taking into account seasonal influences on plant flavonoids, due to the variability of climatic conditions such as rainfall, temperature, and irradiance. The MPx sensor could be easily integrated in online sorting devices using the index of flavonoid content as an additional quality parameter.

Acknowledgement Supported by project "AgroBioTech" of the Operational Programme Research and development, Structural Funds of EU.

References

Ad'hiah AH, Al-Bederi ONH, Al-Sammarrae KW (2013) Cytotoxic effects of *Agrimonia eupatoria* L. against cancer cell lines *in vitro*. J Assoc Arab Univ Basic Appl Sci 14:87–92

Agati G, Pinelli P, Cortes Ebner S, Romani A, Cartelat A, Cerovic ZG (2005) Non-destructive evaluation of anthocyanins in olive (*Olea europaea*) fruits by *in situ* chlorophyll fluorescence spectroscopy. J Agric Food Chem 53:1354–1363

Agati G, Meyer S, Matteini P, Cerovic ZG (2007) Assessment of anthocyanins in grape (*Vitis vinifera* L.) berries using a noninvasive chlorophyll fluorescence method. J Agric Food Chem 55:1053–1061

Agati G, Cerovic ZG, Dalla Marta A, Di Stefano V, Pinelli P, Traversi ML, Orlandini S (2008) Optically assessed preformed flavonoids and susceptibility of grapevine to *Plasmopara viticola* under different light regimes. Funct Plant Biol 35:77–84

Agati G, Cerovic ZG, Pinelli P, Tattini M (2011) Light-induced accumulation of ortho-dihydroxylated flavonoids as nondestructively monitored by chlorophyll fluorescence excitation techniques. Environ Exp Bot 73:3–9

Anokwuru CP, Anyasor GN, Ajibaye O, Fakoya O, Okebugwu P (2011) Effect of extraction solvents on phenolic, flavonoid and antioxidant activities of three nigerian medicinal plants. Nat Sci 9:53–61

Apati P, Szentmihályi K, Balázs A, Baumann D, Hamburger M, Sz. Kristó T, Szőke E, Kéry A (2002) HPLC analysis of the flavonoids in pharmaceutical preparations from Canadian goldenrod (*Solidago canadensis*). Chromatographia 56:65–68

Apáti P, Kéry A, Houghton PJ, Steventon GB, Kite G (2006) *In vitro* effect of flavonoids from *Solidago canadensis* extract on glutathione S-transferase. J Pharm Pharmacol 58:251–256

Arts ICW, Hollman PCH (2005) Polyphenols and disease risk in epidemiologic studies. Am J Clin Nutr 81:317–325

Atanassova M, Georgieva S, Ivancheva K (2011) Total phenolic and total flavonoid contents, antioxidant capacity and biological contaminants in medicinal herbs. J Univ Chem Technol Metall 46:81–88

Baluja J, Diago MP, Goovaerts P, Tardaguila J (2012a) Assessment of the spatial variability of anthocyanins in grapes using a fluorescence sensor: relationships with vine vigour and yield. Precision Agric 13:457–472

Baluja J, Diago MP, Goovaerts P, Tardaguila J (2012b) Spatio-temporal dynamics of grape anthocyanin accumulation in a Tempranillo vineyard monitored by proximal sensing. Aust J Grape Wine Res 18:173–182

Barnes PW, Searles PS, Ballaré CL, Ryel RJ, Caldwell MM (2000) Non-invasive measurements of leaf epidermal transmittance of UV radiation using chlorophyll fluorescence: field and laboratory studies. Physiol Plant 109:274–283

Barnes PW, Flint SD, Slusser JR, Gao W, Ryel RJ (2008) Diurnal changes in epidermal UV transmittance of plants in naturally high UV environments. Physiol Plant 133:363–372

Barthod S, Cerovic Z, Epron D (2007) Can dual chlorophyll fluorescence excitation be used to assess the variation in the content of UV-absorbing phenolic compounds in leaves of temperate tree species along a light gradient? J Exp Bot 58:1753–1760

Bélanger MC, Viau AA, Samson G, Chamberland M (2006) Near-field fluorescence measurements for nutrient deficiencies detection on potatoes (*Solanum tuberosum* L.): Effects of the angle of view. Int J Remote Sens 27:4181–4198

Bélanger MC, Roger JM, Cartolaro P, Viau AA, Bellon-Maurel V (2008) Detection of powdery mildew in grapevine using remotely sensed UV-induced fluorescence. Int J Remote Sens 29:1707–1724

Belasque L, Gasparoto MCG, Marcassa LG (2008) Detection of mechanical and disease stresses in citrus plants by fluorescence spectroscopy. Appl Optics 47:1922–1926

Bengtsson GB, Schöner R, Lombardo E, Schöner J, Borge GIA, Bilger W (2006) Chlorophyll fluorescence for non-destructive measurement of flavonoids in broccoli. Postharvest Biol Tec 39:291–298

Ben Ghozlen N, Moise N, Latouche G, Martinon V, Mercier L, Besançon E, Cerovic ZG (2010) Assessment of grapevine maturity using new portable sensor: Non-destructive quantification of anthocyanins. J Int Sci Vigne Vin 44:1–8

Betemps DL, Fachinello JC, Galarça SP, Portela NM, Remorini D, Massai R, Agati G (2012) Non-destructive evaluation of ripening and quality traits in apples using a multiparametric fluorescence sensor. J Sci Food Agric 92:1855–1864

Bidel LPR, Meyer S, Goulas Y, Cadot Y, Cerovic ZG (2007) Responses of epidermal phenolic compounds to light acclimation: *in vivo* qualitative and quantitative assessment using chlorophyll fluorescence excitation spectra in leaves of three woody species. J Photoch Photobio B 88:163–179

Bilger W, Veit M, Schreiber L, Schreiber U (1997) Measurement of leaf epidermal transmittance of UV radiation by chlorophyll fluorescence. Physiol Plant 101:754–763

Bilger W, Johnsen T, Schreiber U (2001) UV-excited chlorophyll fluorescence as a tool for the assessment of UV protection by the epidermis of plants. J Exp Bot 52:2007–2017

Bilger W, Rolland M, Nybakken L (2007) UV screening in higher plants induced by low temperature in the absence of UV-B radiation. Photoch Photobio Sci 6:190–195

Borges G, Degeneve A, Mullen W, Crozier A (2010) Identification of flavonoid and phenolic antioxidants in black currants, blueberries, raspberries, red currants, and cranberries. J Agric Food Chem 58:3901–3909

Braidot E, Zancani M, Petrussa E, Peresson C, Bertolini A, Patui S, Macrì F, Vianello A (2008) Transport and accumulation of flavonoids in grapevine (*Vitis vinifera* L.). Plant Signal Behav 3:626–632

Bramley RGV, Le Moigne M, Evain S, Ouzman J, Florin L, Fadaili EM, Hinze CJ, Cerovic ZG (2011) On-the-go sensing of grape berry anthocyanins during commercial harvest: development and prospects. Aust J Grape Wine Res 17:316–326

Brestic M, Zivcak M (2013) PSII fluorescence techniques for measurement of drought and high temperature stress signal in plants: protocols and applications. In: Das AB, Rout GR (eds) Molecular stress physiology of plants. Springer, Dordrecht, pp 87–131

Brestic M, Zivcak M, Olsovska K, Repkova J (2008) Functional study of PS II and PS I energy use and dissipation mechanisms in barley wild type and chlorina mutants under high light conditions. In: Allen JA, Gantt E, Golbeck JH, Osmond B (eds) Photosynthesis. Energy from the sun. Springer, Dordrecht, pp 1407–1411

Brestic M, Zivcak M, Kalaji HM, Allakhverdiev SI, Carpentier R (2012) Photosystem II thermostability in situ: environmentally induced acclimation and genotype-specific reactions in Triticum aestivum L. Plant Physiol Biochem 57:93–105

Brestic M, Zivcak M, Olsovska K, Shao HB, Kalaji HM, Allakhverdiev SI (2014) Reduced glutamine synthetase activity plays a role in control of photosynthetic responses to high light in barley leaves. Plant Physiol Biochem 81:74–83

Brestic M, Zivcak M, Kunderlikova K, Sytar O, Shao HB, Kalaji HM, Allakhverdiev SI (2015) Low PSI content limits the photoprotection of PSI and PSII in early growth stages of chlorophyll b deficient wheat mutant lines. Photosynth Res 125(12):151–166

Burchard P, Bilger W, Weissenböck G (2000) Contribution of hydroxycinnamates and flavonoids to epidermal shielding of UV-A and UV-B radiation in developing rye primary leaves as

assessed by ultraviolet-induced chlorophyll fluorescence measurements. Plant Cell Environ 23:1373–1380

Bürling K, Cerovic ZG, Cornic G, Ducruet JM, Noga G, Hunsche M (2013) Fluorescence-based sensing of drought-induced stress in the vegetative phase of four contrasting wheat genotypes. Environ Exp Bot 89:51–59

Bursal E, Köksal E, Gülçin I, Bilsel G, Gören AC (2013) Antioxidant activity and polyphenol content of cherry stem (*Cerasus avium* L.) determined by LC–MS/MS. Food Res Int 51:67–74

Butz P, Hofmann C, Tauscher B (2005) Recent developments in noninvasive techniques for fresh fruit and vegetable internal quality analysis. J Food Sci 70:131–141

Cai YZ, Luo Q, Sun M, Corke H (2004) Antioxidant activity and phenolic compounds of 112 traditional Chinese medicinal plants associated with anticancer. Life Sci 74:2157–2184

Cartelat A, Cerovic ZG, Goulas Y, Meyer S, Lelarge C, Prioul JL, Barbottin A, Jeuffroy MH, Gate P, Agati G, Moya I (2005) Optically assessed contents of leaf polyphenolics and chlorophyll as indicators of nitrogen deficiency in wheat (*Triticum aestivum* L.). Field Crop Res 91:35–49

Cerovic ZG, Moise N, Agati G, Latouche G, Ghozlen NB, Meyer S (2007) New portable optical sensors for the assessment of winegrape phenolic maturity based on berry fluorescence. In: Stafford JV (ed) Precision Agriculture '07 Wageningen Academic Publishers, Wageningen, poster 035, pp 1–6

Cerovic ZG, Ounis A, Cartelat A, Latouche G, Goulas Y, Meyer S, Moya I (2002) The use of chlorophyll fluorescence excitation spectra for the nondestructive *in situ* assessment of UV-absorbing compounds in leaves. Plant Cell Environ 25:1663–1676

Cerovic ZG, Cartelat A, Goulas Y, Meyer S (2005) In-field assessment of wheat-leaf polyphenolics using the new optical leaf-clip Dualex. Precision Agric 5:243–249

Cerovic ZG, Moise N, Agati G, Latouche G, Ben Ghozlen N, Meyer S (2008) New portable optical sensors for the assessment of winegrape phenolic maturity based on berry fluorescence. J Food Comp Anal 21:650–654

Cerovic ZG, Masdoumierd G, Ghozlen NB, Latouche G (2012) A new optical leaf-clip meter for simultaneous non-destructive assessment of leaf chlorophyll and epidermal flavonoids. Physiol Plant 146:251–260

Chalker-Scott L (1999) Environmental significance of anthocyanins in plant stress responses. Photochem Photobiol 70:1–9

Chan PK (2003) Inhibition of tumor growth *in vitro* by the extract of *Fagopyrum cymosum* (fagoc). Life Sci 72:1851–1858

Clifford MN (2000) Chlorogenic acids and other cinnamates. Nature, occurrence, dietary burden, absorption and metabolism. J Sci Food Agric 80:1033–1043

Cockell CS, Knowland J (1999) Ultraviolet radiation screening compounds. Biological Reviews 74:311–345

Cuadra P, Harborne JB (1996) Changes in epicuticular flavonoids and photosynthetic pigments as a plant response to UV-B radiation. Z Naturforsch C 51:671–680

Datko M, Zivcak M, Brestic M (2008) Proteomic analysis of barley (Hordeum vulgare L.) leaves as affected by high temperature treatment. In: Allen JF, Gantt E, Goldbeck JH, Os-mond B (eds) Photosynthesis Energy from the sun. 14th International congress on photosynthesis. Springer, Dordrecht, pp 1523–1527

Day TA, Howells BW, Rice WJ (1994) Ultraviolet absorption and epidermal-transmittance spectra in foliage. Physiol Plant 92:207–218

Demkura PV, Abdala G, Baldwin IT, Ballaré CL (2010) Jasmonate-dependent and-independent pathways mediate specific effects of solar ultraviolet B radiation on leaf phenolics and antiherbivore defense. Plant Physiol 152:1084–1095

Demotes-Mainard S, Boumaza R, Meyer S, Cerovic ZG (2008) Indicators of nitrogen status for ornamental woody plants based on optical measurements of leaf epidermal polyphenol and chlorophyll contents. Sci Hortic 115:377–385

Djeridane A, Yousfi M, Nadjemi B, Boutassouna D, Stocker P, Vidal N (2006) Antioxidant activity of some Algerian medicinal plants extracts containing phenolic compounds. Food Chem 97:654–660

El-Mousallamy AM, Hussein SA, Irmgard M, Nawwar MA (2000) Unusual phenolic glycosides from *Cotoneaster orbicularis*. Phytochemistry 53:699–704

Fan L, Fang C, Dubé C, Tremblay N, Khanizadeh S (2011) A non-destructive method to predict polyphenol content in strawberry. J Food Agric Environ 9:59–62

Fernie AR (2003) Review: metabolome characterisation in plant system analysis. Funct Plant Biol 30:111–120

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18:1157–1161

Fu L, Xu BT, Xu XR, Gan RY, Zhang Y, Xia EQ, Li HB (2011) Antioxidant capacities and total phenolic contents of 62 fruits. Food Chem 129:345–350

Galambošová J, Macák M, Živčák M, Rataj V, Slamka P, Olšovská K (2014) Comparison of spectral reflectance and multispectrally induced fluorescence to determine winter wheat nitrogen deficit. Adv Mater Res 1059:127–133

Gan RY, Kuang L, Xu XR, Zhang YA, Xia EQ, Song FL, Li HB (2010) Screening of natural antioxidants from traditional Chinese medicinal plants associated with treatment of rheumatic disease. Molecules 15:5988–5997

Garcia LL, Cosme LL, Peralta HR, Garcia BM (1973) Phytochemical investigation of *Coleus blumei* Benth. Philipp J Sci 102:1–12

Ghasemi PA, Rahnama GH, Malekpoor F, Roohi BH (2011) Variation in antibacterial activity and phenolic content of *Hypericum scabrum* L. populations. J Med Plant Res 5:4119–4125

Ghozlen NB, Cerovic ZG, Germain C, Toutain S, Latouche G (2010) Nondestructive optical monitoring of grape maturation by proximal sensing. Sensors 10:10040–10068

Gitelson AA, Merzlyak MN, Chivkunova OB (2001) Optical properties and nondestructive estimation of anthocyanin content in plant leaves. Photochem Photobiol 74:38–45

González-Gallego J, Sánchez-Campos S, Tuñón MJ (2007) Anti-inflammatory properties of dietary flavonoids. Nutr Hosp 22:287–293

Gorham J (1995) The biochemistry of the stilbenoids. Chapman & Hall, London

Goulas Y, Cerovic ZG, Cartelat A, Moya I (2004) Dualex: a new instrument for field measurements of epidermal ultraviolet absorbance by chlorophyll fluorescence. Applied Optics 43:4488–4496

Grandmaison J, Ibrahim RK (1996) Evidence for nuclear binding of flavonol sulphate esters in *Flaveria chloraefolia*. J Plant Physiol 147:653–660

Hagen SF, Borge GIA, Solhaug KA, Bengtsson GB (2009) Effect of cold storage and harvest date on bioactive compounds in curly kale (*Brassica oleracea* L. var. acephala). Postharvest Biology and Technology 51(1):36–42

Hagen SF, Solhaug KA, Bengtsson GB, Borge GIA, Bilger W (2006) Chlorophyll fluorescence as a tool for nondestructive estimation of anthocyanins and total flavonoids in apples. Postharvest Biol Technol 41:156–163

Häkkinen S, Heinonen M, Kärenlampi S, Mykkänen H, Ruuskanen J, Törrönen R (1999) Screening of selected flavonoids and phenolic acids in 19 berries. Food Res Int 32:345–353

Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol 169:453–468

Hall RD (2011) Plant metabolomics in a nutshell: potential and future challenges. Ann Plant Rev Biol Plant Metab 43:1–24

Han Q, Shinohara K, Kakubari Y, Mukai Y (2003) Photoprotective role of rhodoxanthin during cold acclimation in Cryptomeria japonica. Plant, Cell and Environment 2:715–723

Haq M, Sani W, Hossain ABMS, Taha RM, Monneruzzaman KM (2011) Total phenolic contents, antioxidant and antimicrobial activities of *Bruguiera gymnorrhiza*. J Med Plant Res 5:4112–4118

Harborne JB (1976) Function of flavonoids in plants. In: Goodwin TW (ed) Chemistry and biochemistry of plant pigments. Academic, London

Harborne JB (1989) General procedures and measurement of total phenolics. In: Dey PM, Harborne JB (eds) Plant phenolics. Academic, London

Hazir MHM, Shariff ARM, Amiruddin MD (2012a) Determination of oil palm fresh fruit bunch ripeness—based on flavonoids and anthocyanin content. Ind Crop Prod 36:466–475
Hazir MHM, Shariff ARM, Amiruddin MD, Ramli AR, Iqbal Saripan M (2012b) Oil palm bunch ripeness classification using fluorescence technique. J Food Eng 113:534–540
Hong YP, Lin SQ, Jiang YM, Ashraf M (2008) Variation in contents of total phenolics and flavonoids and antioxidant activities in the leaves of 11 *Eriobotrya* species. Plant Food Hum Nutr 63:200–204
Ibrahim RK (1992) Immunolocalization of flavonoid conjugates and their enzymes. In: Stafford HA, Ibrahim RK (eds) Phenolic metabolism in plants. Plenum, New York, NY
Ishii S, Katsumura T, Shiozuka C, Ooyauchi K, Kawasaki K, Takigawa S, Fukushima T, Tokuji Y, Kinoshita M, Ohnishi M, Kawahara M, Ohba K (2008) Anti-inflammatory effect of buckwheat sprouts in lipopolysaccharide-activated human colon cancer cells and mice. Biosci Biotech Biochem 72:3148–3157
Jonadet M, Meunier MT, Villie F, Bastide JP, Lamaison JL (1986) Flavonoids extracted from Ribes nigrum L. and Alchemilla vulgaris L.: 1. in vitro inhibitory activities on elastase, trypsin and chymotrypsin. 2. Angioprotective activities compared in vivo. J Pharmacol 17:21–27
Jung HA, Park JC, Chung HY, Kim J, Choi JS (1999) Antioxidant flavonoids and chlorogenic acid from the leaves of *Eriobotrya japonica*. Arch Pharm Res 22:213–218
Jurd L (1957) The detection of aromatic acids in plant extracts by ultraviolet absorption spectra of their ions. Arch Biochem Biophys 66:284–288
Kagan J (1968) The flavonoid pigments of *Liatris spicata*. Phytochemistry 7:1205–1207
Kalaji HM, Schansker G, Ladle RJ, Goltsev V, Bosa K, Allakhverdiev SI, Brestic M, Bussotti F, Calatayud A, Dabrowski P, Elsheery NI, Lorenzo L, Guidi L, Hogewoning SW, Jajoo A, Misra AN, Nebauer SG, Pancaldi S, Penella C, Poli DB, Pollastrini M, Romanowska-Duda ZB, Rutkowska B, Serôdio J, Suresh K, Szulc W, Tambussi E, Yanniccari M, Zivcak M (2014) Frequently asked questions about in vivo chlorophyll fluorescence: practical issues. Photosynth Res 122:121–158
Kamal J (2011) Quantification of alkaloids, phenols and flavonoids in sunflower (*Helianthus annuus* L.). Afr J Biotechnol 10:3149–3315
Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30:42–46
Karen R, Polomski B (1999) Coleus. HGIC 1162: Extension. South Carolina: Clemson University; Clemson University, June 1999. Web. 03 May 2013
Kaushik R, Pradeep N, Vamshi V, Geetha M, Usha A (2010) Nutrient composition of cultivated stevia leaves and the influence of polyphenols and plant pigments on sensory and antioxidant properties of leaf extracts. J Food Sci Technol 47:27–33
Kawa JM, Taylor CG, Przybylski R (2003) Buckwheat concentrate reduces serum glucose in streptozotocin-diabetic rats. J Agric Food Chem 51:7287–7291
Kayashita J, Shimaoka I, Nakajoh M, Yamazaki M, Norihisa K (1997) Consumption of buckwheat protein lowers plasma cholesterol and raises fecal neutral sterols in cholesterol-fed rats because of its low digestibility. J Nutr 127:1395–1400
Keski-Saari S, Julkunen-Tiitto R (2003) Early developmental responses of mountain birch (*Betula rubescens* subsp. Czerepanovii) seedlings to different concentrations of phosphorus. Tree Physiol 23:1201–1208
Khattak MMAK, Taher M (2011) Bioactivity-guided isolation of antimicrobial agent from *Coleus amboinicus Lour* (*Torbangun*). Technical report. Submitted IIUM RMC. IIUM, Kuala Lumpur, http://irep.iium.edu.my/3985/
Kim HK, Choi YH, Verpoorte R (2011) NMR-based plant metabolomics: where do we stand, where do we go? Trends Biotechnol 29:267–275
Kolb CA, Käser MA, Kopecký J, Zotz G, Riederer M, Pfündel EE (2001) Effects of natural intensities of visible and ultraviolet radiation on epidermal ultraviolet screening and photosynthesis in grape leaves. Plant Physiol 127:863–875
Kolb CA, Kopecký J, Riederer M, Pfündel EE (2003) UV screening by phenolics in berries of grapevine (*Vitis vinifera*). Funct Plant Biol 30:1177–1186

Kolb CA, Schreiber U, Gademann R, Pfundel EE (2005) UV-A screening in plants determined using a new portable fluorimeter. Photosynthetica 43:371–377

Koostra A (1994) Protection from UV-B-induced DNA damage by flavonoids. Plant Mol Biol 26:771–774

Kubínová R, Jankovská D, Bauerová V (2012) Antioxidant and α-glucosidase inhibition activities and polyphenol content of five species of *Agrimonia* genus. Acta Fytotech zootech 15:38–41

Kuckenberg J, Tartachnyk I, Noga G (2008) Evaluation of fluorescence and remission techniques for monitoring changes in peel chlorophyll and internal fruit characteristics in sunlit and shaded sides of apple fruit during shelf-life. Postharvest Biol Technol 48:231–241

Lai A, Santangelo E, Soressi GP, Fantoni R (2007) Analysis of the main secondary metabolites produced in tomato (*Lycopersicon esculentum*, Mill.) epicarp tissue during fruit ripening using fluorescence techniques. Postharvest Biol Technol 43:335–342

Laitinen ML, Julkunen-Tiitto R, Rousi M (2002) Foliar phenolic composition of European white birch during bud and leaf development. Physiol Plant 114:450–460

Latouche G, Bellow S, Poutaraud A, Meyer S, Cerovic ZG (2013) Influence of constitutive phenolic compounds on the response of grapevine (*Vitis vinifera* L.) leaves to infection by *Plasmopara viticola*. Planta 237:351–361

Le Moigne M, Florin L, Rigaud S, Cerovic ZG (2010) Anthocyanin assessment at grape reception in a winery using a fluorescence optical remote sensor. In: Macrowine 2010: third international symposium on macromolecules and secondary metabolites of grapevine and wine, p 85

Lee TT, Huang CC, Shieh XH, Chen CL, Chen LJ, Yu B (2010) Flavonoid, phenol and polysaccharide contents of *Echinacea purpurea* L. and its immunostimulant capacity *in vitro*. IJESD 1:5–9

Lei Z, Huhman DV, Sumner LW (2011) Mass spectrometry strategies in metabolomics. J Biol Chem 286:25435–25442

Lenk S, Buschmann C (2006) Distribution of UV-shielding of the epidermis of sun and shade leaves of the beech (*Fagus sylvatica* L.) as monitored by multi-colour fluorescence imaging. J Plant Physiol 163:1273–1283

Lenk S, Buschmann C, Pfündel EE (2007) *In vivo* assessing flavonols in white grape berries (*Vitis vinifera* L. cv. Pinot Blanc) of different degrees of ripeness using chlorophyll fluorescence imaging. Funct Plant Biol 34:1092–1104

Leporatti ML, Ivancheva S (2003) Preliminary comparative analysis of medicinal plants used in the traditional medicine of Bulgaria and Italy. J Ethnopharmacol 87:123–142

Lins EC, Belasque Junior J, Marcassa LG (2009) Detection of citrus canker in citrus plants using laser induced fluorescence spectroscopy. Precision Agric 10:319–330

Louati S, Simmonds MSJ, Grayer RJ, Kite GC, Damak M (2003) Flavonoids from *Eriobotrya japonica* (*Rosaceae*) growing in Tunisia. Biochem Syst Ecol 31:99–101

Louis J, Cerovic ZG, Moya I (2006) Quantitative study of fluorescence excitation and emission spectra of bean leaves. J Photoch Photobio B 85:65–71

Louis J, Meyer S, Maunoury-Danger F, Fresneau C, Meudec E, Cerovic ZG (2009) Seasonal changes in optically assessed epidermal phenolic compounds and chlorophyll contents in leaves of sessile oak (*Quercus petraea*): towards signatures of phenological stage. Funct Plant Biol 36:732–741

Luqman S, Rizvi SI (2006) Protection of lipid peroxidation and carbonyl formation in proteins by capsaicin in human erythrocytes subjected to oxidative stress. Phytother Res 20:303–306

Ma MS, Bae IY, Lee HG, Yang CB (2006) Purification and identification of angiotensin I-converting enzyme inhibitory peptide from buckwheat (*Fagopyrum esculentum* Moench.). Food Chem 96:36–42

Ma F, Cheng L (2004) Exposure of the shaded side of apple fruit to full sun leads to up-regulation of both the xanthophyll cycle and the ascorbate–glutathione cycle. Plant Science 166: 1479–1486

Mabry TJ, Markham KR, Thomas MB (1970) The systematic identification of flavonoids. Springer, Heidelberg

Maha MS, Zeinab AKTI (2012) Cytotoxic compounds from the leaves of *Gaillardia aristata* Pursh. growing in Egypt. Nat Prod Res: formerly. Nat Prod Lett 26(22):2057–2062. doi:10.1080/14786419.2011.606219

Maoz I, Bahar A, Kaplunov T, Zutchi Y, Daus A, Lurie S, Lichter A (2014) The effect of the cytokinin forchlorfenuron on the tannin content of Thompson Seedless table grapes. Am J Enol Viticult 65(2):230–237

Matese A, Capraro F, Primicerio J, Gualato G, Di Gennaro SF, Agati G (2013) Mapping of vine vigor by UAV and anthocyanin content by a non-destructive fluorescence technique. In: Precision agriculture'13. Wageningen Academic Publishers, Wageningen, pp 201–208

Mendes Novo J, Iriel A, Lagorio MG (2012) Modelling chlorophyll fluorescence of kiwi fruit (*Actinidia deliciosa*). Photochem Photobiol Sci 11:724–730

Mercure S-A, Daoust B, Samson G (2004) Causal relationship between growth inhibition, accumulation of phenolic metabolites, and changes of UV-induced fluorescences in nitrogen-deficient barley plants. Can J Bot 82:815–821

Merzlyak MN, Melø TB, Naqvi KR (2008) Effect of anthocyanins, carotenoids, and flavonols on chlorophyll fluorescence excitation spectra in apple fruit: signature analysis, assessment, modelling, and relevance to photoprotection. J Exp Bot 59:349–359

Merzlyak M, Solovchenko A, Pogosyan S (2005) Optical properties of rhodoxanthin accumulated in Aloe arborescens Mill. leaves under high-light stress with special reference to its photoprotective function. Photochemical and Photobiological Sciences 4:333–400

Merzlyak MN, Solovchenko AE (2002) Photostability of pigments in ripening apple fruit: a possible photoprotective role of carotenoids during plant senescence. Plant Science 163: 881–888

Meyer S, Louis J, Moise N, Piolot T, Baudin X, Cerovic ZG (2009) Developmental changes in spatial distribution of *in vivo* fluorescence and epidermal UV absorbance over *Quercus petraea* leaves. Ann Bot 104:621–633

Morales LO, Tegelberg R, Brosché M, Lindfors A, Siipola S, Aphalo PJ (2011) Temporal variation in epidermal flavonoids due to altered solar UV radiation is moderated by the leaf position in *Betula pendula*. Physiol Plant 143:261–270

Moskowitz AH, Hradzina G (1981) Vacuolar contents of fruit subepidermal cells from *Vitis* sp. Plant Physiol 68:686–692

Moyer RA, Hummer KE, Finn CE, Frei B, Wrolstad RE (2002) Anthocyanins, phenolics, and antioxidant capacity in diverse small fruits: *Vaccinium*, *Rubus,* and *Ribes*. J Agric Food Chem 50:519–525

Muller V, Lankes C, Schmitz-Eiberger M, Noga G, Hunsche M (2013) Estimation of flavonoid and centelloside accumulation in leaves of *Centella asiatica* L. Urban by multiparametric fluorescence measurements. Environ Exp Bot 93:27–34

Nicholson JK, Lindon JC (2008) Systems biology: metabonomics. Nature 455:1054–1056

Nowak R, Gawlik-Dziki U (2007) Polyphenols of *Rosa* L. leaves extracts and their radical scavenging activity. Z Naturforsch C 62:32–38

Ounis A, Cerovic ZG, Briantais JM, Moya I (2001a) Dual-excitation FLIDAR for the estimation of epidermal UV absorption in leaves and canopies. Remote Sens Environ 76:33–48

Ounis A, Cerovic ZG, Briantais JM, Moya I (2001b) DE-FLIDAR: a new remote sensing instrument for estimation of epidermal UV absorption in leaves and canopies. In: Proceedings of European association of remote sensing laboratories (EARSeL)-SIG-workshop LIDAR (EARSeL, 2000), vol 1, pp 196–204

Ozimina II (1979) Flavonoids of *Spartium junceum*. 1. Flavones and flavonols. Chem Nat Comp 16:763–764

Palme E, Bilia AR, Morelli I (1996) Flavonols and isoflavones from *Cotoneaster simonsii*. Phytochemistry 42:903–905

Pandey KB, Rizvi SI (2009) Plant polyphenols as dietary antioxidants in human health and disease. Oxid Med Cell Longev 2:270–278

Pandey KB, Mishra N, Rizvi SI (2009) Protective role of myricetin on markers of oxidative stress in human erythrocytes subjected to oxidative stress. Nat Prod Commun 4:221–226

Patel DK (2012) Study on medicinal plants with special reference to family Asteraceae, Fabaceae and Solanaceae in G.G.V.-Campus, Bilaspur (C. G.) in central India. Curr Bot 3:34–38

Pfündel EE, Agati G, Cerovic ZG (2006) Optical properties of plant surfaces. In: Riederer M, Muller C (eds) Biology of the plant cuticle. Annual Plant Reviews, Vol. 23. Blackwell Publishing, Oxford, pp 216–249

Pfündel EE (2003) Action of UV and visible radiation on chlorophyll fluorescence from dark-adapted grape leaves (*Vitis vinifera* L.). Photosynth Res 75:29–39

Pfündel EE, Ghozlen NB, Meyer S, Cerovic ZG (2007) Investigating UV screening in leaves by two different types of portable UV fluorimeters reveals *in vivo* screening by anthocyanins and carotenoids. Photosynth Res 93:205–221

Pinelli P, Romani A, Fierini E, Remorini D, Agati G (2013) Characterisation of the polyphenol content in the kiwifruit (*Actinidia deliciosa*) exocarp for the calibration of a fruit-sorting optical sensor. Phytochem Anal 24:460–466

Pokorný J (2000) Natural antioxidants. In: Zeuthen P, Bøgh-Sørensen L (eds) Food preservation techniques. Woodhead Publishing, Cambridge

Pokorný J (2007) Are natural antioxidants better- and safer-than synthetic antioxidants? Eur J Lipid Sci Technol 109:629–642

Potters G, Pasternak TP, Guisez Y, Palme KJ, Jansen MAK (2007) Stress-induced morphogenic responses: growing out of trouble? Trends Plant Sci 12:98–105

Raal A, Kirsipuu K (2011) Total flavonoid content in varieties of *Calendula officinalis* L. originating from different countries and cultivated in Estonia. Nat Prod Res 25:658–662

Reinold S, Hahlbrock K (1997) *In situ* localization of phenylpropanoid biosynthetic mRNAs and proteins in parsley (*Petroselinum crispum*). Bot Acta 110:431–443

Repkova J, Brestic M, Zivcak M (2008) Bioindication of barley leaves vulnerability in conditions of water deficit. Cereal Res Commun 36:1747–1750

Robards K, Antolovich M (1997) Analytical biochemistry of fruit flavonoids, a review. Analyst 122:11–34

Roessner U, Willmitzer L, Fernie A (2002) Metabolic profiling and biochemical phenotyping of plant systems. Plant Cell Rep 21:189–196

Rop O, Mlcek J, Jurikova T, Neugebauerova J, Vabkova J (2012) Edible flowers – a new promising source of mineral elements in human nutrition. Molecules 17:6672–6683

Saure MC (1990) External control of anthocyanin formation in apple. Sci Hortic 42:181–218

Scalbert A, Manach C, Morand C, Remesy C (2005) Dietary polyphenols and the prevention of diseases. Crit Rev Food Sci 45:287–306

Schmutz A, Buchala A, Jenny T, Ryser U (1994) The phenols in the wax and in the suberin of green cotton fibres and their function. Acta Hortic 381:269–275

Schnabl H, Weissenböck G, Scharf H (1986) *In vivo* microspectro photometric characterisation of flavonol glycosides in *Vicia faba* guard and epidermal cells. J Exp Bot 37:61–72

Schnabl H, Weissenböck G, Sachs G, Scharf H (1989) Cellular distribution of UV-absorbing compounds in guard and subsidiary cells of *Zea mays* L. J Plant Physiol 135:249–252

Schnitzler JP, Jungblut TP, Heller W, Hutzler P, Heinzmann U, Schmelzer E, Ernst D, Langebartels C, Sandermann H (1996) Tissue localization of UV-B screening pigments and chalcone synthase mRNA in Scots pine (*Pinus sylvestris* L.) needles. New Phythol 132:247–258

Scogings P, Siko S, Taylor R (2014) Calibration of a hand-held instrument for measuring condensed tannin concentration based on UV-and red-excited fluorescence. Afr J Range For Sci 31:55–58

Senejoux F, Demougeot C, Karimov U, Muyard F, Kerramb P, Aisa HA, Girard-Thernier C (2013) Chemical constituents from *Echinops integrifolius*. Biochem Syst Ecol 47:42–44

Shaw AK, Ghosh S, Kalaji HM, Bosa K, Brestic M, Zivcak M, Hossain Z (2014) Nano-CuO stress induced modulation of antioxidative defense and photosynthetic performance of Syrian barley (*Hordeum vulgare* L.). Environ Exp Bot 102:37–47

Shaza AM, Nadia MS, Omyma El-G, Zeinab YA, Iman MA (2012) Phytoconstituents Investigation, Anti-diabetic and Anti-dyslipidemic Activities of *Cotoneaster horizontalis* Decne Cultivated in Egypt. Life Science Journal 9(2s):394–403

Sheahan JJ (1996) Sinapate esters provide greater UV-B attenuation than flavonoids in *Arabidopsis thaliana* (*Brassicaceae*). Am J Bot 83:679–686
Shimazaki K-I, Igarashi T, Kondo N (1988) Protection by the epidermis of photosynthesis against UV-C radiation estimated by chlorophyll a fluorescence. Physiol Plant 74:34–38
Singh RP, Pandey VB (1994) Further flavonoids of *Echinops niveus*. Fitoterapia 65:374
Singh S, Upadhyay RK, Pandey MB, Singh JP, Pandey VB (2006) Flavonoids of *Echinops echinatus*. J Asian Nat Prod Res 8:197–200
Smith J, Markham KR (1998) Tautomerism of flavonol glucosides: relevance to plant UV protection and ower colour. Journal of Photochemistry and Photobiology (A) 11:99–105
Steyn WJ, Wand SJE, Holcroft DM, Jacobs G (2002) Anthocyanins in vegetative tissues: a proposed unified function in photoprotection. New Phytologist 155:349–361
Strack D, Heilemann J, Klinkott JS (1988) Cell wall-bound phenolics from Norway spruce (*Picea abies*) needles. Zeitschrift für Naturforschung C, A Journal of Biosciences 43(1–2):37–41
Strid A, Chow WS, Anderson JM (1994) UV-B damage and protection at the molecular level in plants. Photosynth Res 39:475–489
Sytar O, Bruckova K, Hunkova E, Zivcak M, Konate K, Brestic M (2015) The application of multiplex fluorimetric sensor for the analysis of flavonoids content in the medicinal herbs family *Asteraceae*, *Lamiaceae*, *Rosaceae*. Biol Res 48:5
Tavarini S, Degl'Innocenti E, Remorini D, Massai R, Guidi L (2008) Antioxidant capacity, ascorbic acid, total phenols and carotenoids changes during harvest and after storage of Hayward kiwifruit. Food Chem 107:282–288
Terfa MT, Solhaug KA, Gislerød HR, Olsen JE, Torre S (2013) A high proportion of blue light increases the photosynthesis capacity and leaf formation rate of *Rosa* × *hybrida* but does not affect time to flower opening. Physiol Plant 148:146–159
Trendafilovaa A, Todorovaa M, Nikolovab M, Gavrilovab A, Vitkovab A (2011) Flavonoid constituents and free radical scavenging activity of *Alchemilla mollis*. Nat Prod Commun 6:1851–1854
Treutter D (1989) Chemical reaction detection of catechins and proanthocyanins with 4-dimethylamino-cinnamaldehyde. J Chromatogr 467:185–193
Tulipani S, Mezzetti B, Capocasa F, Bompadre S, Beekwilder J, Ric de Vos CH, Capanoglu E, Bovy A, Battino M (2008) Antioxidants, phenolic compounds, and nutritional quality of different strawberry genotypes. J Agric Food Chem 56:696–704
Tulyathan V, Boondee K, Mahawanich T (2005) Characteristics of starch from water chestnut (*Trapa bispinosa* Roxb.). J Food Biochem 29:337–348
Veit M, Beckert C, Höhne C, Bauer K, Geiger H (1995) Interspecific and intraspecific variation of phenolics in the genus Equisetum subgenus Equisetum. Phytochemistry 38:881–891
Verhoeven HA, de Vos CR, Bino RJ, Hall RD (2006) Plant metabolomics strategies based upon quadrupole time of flight mass spectrometry (QTOF-MS) in plant metabolomics. Springer, Berlin
Verpoorte R, Choi YH, Kim HK (2005) Ethnopharmacology and systems biology: a perfect holistic match. J Ethnopharmacol 100:53–56
Vitrac X, Moni JP, Vercauteren J, Deffieux G, Mérillon JM (2002) Direct liquid chromatography analysis of resveratrol derivatives and flavanonols in wines with absorbance and fluorescence detection. Anal Chim Acta 458:103–110
Vogt T, Pollak P, Tarlyn N, Taylor LP (1994) Pollination- or wound-induced kaempferol accumulation in petunia stigmas enhances seed production. Plant Cell 6:11–23
Wagner H, Gilbert M, Wilhelm C (2003) Longitudinal leaf gradients of UV-absorbing screening pigments in barley (*Hordeum vulgare*). Physiol Plant 117:383–391
Wang B, Ji C (2008) Tannin concentration enhances seed caching by scatter-hoarding rodents: an experiment using artificial seeds. Acta Oecol 34:379–385
Wang SY, Lin H-S (2000) Antioxidant activity in fruits and leaves of blackberry, raspberry, and strawberry varies with cultivar and developmental stage. J Agric Food Chem 48:140–146
Wojdyło A, Oszmiański J, Czemerys R (2007) Antioxidant activity and phenolic compounds in 32 selected herbs. Food Chem 105:940–949

Wollenweber E, Dietz VH (1981) Occurrence and distribution of free flavonoid aglycones in plants. Phytochemistry 20:869–932

Wu X, Gu L, Prior RL, McKay S (2004) Characterization of anthocyanins and proanthocyanidins in some cultivars of *Ribes*, *Aronia*, and *Sambucus* and their antioxidant capacity. J Agric Food Chem 52:7846–7856

Wulf JS, Rühmann S, Rego I, Puhl I, Treutter D, Zude M (2008) Nondestructive application of laser-induced fluorescence spectroscopy for quantitative analyses of phenolic compounds in strawberry fruits (*Fragaria x ananassa*). J Agric Food Chem 56:2875–2882

Xu X, Qi X, Wang W, Chen G (2005) Separation and determination of flavonoids in *Agrimonia pilosa* Ledeb. by capillary electrophoresis with electrochemical detection. J Sep Sci 28:647–652

Yanishlieva N (2001) Inhibiting oxidation. In: Pokorny J, Yanishlieva N, Gordon M (eds) Antioxidants in food. Woodhead Publishing Ltd, Cambridge

Yanishlieva N, Marinova E, Pokorný J (2006) Natural antioxidants from herbs and spices. Eur J Lipid Sci Tech 108:776–793

Yoo KM, Lee CH, Lee H, Moon B, Lee CY (2008) Relative antioxidant and cytoprotective activities of common herbs. Food Chem 106:929–936

Zainuddin A, Pokorny J, Venskutonis R (2002) Antioxidant activity of sweetgrass (*Hierochloë odorata* Wahlnb.) extract in lard and rapeseed oil emulsions. Nahrung 46:15–17

Zhou C, Sun C, Chen K, Li X (2011) Flavonoids, phenolics, and antioxidant capacity in the flower of *Eriobotrya japonica* Lindl. Int J Mol Sci 12(5):2935–2945

Zivcak M, Brestic M, Olsovska K (2008a) Application of photosynthetic parameters in screening of wheat (*Triticum aestivum* L.) genotypes for improved drought and high temperature tolerance. In: Allen JF, Gantt E, Goldbeck JH, Osmond B (eds) Photosynthesis. Energy from the sun: 14th international congress on photosynthesis. Springer, Dordrecht

Zivcak M, Brestic M, Olsovska K (2008b) Physiological parameters useful in screening for improved tolerance to drought in winter wheat (Triticum aestivum L.). Cereal Res Commun 36:1943–1946

Zivcak M, Brestic M, Olsovska K, Slamka P (2008c) Performance index as a sensitive indicator of water stress in Triticum aestivum. Plant Soil Environ 54:133–139

Zivcak M, Brestic M, Balatová Z, Drevenaková P, Olsovska K, Kalaji HM, Allakhverdiev SI (2013) Photosynthetic electron transport and specific photoprotective responses in wheat leaves under drought stress. Photosynth Res 117:529–546

Zivcak M, Olšovská K, Slamka P, Galambošová J, Raraj V, Shao HB, Kalaji MH, Brestič M (2014a) Measurements of chlorophyll fluorescence in different leaf positions may detect nitrogen deficiency in wheat. Zemdirbyste Agric 101:437–444

Zivcak M, Brestic M, Kalaji HM, Govindjee (2014b) Photosynthetic responses of sun-and shade-grown barley leaves to high light: is the lower PSII connectivity in shade leaves associated with protection against excess of light? Photosynth Res 119:339–354

Zivcak M, Kalaji HM, Shao HB, Olšovská K, Brestič M (2014c) Photosynthetic proton and electron transport in wheat leaves under prolonged moderate drought stress. J Photochem Photobiol B 137:107–115

Zivcak M, Olšovská K, Slamka P, Galambošová J, Rataj V, Shao HB, Brestič M (2014d) Application of chlorophyll fluorescence performance indices to assess the wheat photosynthetic functions influenced by nitrogen deficiency. Plant Soil Environ 60:210–215

Plant Glycomics

M. Asif Shahzad, Aimal Khan, Maria Khalid, and Alvina Gul

Contents

M.A. Shahzad • A. Khan • M. Khalid • A. Gul (✉)
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_19

Abstract Glycomics is the comprehensive study of glycomes (the entire complement of sugars, whether free or present in more complex molecules of an organism), including genetic, physiologic, pathologic, and other aspects in living organisms. Carbohydrates being most abundant macromolecules are found in all organisms' organs, tissues, and cells. They are present in almost every organelle of the cell in varying amounts depending on the type of organelle. Carbohydrates are not directly synthesized by genes; rather they are formed by gene products. Sometime carbohydrates are present in free form and also exist in the form of conjugates. Plants being the largest producer of carbohydrates on earth are of particular importance. Plants are rich in complex carbohydrates molecules. Complex biopolymers like cellulose, lignin, and hemicelluloses are being studied along with their structure and type of linkages present between them. The presence of these carbohydrates is somewhat linked to the survival of these plants under extreme conditions and stresses. Understanding these carbohydrates has allowed us to find answers on how plants survived severe climate changes in the past. These complex molecules form linkages with non-carbohydrate molecules and understanding the structure of these conjugates is a challenging task to the scientific community. Glycomics approach regarding the structural and functional analysis of these carbohydrates has been revolutionized by the modification in techniques like mass spectrophotometry, high pressure liquid chromatography, and capillary electrophoresis. Still some improvements are needed in these techniques to make glycomic approach less time-consuming and more specific and sensitive.

Keywords Cell wall glycome • Glycoconjugates • Pectin • Lectins • Glycosides • Microarray • Cell membrane glycome • Chloroplast glycome • Mitochondrial glycome

1 Introduction

Glycomics is the structural and functional study of carbohydrates in a biological system. "Glyco" means sugar/carbohydrates/glycans or saccharides and "omics" means study of the totality of something. So glycomics is the study of all the processes in a living organism in which glycans are involved (glycobiology). Carbohydrates, being the most abundant macromolecule on earth, is also present in the form of conjugate molecules, such as glycoproteins (a repeating structure is usually absent except in poly-*N*-acetyllactosamine and polysialic acid), proteoglycans (in proteoglycans, carbohydrates are composed of repeats of disaccharides units and are comparatively larger than 3000 Da), and glycolipids (conjugated molecules of carbohydrates with lipids; Varki et al. 1999). Human galactin has been reported to be involved in a number of immunogenic responses along with interaction with other macromolecules (Krzeminski 2011).

Carbohydrates are polyhydroxy aldehydes or ketones or those compounds which on hydrolysis yield these (aldose or ketose) compounds. Carbon, hydrogen, and oxygen are the basic elements present in carbohydrates with the empirical formula $(CH_2O)_n$, while sometime they also contain sulfur, phosphorous, or nitrogen. Carbohydrates are very complex relative to DNA or protein because of its branching nature. Furthermore, the presence of more than 30 basic building block units in carbohydrates adds more to its complexity in relation to DNA and protein, which has 4 and 20 basic units, respectively. Transfer of information from microtubules to extracellular matrix also involves glycoproteins (Baskin and Gu 2012).

Carbohydrates are present in plants as either free reducing sugars or in a glycoconjugated form. In the conjugated form they exist as proteoglycans, glycolipids, and glycoproteins, and they play an important role in cell signaling cascade, immune response, hormone action, and viral infection (Varki 1993; Brockhausen et al. 1998; Dennis et al. 1999; Haltiwanger and Lowe 2004; Taniguchi et al. 2006). Being important macromolecules, they are also involved in many cellular processes which include pathogen interaction, motility, and adhesion (Ohtsubo and Marth 2006; Marth and Grewal 2008; Sharon 2006; Dube and Bertozzi 2005). Pollen tube in plants is also rich in various polysaccharides that determine its structure and development (Chebli et al. 2012). O-acetylation of various carbohydrates leads to modification into various polysaccharides essential for proper development (Gille and Pauly 2012). The content of polysaccharides varies with different stages of a plant. Biosynthesis of different polysaccharides is dependent upon the plant stage of growth (Mollet et al. 2013).

In all living organisms carbohydrates are either present in conjugated or glycosylated form. Glycosylation is more rampant than phosphorylation, acetylation, or methylation, and it occurs on both lipids and proteins (Apweiler et al. 1999). 50 % of the total proteins are estimated to be glycosylated (Haltiwanger and Lowe 2004). Glycoproteomics is another emerging aspect of analyzing many macromolecules using glycobiomarkers, and their respective derivatives (Gabius 2011). Galacturonans act as a regulator of growth and development in a number of plants (Ferrari et al. 2013). Polygalacturanases play an important role in regulation during the process of fruit ripening and development as well (Roongsattham et al. 2012). They are also involved in the signaling during the degradation of cell wall (Vallarino and Osorio 2012; Yang et al. 2013).

Almost 500 known glycolipids are involved in cell–cell interaction and membrane microdomain association in which glycoglycerolipids (sugars bonded with diacylglycerol) are mainly found in plants (Holzl and Dormann 2007). Glycosylation results in interlipid hydrogen bonding which as a result brings about structural integrity to membranes of organisms and other many more functions such as intercellular communication and as receptors and signal transducer components (Ochoa-Villarreal et al. 2012). O-glycosylated proteins also play a significant role in determining the hair growth of plants (Velasquez et al. 2011).

Membranes of membrane-bound organelles (Golgi bodies, endosomes, lysosomes, nuclear membrane, endoplasmic reticulum, and mitochondria) have two-thirds of glycolipids (Gillard et al. 1993). Golgi bodies add ceramide to sugar unit

(saccharides) to synthesize glycolipids (Edidin 2003). Golgi movement is mostly controlled by the type of orientation of actin filaments (Miriam et al. 2011). Luminal side of organelles and leaflets of plasma membrane are distributed with glycolipids. Signaling of many components among different regions of cell is also mediated by sugar molecules (Xiang et al. 2011).

2 Plant Cell Wall

Cellulose is the abundant component of plant cell walls. A strong fibrous network of cell wall not only protects the cell (Underwood 2012) but also provides defense, strength, growth, water movement, etc. Being the most abundant biopolymer, cellulose is a raw material for manufacturing of paper, textiles, lumber, and other products. Cell walls isolated from stipules of cold-acclimated and non-acclimated plants showed that cold temperatures induce changes in polymers containing xylose, arabinose, galactose, and galacturonic acid residues (Baldwin et al. 2014). Distribution of these polysaccharides has different percentage in roots, stems, and leaves (Cao et al. 2014).

2.1 Cell Wall Glycome

Plant cell wall being the outermost layer helps in the protection and survival of cells. It performs a number of functions including provision of specific shape to different plant tissues and organs which are essential for proper functioning of tissues and aid in interaction with other macromolecules present inside and outside of cells (Terao et al. 2013). It also aid in plant–microbe interaction and defense related aspects of plants (Gough and Cullimore 2011).

Outside the cell wall there is a waxy cuticle layer that is an important barrier in protecting the cell from excessive water loss. It is reported that some of the plants have a hydrocarbon polymer in cuticle layer called cutan (Schmidt and Schönherr 1982). Cuticle is composed of cuticular waxes and mixtures of water-insoluble compounds along with hydrocarbons with varying chain length from C16 to C36 (Baker 1982). Cuticle resists the external invasion of different microbes and pathogenic agents, ultimately leading to protection of plant tissues. Diffusion of sugars across cell wall also determines the stability of cell structure (Van Der Wal and Leveau 2011).

Cell wall is divided into three further layers:

1. Middle lamella.
2. Primary cell wall.
3. Secondary cell wall.

2.1.1 Middle Lamella

It is the layer present between two adjacent cells that help in joining and exchange of materials along with providing stability necessary for plasmodesmata formation between cells. It is composed mainly of pectin that is a set of complex polysaccharides that is also present in primary cell wall. In some plants it is difficult to identify middle lamella due to the thickening of secondary cell wall (Raven 2005).

Pectin

Pectin is an important constituent of both middle lamella and primary cell wall in most of the plants. Pectin, also referred to as pectic polysaccharide, is mainly composed of about 200 monomers of D-galacturonic acid (Pornsak 2003). Pectin is essential for determining specific morphology of various parts of plants (Palin and Geitmann 2012). In the phenomenon of fruit ripening, pectin present in middle lamella is broken down by the action of an enzyme called pectinase. As a result the middle lamella is dissolved, leading to the softening of fruit. Under high auxin level, the amount of pectin in cell wall is reported to be increased considerably (Braybrook and Peaucelle 2013). When the fruit is under the process of ripening, the expression of pectin methylesterase is also increased (Cação et al. 2012). When deacetylation of pectin occurs, it results in abnormal cell elongation, germination, and reproduction (Gou et al. 2012). Demethylesterification of pectin in cell wall plays a significant role in seed germination by weakening the seed coat pectin and softening the coat (Müller et al. 2013). There are three different classes of pectin which are important structural components of both middle lamella and primary cell wall. Pectin lyase that degrades the pectin molecule is used commercially for starch decomposition and also for protoplast formation in cell culture techniques (Cao 2012). These three pectins include homogalacturonans, rhamnogalacturonans, and substituted galacturonans (Buchanan et al. 2000). They are discussed in detail in the next section.

2.1.2 Primary Cell Wall

It consists of growing cells or cells that have the ability to grow. This is synthesized after the formation of middle lamella, and it comprises a hard cross-linked network of cellulose fibrils that are fixed into the matrix mainly formed by pectins, hemicelluloses, and some glycoproteins. These cells are present on the outer side of plant stem. It is composed of cellulose, hemicelluloses, and pectin. Cellulose being the major component of cell wall plays an important role in defining the shape and integrity of cell wall. Under hydrothermal treatment, the content of cellulose and lignin increased by 43 % and 29 %, respectively, in wheat (Merali et al. 2013). Removal of enzymes that are responsible for the transfer of galacturonan results in abnormal development of pollen tube (Wang et al. 2013a, b).

Cellulose

Cellulose is an important component of primary cell wall that leads to the stability and control of growth of plant cells and tissues. In fact it the most abundant organic polymer present on earth. Cellulose is complex polymer consisting of about tens of thousands of residues of β(1 → 4) linked D-glucose (Crawford 1981). In primary cell wall this cellulose polymer makes up a complex network of residues leading to microfibrillar structure that help in providing specific shape to the cell (Brett and Waldron 1990). Complex chains of cellulose interlink with H-bonds, leading to the formation of crystalline microfibrils. Further X-ray studies also revealed that there are some other polysaccharides which are covalently bonded to the cellulose complex. These polysaccharides were predicted to be xylans and mannans due to the fact that the amorphous phase developed by these molecules had a lot of resemblance to the X-ray study conducted (Koller et al. 1991). It was revealed that bonding to these molecules makes cellulose more resistant and cannot be broken even if treated with a strong acid (Taiz 1984).

Cellulose has also been reported in other organisms like bacteria and algae but the difference is its existence as mixture or as pure. The plant cellulose exists as mixture with hemicelluloses, pectin and lignin but cellulose in bacteria exist in pure form with greater water content. Celluloses are categorized on the nature and arrangement of H-bonds. Natural cellulose is cellulose I, which is further categorized into Iα and Iβ. Cellulose Iα is present in bacteria, while cellulose Iβ is a component of plants (Serge Pérez, William Mackie 2001 CERMAV).

Hemicelluloses

Hemicellulose is the second most abundant polysaccharides present in nature. It is a complex heteropolymer of pentose and hexose saccharides. Mostly pentose sugar includes xylan and arabinose, while hexose sugars are mannose, glucose, and galactose. In plant stems, softwood hemicelluloses are made of glucomannans and hardwood contains mostly xylans (McMillan 1993).

Xylan is a heteropolysaccharide comprised of chain of 1,4-linked β-D-xylopyranose units. It has also been reported that in addition to xylopyranose, it also contain some other units that includes arabinose and glucuronic acid. In different plants the composition of xylan may differ (Aspinall 1980; Carbohydrates: structure and function. New York: Academic). Xylan content is different in different plants depending upon the species in which it is present (Shibuya and Iwasaki 1985).

Pectin

As discussed earlier, pectin is an important part of middle lamella, primary cell wall, and secondary cell wall in plants. Matrix present in cell wall is formed by pectin and hemicelluloses in which cellulose microfibrils are anchored. It is a

complex macromolecule made up of D-galacturonic acid residues (Pornsak 2003). Pectin is considered to be the most complex macromolecule present as it can consist of about 17 different residues of monosaccharides linked by 20 different kinds of linkages (Ridley et al. 2001a, b). It is also involved in the coloring and pigmentation of many grapes species (Ochoa-Villarreal et al. 2011). In most of dicotyledonous plants, pectin is about 35 % of whole plants (Fry 1988). Most of the grasses have about 2–10 % pectin, while in fruits of some plants cell wall has a higher content of pectin compared to other parts of plant (Fry 1988). It has also been reported that pectin is also involved in the activation of defense response against different pathogenic stimuli by releasing specific components known as phytoalexins that are known for their antimicrobial activity (Hahn et al. 1981; McCarthy et al. 2014). The structure of pectin is quite complex as it includes a number of other residues that are complex and give primary cell wall stability and ability of signal transduction (Yapo 2011). It includes other peptic polysaccharides like homogalacturonan, xylogalacturonan, rhamnogalacturonan I, rhamnogalacturonan II, arabinogalactan I, arabinogalactan II, and arabinan (Atmodjo et al. 2013).

Homogalacturonan is the most abundant pectic polysaccharide present which accounts for about 60 % of total pectin. It is a linear polymer of α-1,4-linked galacturonic acid residues. Some modifications like methyl esterification at C-6 or O-acetylated at O-2 and O-3 help in defining the physical nature of pectin (Ishii 1995). In addition to methyl esterification, distribution of ester does affect physical properties of peptic macromolecules. Galactouronan is also involved in the accumulation of nutrients in many plants (Camejo et al. 2011). Green tea being widely used is also rich in homogalacturonan (HG) pectins consisting of a backbone of 1,4-linked α-D-galacturonic acid (GalA) residues with 28.4 % and 26.1 % of carboxyl groups as methyl ester respectively (Wang et al. 2014). These HG components are essential macromolecules of green tea that play beneficial role in defense against pathogens (Wang et al. 2013a, b).

Rhamnogalacturonan I is another structural component of pectin which consists of a-(1,5)-L-arabinan, b-(1,4)-galactan, and type-I arabinogalactan. It accounts for about 20–35 % of pectin. It has been found in cell wall with varying percentage in different plants. In potato cell wall dry weight, it accounts for 35 % of polysaccharides (Obro et al. 2004). RGI is mostly substituted with neutral sugars at O-4 and the side chains that are bound to the RGI includes arabinogalactan and arabinan residues (Lerouge et al. 1993). Different organs of plants have varying concentrations of RGI in cell walls depending upon the type of function performed by each organ (Lee et al. 2013).

Rhamnogalacturonan II is one of the most highly conserved polysaccharide structures in plants. Its structure consists of four side chains with very different residues including apiose, aceric acid, 3-deoxy-lyxo-2-heptulosaric acid, and 3-deoxy-manno-2-octulosonic acid (Pabst et al. 2013). These very peculiar residues are attached to the nine molecules of homogalacturonan which are most of the time esterified (O'Neill et al. 2001). RGII has the ability to bind with the boron making a complex by forming borate-diol-ester which binds two homogalacturonan molecules by cross-linking them, hence resulting in an increased stability to cell wall

(Ishii et al. 1999). Borate is specifically important in providing strength to cell wall. In addition to that borate also specifies the meristematic and reproductive systems of plants. If either RGII loses its ability to bind borate or if borate fails to dimerize two HG molecules, it results in damage to the reproductive as well as meristematic parts of plants (Blevins 1998). It is also involved in the elongation of pollen tube in many plants (Dumont et al. 2014).

Xylogalactouronan is the polymer of a-(1,4)-linked D-galacturonic acid that has substitution of a residue β-D-xylose at the position of O-3. It has been reported that these macromolecules are present in cell wall of most plants although the degree of their presence varies among different plants (Nakamura et al. 2002). Different studies regarding the presence of XGA report that it is present in storage tissues and reproductive parts as well. A study conducted on duckweed plant showed that about 20.3 % pectin constitutes three main macromolecules, galacturonan, xylogalacturonan, and rhamnogalacturonan (Zhao et al. 2014). Now it is believed that it also an important macromolecule that help in defining the peptic structure in Arabidopsis thaliana (Gardner et al. 2002; Dilokpimol et al. 2014).

Arabinogalactan I in simple words is a polymer made up of arabinose and galactose residues. α-L-arabinofuranose residues is substituted at O-3 of galactosyl residues (Mohnen 1999). In plants, it is mostly present in gums and it has also been found to attach proteins forming arabinogalactan protein (AGP). β-galactose is also identified to be substituted at O-6 with galactan residues of arabinogalactan (Van de Vis 1994). Arabinogalactan proteins (AGPs) are a highly diverse class of cell surface proteoglycans that are commonly found in most plant species. AGPs play important roles in many cellular processes during plant development, such as reproduction, cell proliferation, pattern formation, and growth (Knoch et al. 2014). An enzyme hydroxyproline-o-glycosyltransferase play key role in the production of arabinogalactan proteins in many plants (Basu et al. 2013).

Arabinogalactan II is made up 1,3-linked β-D-galacturonic acid backbone. There are some linked chains of α-L-arabinofuranose (Ridley et al. 2001a, b). This type of arabinogalactan is mainly found to be binding to proteins forming arabinogalactan protein (AGP). Proteins are mostly enriched of specific amino acids like proline, alanine, serine, and threonine (Gaspar et al. 2001). It has also been reported that carrot cell wall AGP is linked to the pectin (Immerzeel et al. 2006). Galactosyl transferases enzymes have been identified as the key player in the synthesis of arabinogalactan type II molecule (Dilokpimol et al. 2014). This enzyme is thus critical for normal embryo development in plants as reported in *Arabidopsis thaliana* (Geshi et al. 2013).

2.1.3 Secondary Cell Wall

It is a thick layer present inside primary cell wall that is formed when the cell has fully grown. This layer helps in providing rigidity and support to the plant cell. It is mainly composed of cellulose, hemicelluloses, and an additional biopolymer termed as lignin. Cellulose and hemicelluloses already discussed earlier are important components of entire plant cell wall.

Lignin

Lignin being an important part of cell wall of plants and algae, it is the second most abundant organic polymer present on earth exceeded only by cellulose. It constitutes about 15–25 % of dry weight of plants (Boerjan et al. 2003). In cell wall, lignin cross-link with other macromolecules like cellulose and hemicelluloses and fills the spaces present in cell wall. It forms covalent linkages to hemicelluloses, thus providing greater support and enhanced mechanical strength to plants (Chabannes et al. 2001). Lignin is abundant in xylem tracheids and sclereid cells. It plays an important role transportation of water in stems of plants. Most of the polysaccharides in plants are hydrophilic in nature, while lignin is hydrophobic, so it does not allow the movement of water across the wall, hence acting as good water conducting channels. Lignin has also been involved in defense against the enzymes that are destructive by stopping their penetration through cell wall.

Lignins can be divided into three different categories: softwood lignin that is mostly present in gymnosperms, hardwood lignin that is part of angiosperms, and grass lignin present in graminaceous plants (Pearl 1967). Softwood lignin is mainly composed of guaiacyl lignin which is made up of coniferylmonolignol. Hardwood lignin is rich in guaiacyl-syringyl lignin which is composed mainly of coniferyl and sinapylmonolignols. Grass lignin consists of p-coumarylmonolignols (Freudenberg and Neish 1968). In addition to these monolignols, lignin also has a variety of functional groups that define its nature. These functional groups include methoxyl and some of aldehyde groups. Studies done on lignin clarify that there are covalent linkages present between lignin and hemicelluloses in wood (Erikkson and Lindgren 1977).

2.2 Cellulose Synthesis

CesA genes code for plant cellulose synthase complex, but this complex requires three different CesA genes (Taylor et al. 2003; Burn et al. 2002; Burton et al. 2004; Scheible and Pauly 2004). There are different sets of genes involved for primary (CesA-1, CesA-3, CesA-6) and secondary (CesA-4, CesA-7, CesA-8) cell wall biosynthesis (Arioli et al. 1998; Fagard et al. 2000; Taylor et al. 2003). CesA genes code for a protein called CesA protein. This CesA protein is present in hexameric form, embedded in cell membrane (Kimura et al. 1999).

Cellulose is composed of many small subunits called microfibrils. CesA proteins make many (1,4)-linked β-D-glucan chains and then crystallize it, spontaneously forming microfibrils. The length of a microfibril is very long relative to the size of cell, and it can wrap around the cell many times (Hayashi 1989). A family of enzyme is collectively responsible for the synthesis of cellulose network. This cellulose synthase family is critical for the proper organization of cell wall in cellulose. Model organisms like *Arabidopsis thaliana*, despite a very small genome, have a very complex cellulose synthase family (Carroll and Specht 2011).

Microtubules also significantly affect the enzymes that are responsible for the synthesis of cellulose network (Fujita et al. 2012).

It is possible that sterol glucoside, which is formed in cell membrane, starts glucan chain synthesis by using sterol glucoside and uridine 52-diphosphate-glucose to form short sterol-linked glucans in the presence of CesA and then CesA enzyme adds glucose residues to the already present glucans (Peng et al. 2002). In order to determine the functionality of the enzyme, a mutation was induced that showed significant reduction in cellulose crystallinity and also the activity of other cellulose synthases (Fujita et al. 2013). Cellulose synthase 5 is involved in the synthesis of cellulose network in seed coat and also affects the pectin structure as well (Harpaz-Saad et al. 2012). Mutation in two domains of cellulose synthase also results in significant reduction of microfibril crystallinity of cellulose (Harris et al. 2012).

Korrigan (KOR) is a membrane-bound endonuclease enzyme which is thought to be required for cellulose synthesis (Lane et al. 2001; Nicol et al. 1998). It is also confirmed that KOR enzyme is involved in proper crystallization of microfibrils because mutants are seen having defects in cell elongation and cytokinesis (Bashline et al. 2013). Even though mutants make (1,4)-linked β-D-glucan, they cannot crystallize microfibril properly (Bashline et al. 2011).

2.3 Synthesis of Other Polysaccharides

Along with cellulose other polysaccharides are also synthesized in the matrix of the cell; they are hemicellulose and pectin. They are more complex with varieties of sugar residues and glycosidic linkages. Twenty of the genes which code for glycosyltransferase are proved to be tangled in the formation of matrix polysaccharides (Scheible and Pauly 2004).

Cellulose synthase-like (CSL) is a family of genes which includes CesA gene (Richmond and Somerville 2000). CSL synthesizes beta-D-glycan (the back bone of hemicellulose) in Golgi bodies and other beta D-glycans in cell wall.

CSLA, which is a subgroup of CSL genes, encodes β-mannan synthase, and this enzyme helps in the formation of mannan backbone of some hemicelluloses.

Galactomannan, which has b-mannan backbone with side chains of galactose, is stored in guar-bean seeds and synthesis of polysaccharide is done by CSLA gene product (Dhugga et al. 2004). CSLA gene synthesizes proteins which take part in the synthesis of glucomannans and mannans of the growing cell wall (Liepman et al. 2005). Genetic factors responsible for the formation of complex cell wall network have been determined which shows that this complex network is formed by interrelated genomic signaling between many molecules (Brabham and Debolt 2012). Pectic-β(1,4)-galactan, extensin, and arabinogalactan play essential roles in the synthesis of cell wall for grape species differentially (Moore et al. 2014).

Synthases and glycosyltransferase enzymes form the backbone of polysaccharides and glycan branches, respectively. Genes for these enzymes have been recently identified.

2.4 Formation of Cell Wall Network

A strong and extensible synthesis of cell wall network is the result of both enzymatic cross-linking and interaction between cell wall polysaccharides. When the polysaccharides are synthesized in the cell, then they are secreted outside the cell, i.e., into the cell wall. There it gets associated with preexisting cell wall polymer and newly formed cellulose microfibrils (Talbott and Ray 1992). How does cell wall expand? The answer for this question is still not clear because of lack of information in polymer–polymer interaction in cell wall. Extensin enzyme plays a significant role in defining the architecture of plant cell wall (Lamport et al. 2011). For effective and accurate synthesis of cell wall, interaction of different enzymes must be appropriate (Bringmann et al. 2012).

Primary cell wall has xyloglucan (abundant hemicellulose), and it is involved in direct or indirect cross-linkage of microfibrils (Carpita and Gibeaut 1993; Fry 1989; Talbott and Ray 1992). An experimentally confirmed physical model of cell wall growth was proposed on the basis of thermodynamics of hydrogen-bonded networks (Veytsman and Cosgrove 1998) and the effect of abundance and size of xyloglucan on the cell wall enlargement (Takeda et al. 2002; Wolf et al. 2012).

Xyloglucan endotransglucosylase (XET) can cut the older bonds present between existing network of polysaccharides and ligate glycosidic bond to the older polysaccharides to combine existing and newly synthesized polysaccharides in xyloglucan backbone (Nishitani and Tominaga 1992; Steele et al. 2001; Purugganan et al. 1997; Yokoyama and Nishitani 2001; Rose et al. 2002).

XET adds new xyloglucans to the existing cell wall network and increases strength of cell wall by adding xyloglucans to the chain (Takeda et al. 2002).

In short, there is interaction between microfibrils with matrix polysaccharides and the interaction is non-covalent in cellulose. So there are many sites where cell wall may be loosened and these sites are the regions from where cell wall may be expanded.

2.4.1 Factors Involved in the Expansion of Cell Wall

Extension of cell wall is energized by mechanical energy of turgor pressure by cell wall polymers. Cell wall experiences a pressure of 100–1000 atm, but the question is how such a thin cell wall polymer under such high pressure could go a process of remodeling without risking a cell. This happens by a highly controlled mechanism in which cellulose microfibrils are separated (Marga et al. 2005) and then synthesized cell wall materials are integrated and this prevents thinning to the point of instability.

So there are four factors involved in the expansion of cell wall. They are: expansin, xyloglucan endotransglycolase, endo-(1,4)-β-D-glucanase, and hydroxyl radicals.

Expansin

Expansins are pH-dependent wall loosening proteins (McQueen-Mason et al. 1992). Change in pH is a stimulus for plants to grow (Rayle and Cleland 1970; Cosgrove 1989; Hager et al. 1971). pH 4.5–6 can activate expansin, so it means that low pH activates while high pH inactivates expansin. So there are many factors which alter cell growth, i.e., H+-ATPase pump in plasma membrane is one of the internal factors to change pH of cell wall while external factors include: salt stress, water deficits, early outgrowth of root hairs, root tropisms, light, hormone, and fusicoccin (fungal toxin).

Mechanism of Function of Expansin

Expansins are the loosening agent of primary cell wall and can alone restore the lost extension of cell wall. These two functions were proved by McQueen et al. by using enzyme protease/heat for the removal of acid-growth behavior of cell wall and then expansins restored the property of cell wall extensibility. If expansin is exogenously added to growing cells, it will result in cell wall expansion. Furthermore, if genes responsible for the production of expansin are overly expressed, then plant growth is stimulated and if these genes are silenced, then plant growth is ceased (Pien et al. 2001; Cho and Cosgrove 2000; Choi et al. 2003; Zenoni et al. 2004). There is one more way to confirm that expansin plays an important role in cell wall expansion; it is the expression of expansin-gene endogenously which will be compared with the onset either to increase or cease cell growth (Cho and Kende 1997; Reinhardt et al. 1998; Brummell et al. 1999; Vriezen et al. 2000; Lee and Kende 2001; Wu et al. 2001; Cho and Cosgrove 2002). These were some proofs which support the mechanism of action of expansins.

Expansin relaxes wall stress and it does not break non-covalent bonds among the polysaccharides of cell wall. Expansins initiate the process by weakening the hydrogen bonds between the networks of cellulose fibrils and then the hydrolysis of crystalline cellulose by cellulase. The moment expansin enters the cell wall, it stimulates extension and when expansin is removed from the cell wall, then cell wall gets back to its original inextensible state (McQueen-Mason and Cosgrove 1995), indicating no change in cell wall structure and degree of cross-linking. By discussing these results we can conclude that expansin is involved in the dissociation of a polysaccharide complex that joins microfibrils together (McQueen-Mason and Cosgrove 1994; McQueen-Mason and Cosgrove 1995).

Xyloglucan Endotransglucosylase/Hydrolase

This enzyme has many functions (Rose et al. 2002): wall loosening (Fry et al. 1992), wall strengthening (Antosiewicz et al. 1997), incorporating newly synthesized xyloglucans into the wall (Thompson and Fry 2001), rearranging loose xyloglucan strands to the surface of cellulose (Thompson and Fry 1997), fruit softening

(Redgwell and Fry 1993), and during the xylem formation it hydrolyzes xyloglucan (Farkas et al. 1992; Fanutti et al. 1993; Bourquin et al. 2002; Matsui et al. 2005).

Endo-(1,4)-β-D-Glucanase

These glycoside hydrolase family 9 enzymes are also known as cellulases. There are 25 family members of these enzymes in *A. thaliana* in which three members are involved in cellulose formation and remaining enzymes are of unknown function.

Hydroxyl Radical

Hydroxyl group, being highly active form of reactive oxygen species, has vital roles in cell death and cell signaling. In order to lose cell wall, this radical is attached to growing cells and can act as a stimulant for cell enlargement (Fry et al. 2002; Fry 1998; Liszkay et al. 2003; Schopfer et al. 2002). Hydroxyl radical can break polysaccharides by removing hydrogen atoms non-enzymatically (Fry 1998). If -OH group is artificially induced, then it results in the extension of cell walls (Liszkay et al. 2003).

Endogenous production of -OH is non-enzymatic and is due to copper which is bound to cell wall (Fry et al. 2002) or might be due to cell wall peroxidases (Liszkay et al. 2003) from hydrogen peroxide and superoxide anion. Synthesis of all these carbohydrates is controlled by a number of enzymes like ATPases (Bonza and De Michelis 2011).

3 Cell Membrane

Cell membrane separates the cell from its surroundings, so it has quite important functions regarding the communication with its interior and its external surrounding. It also provides constant environment for the reactions taking place inside the cell and helps in exchange of metabolites with its surroundings.

Twenty percent carbohydrates is the amount expected to be present in cell membrane (H.D. Grimes, unpublished results; cited from) with 30–40 % proteins (Kjellbom and Larsson 1984) and remaining 40–50 % lipids by weight making the composition of cell membrane.

3.1 Cell Membrane Glycome

Unlike other macromolecules, carbohydrates are very unique in their structure and are linked with each other in more than one form. They are connected with each other by either α or β-linkages with neighboring sugars' 3–4 different positions of

hydroxyl groups. Due to these contrasting features compared to other macromolecules, carbohydrates can have unlimited variations.

Due to this unique property it can act as a ligand for recognition by other molecules. Carbohydrates are not directly synthesized by genes; rather they are synthesized by gene products, i.e., glycosyltransferases. So for every carbohydrate there must be a gene and because there are not many genes for every new carbohydrate; this is the reason why there is not that much variation of carbohydrates in living organisms. β-1-3 glucan is reported to be involved in defense response along with change in its structure in tobacco plants (Fu et al. 2011).

3.1.1 Cell Surface

Carbohydrates present on the outer cell surface are the major components of plant cells and all living organisms, but on maturity specific carbohydrates are restricted on the cell surface of organisms.

Glycosaminoglycans (GAGs) are complex polysaccharides which is one of the components present on cell surface and extra cellular matrix interacting with many proteins. They are involved in many biological process such as cell development and growth.

3.1.2 Glycosaminoglycans (GAGs)

When hexuronic acid and hexosamine are linked together to form disaccharide and then unbranched repeating units of these disaccharides results into heterogeneous polysaccharides called *GAGs*. Carbohydrate portion this molecule is synthesized in Golgi bodies which is O-linked to a protein forming a conjugate molecule called proteoglycan (Silbert and Sugumaran 2002; Sugahara and Kitagawa 2002). Integral plasma membrane synthase synthesize hyaluronic acid which immediately secrets newly synthesized chain (Itano and Kimata 2002). Keratan sulfate, which is sulfated glycosaminoglycans, can be linked to core protein either by N-linkage or O-linkage.

4 Mitochondrial Glycome

Mitochondria also known as power house of cell is an important compartmented organelle that help in the production of energy that is utilized in most of the plant functions including transport, catabolism, and many other tasks like these. They are also involved in signaling, defense, and cell death (McBride et al. 2006). These compartments have different layers present on outer side which help in performing a number of functions. Mitochondrion is a site of a number of reactions like the electron transport chain. The mitochondrial membranes are mostly rich in proteins

and phospholipids, but there have been some traces of glycoconjugates in its structure. Glycosylation has been found to be present at outer membrane of mitochondria. Those molecules that are found to be glycosylated in acceptor proteins are *N*-acetylglucosamine and mannose (Hubbard and Ivatt 1981). These glycoproteins help in signaling of mitochondria to external environment.

5 Chloroplast Glycome

Chloroplast is an important organelle present only in plants and other photosynthetic organism. Chloroplast is site of photosynthesis in which plant utilize light to make up glucose. Chloroplast is double membrane compartment, outer and inner membrane. A number of important chemical reactions takes place inside the chloroplast which ultimately leads to the formation of glucose. Posttranslational modifications also play significant role in defining the functionality of carbonic anhydrases in model plants (Burén et al. 2011).

5.1 Envelope Membrane

Chloroplast is compartmented organelle that membranes providing stability and shape to the organelle. Envelope membranes can be further categorized into inner and outer membrane. Outer membrane is rich in lipids. Carbohydrates are present only in the form of conjugates with lipids. Inner membrane contains 48 % phospholipids, 46 % galactolipids, and the rest sulfolipids (Block et al. 1983). Outer membranes constitute higher galactolipids content as compared to the one present in inner membrane.

5.2 Thylakoid Membrane

It is rich in galactolipids comprising 78 % along with phospholipids 15.5 % and sulfolipids 6.5 %. This membrane is the site of light absorption and ATP synthesis (Block et al. 1983). Although there are some galactoses residues present in membrane but still it is considered as phospholipid layer.

Galactolipids present in membrane are mostly of two types. monogalactosyldiacylglycerol (MGD) and digalactosyldiacylglycerol (DGD). In case of MGD, the galactose residue is bound to the glycerol at position 3, while in case of DGD it is bound at the terminal side (Kelly and Dormann 2004).

5.3 Starch

Starch granules are very common in chloroplast of plants and other photosynthetic organisms making up 15 % of chloroplast volume (Austin et al. 2006). As the sugar is synthesized by chloroplast, it accumulates in the form of starch and consumed during respiration at night.

6 Analytical Tools in Plant Glycomics

Carbohydrates being an important structural component and major class of biological macromolecules of plant cell are important target of studies for the modern scientific community. They play an important role in cell adhesion and other defense mechanisms (Sharon 2006). In the last decade, a number of modern analytical tools been introduced to glycomic approaches to make the process more sensitive and efficient. Carbohydrates are present in both pure and conjugate form. Analysis of carbohydrates present in conjugate of protein or lipids has been a challenge for scientist. Even some of the conjugates are complex with multiple residues of different kinds of linkages (Liu et al. 1992). In the study of molecular organization of cell wall, NMR spectroscopy has played a significant role (Bootten et al. 2011).

Nature of each glycoconjugate is solely dependent upon the part to which that glycan is attached and what kind of linkage is present. In different parts of plant, glycan content is different. Carbohydrates are mostly hydrophilic in nature but they can be hydrophobic if bound to the water repellent molecule.

Analysis of different glycans requires different kinds of strategies to overcome the hurdles that obstruct the glycomic approach. Cellulose being one of the major components of whole plant and is present in cell wall. Cellulose was first discovered by French scientist in 1838 (Payen 1838). Cellulose was first analyzed with the help of different chemical treatment to partially digest the cellulose (Hon 1994) and then organic chemistry help in understanding the linkages and residues present in cellulose that were similar to the one in starch. X-ray diffraction helped in designing the structure of cellulose. In modern era, latest equipments with surprising technology have made analysis of glycans much more rapid and sensitive. There are many different classes of Glycans depending upon the nature of conjugates or the position at which they are glycosylated. Currently, scientific community is focusing one of the most challenging perspectives of understanding the structure of glycoconjugates and linkages present between them. Whenever the glycomic analysis starts, it begins with the fractionation and purification of the target glycan.

6.1 Fractionation

The first step on the way to carbohydrate analysis is separation of our target glycan. The method followed depends upon the nature, i.e., monosaccharide, disaccharide, or glycoconjugate. As carbohydrates are soluble in alcohol, we use ethanol to solubilize the carbohydrate and then filter by filter paper. Sometime, carbohydrates can be polar or nonpolar, so we can utilize ion exchange columns for separation of these molecules (Smith 1967).

When targeting non-structural carbohydrates in plants without any involvement in plant structure. Their content differs for different seasons due to adaptive behavior of plants (Teace and Fogel 2007). In that case we use grinding of material in liquid nitrogen that will stop all kinds of activity in plants. It is ground till it becomes too small less that 50 μm as it will facilitate extraction of soluble carbohydrates with best yield (Gomez et al. 2003).

6.2 Hydrolysis

It is one of the oldest and most basic procedures carried out for the oligosaccharides and polysaccharides to be converted into small subunits. When targeting starch we can also utilize enzymatic cleavage. Enzymes used are amylase or amyloglucosidase (Palacio et al. 2007). Chemical hydrolysis can also be done if target carbohydrates are not sensitive to chemicals. Mostly concentrated acids are used to break down polysaccharides, but such acids can also damage the monomer that helps in quantification, leading to error in analyses. Using a dilute acid might be helpful as in the case of 1 % HCl (Raessler et al. 2008).

6.3 Purification

Before doing mass spectrometric analysis on glycans, it is good to purify our desired glycan from the crude hydrolyzed sample. Carbohydrates are less tolerant to salts and other chemicals as compared to proteins. A very small fraction of metallic alkali is required for proper ionization, but it can affect resolution (Chen et al. 1997). Glycans can be purified by the use of solid phase extraction as well (Kussmann et al. 1997).

6.4 Analysis of Carbohydrates

It is the last step done in glycomic approach. This step depends upon the type of carbohydrates we are going to analyze. If we are studying the glycan portion of a glycoprotein or glycolipid, it would be different than the study conducted on simple glucose, starch, or any other sugar.

6.5 Separation Techniques

Most of the separation techniques involve the use of SDS-PAGE, thin layer chromatography, and column chromatography. Carbohydrates are hydrophilic in nature and bear a number of isomers, so these molecules require a specialized type of equipment.

6.5.1 Capillary Electrophoresis

Capillary electrophoresis is one of the most widely used separation techniques for the analysis of biomolecules. It has a great degree of improvement and modification for any desired purpose. Capillaries are narrow channels that provide high resistance to the molecules under high voltage, resulting in a very quick separation of the target molecule (Houel et al. 2014) It has mostly been used for the separation of N-glycans (Roxana et al. 2014) This technique is widely used for the analysis of nucleic acids like DNA but for glycomic study, already purified sample is tagged with a probe mostly laser-induced fluorescent probes, and these probes facilitate the detection of glycan at neutral pH even at very low concentrations (Vanderschaeghe et al. 2010). Glycans lack the ability to absorb light with wavelengths above 200 nm, and labeling of these molecules is essential for detection even if they are present in very small quantities. The major class of glycans is neutral at certain pH values, so these molecules can be charge-tagged to enable their movement in an electric field (Reusch et al. 2014). Many fluorophores are recommended for the separation which include 8-aminonaphthalene-1,3,6-trisulfonic acid (Guttman and Pritchett 1995; Guttman and Starr 1995) and 8-aminopyrene-1,3,6-tri-sulfonic acid (APTS; Guttman and Pritchett 1995).

With the advancement in technique, a latest fluorophore 5-aminonaphthalene-2-sulfonic acid (ANSA) is suggested for the labeling of glycan molecules (Briggs et al. 2009). APTS is mostly preferred for glycan analysis for bearing the trisulfate group, hence providing greater negative charge to molecules and leading to rapid separation. ASNA has a single sulfate group, and it has been suggested that it gives better results in capillary electrophoresis and is compatible with cadmium/helium laser. When performing CE along with HPLC and MS analysis, ANSA is preferred, while if the only objective is separation of molecules, then it is better to use APTS fluorophore. In the analysis of glycoconjugates, the glycan part is released from its aglycone part with the help of enzymes that break down the peptide linkage present, and the most common enzyme used for this is peptide N-glycosidase as it has a very broad substrate affinity (Kita et al. 2007). O-glycans are specifically more susceptible to degradation after being released from the counterpart, so they need to be reduced in order to get accurate results; but recently a new method has been introduced that rapidly releases O-glycan part within minutes (Yamada et al. 2007). Glycans separated in the CE are directly subjected to detection by laser detector in the form of different peaks depending upon the retention time of each glycan and is compared to a standard like maltose used as ladder.

Another simple version of CE is capillary zoned electrophoresis which makes use of direct UV detection and allows the quantification of carbohydrates as well. It is mostly utilized for the separation and quantification of basic structural carbohydrates like cellulose, hemicelluloses, and other pectin compounds (Fengel and Wegener Wood 1989). After hydrolytic and enzymatic breakdown of these complex carbohydrates, the monosaccharides are subjected to detection as they have a significant role after separation of carbohydrates. Carbohydrates consist of a number of OH groups, so these molecules are subjected to ionization and for this purpose a strong alkali is used. Direct UV detection involves complex formation with borate that increases the absorption of UV by many monosaccharides and disaccharides (Hoffstetter-Kuhn et al. 1991). A complex can also be formed with copper (Cu) to enhance the absorbance of UV, hence leading to detection of very small amount of glycans (Townsend 1993). Capillary electrophoresis techniques have proven a valuable asset for glycomic analysis of plant materials with ease of separation and detection involving a number of modifications and improvement according to the desired analyses. It involves a number of detection systems that are more specific and sensitive to any target.

6.5.2 High Pressure Liquid Chromatography

High pressure liquid chromatography HPLC is the technique developed for analyses and characterization of carbohydrates present in different organisms especially in plants (Sharma et al. 2010). Separation of carbohydrate molecules has been a challenging aspect in glycomics due to the presence of isomers and anomeric forms (Dvorackova et al. 2014). In addition to these, complex branching and cross-linking within or with other side residues make it even more difficult to analyze and separate these molecules (Ikegami et al. 2008). It is one of the most common separation techniques in glycomic research for the separation of both neutral and charged glycans. It has an increased resolving power as compared to capillary electrophoresis. It can be modified according to other techniques either before or after them (Godin et al. 2013). Other separation techniques can also be applied for the purpose of plant glycome analysis, but best results have been obtained by usage of HPLC (Wei and Ding 2000). In the analyses of carbohydrates, HPLC has an advantageous property that it bears the utilization of a wide range of stationary phases in order to make the analyses more accurate and according to the target analyte (Shanmugavelan et al. 2013). It may utilize an anion or cation exchange column depending on what type of carbohydrates we are going to analyze (Schuster-Wolff-Buhring et al. 2011; Suzuki 2014).

The detection system plays a critical role in analyses as it entirely depends on the type saccharide to be analyzed and nature of side residues present. In capillary electrophoresis, a number of detectors have been discussed that have some flaws and requirement of specific probes for the detection of complex carbohydrates. In HPLC, refractive index of molecules is an important phenomenon in order to detect and analyze them. HPLC-RID is a simple detection method, and it has a broad range

with respect to its compatibility to a number of complex saccharides, but this detection method also has some drawbacks including the lengthy procedure in stabilizing the baseline in separation column and very poor sensitivity (Wei et al. 1996). Another widely used detection method is evaporative light-scattering detection (ELSD), mostly used for the compounds that are less volatile as compared to the mobile phase used. So using a mobile phase that is highly volatile can help in the separation of a number of carbohydrates. It is applied in the analyses of compounds that are not compatible to strong fluorophores and also in case when UV detection cannot be applied to analytes (Kuang et al. 2011). A comparison done by Hernandez et al. (1998) between RID and ELSD detection methods with respect to the detection of major saccharides present in plant showed clear evidence that ELSD was far more better in the perspective of sensitivity and detection limit ranges. Another important factor that makes ELSD preferable over RID is the short time analysis as compared to laborious procedure of the latter one.

ELSD detector brings about detection by taking into account a number of factors including gas flow rate and drift tube temperature. The first part of detection methodology is nebulization in which the eluent in the column is converted into droplets and then subjected to drying of droplets with the help of air flow or that of nitrogen. After being nebulized and dried they are subjected to the drift tube where evaporation of the mobile phase takes place, leaving behind particles containing our target mixture. Temperature provided for the drying and evaporation of the droplets and the mobile phase respectively depends upon the size of droplets and nature of the mobile phase (Kohler et al. 1997). Then the mixture is subject to radiations from a light source, leading to the scattering of the radiations depending solely on the mass of molecules, thus detecting the carbohydrates in the form of spectra depicting the degree of scattering with respect to mass (Trones et al. 1998).

The stationary phase also plays a crucial role in chromatographic analyses and detection of carbohydrates. A number of stationary phases have been used depending on the compatibility with the mobile phase used. When our target molecule is a monosaccharide or disaccharide, we use amine bonded silica gel column as it is more compatible with RID detector. For ELSD, there are other stationary phases suggested to make the analytical procedure simple and free from any kind of flaws (Hernandez et al. 1998).

6.6 Mass Spectrometry in Glycomic Analysis

Mass spectrometry has been widely applied in proteomic studies of plants and other organisms (Packer and Harrison 1998). MS mostly utilizes the mass-to-charge ratio; carbohydrate isomers have the same mass with a difference in structure so it might be a problem for two isomers of a glycan.

In analysis, most of the glycans are bound to either lipid or protein forming a conjugate. In glycomics we need an in-depth knowledge of the technique we are going to apply and what kind of modifications can be made (Zaia 2010). Sample

preparation for MS is a critical step in glycome analysis, as glycome is quite complex in nature so pre-fractionation must be done to make the MS analysis much easier and accurate (Kailemia et al. 2014). In MS, two techniques are used for glycome analysis: marker assisted laser desorption/ionization and electro-spray ionization (Snovida and Perreault 2007), which are helpful in determining the nature and specific properties of carbohydrate of interest.

Mass spectrometry is used to demonstrate structural characterization of carbohydrates on the basis of their molecular mass and mass of molecular ions (Pagel 2013). First of all molecules are converted into gaseous state, then ionized and then measured by their charge-to-mass ratio which is the primary way of identifying carbohydrates. In order to check the correct identification of carbohydrates masses of molecules which were used in the samples are then used by Virtual Expert Mass Spectrometrist (VEMS) program to search the database of carbohydrates.

If the molecule which is going to be analyzed is a conjugate, then first the carbohydrate portion from glycoconjugate is removed either by using enzymes (enzymatic method) or by chemicals (chemical method) (Mass spectrometry data analysis in proteomics, "Plant Metabolomics and Strategies" chapter).

For the determination of specific properties of carbohydrates matrix-assisted laser desorption ionization-mass spectrometry (MALDI-MS) is used extensively (Zhang et al. 2011). This method has the same mechanism and process as discussed above with a difference of using laser beam for ionizing sample molecules instead of high energy electrons and unlike simple mass spectrometry, this technique finally gives singly charged ions. MALDI-mass spectrometry actually is combined with time-of-flight (TOF) in which the velocity of molecular ions/ions is dependent on both mass and charge. MALDI mass spectrometry is not that much sensitive to contaminants in the sample relative to simple MS.

6.7 Microarray Based Methods

Plants are rich in variety of complex carbohydrates that form structural components and maintain the integrity of plant cells and other organelles. Plant cell wall is rich in one of the most abundant molecules on earth (Carpita and Gibeaut 1993). Carbohydrates being the vector of energy in plants play critical role in the development and growth of plant, so analysis of carbohydrates make it possible to predict the plant growth level under different conditions and stresses. Unlike nucleic acids, carbohydrates cannot be sequenced and hence it requires a different strategy to characterize and understand the interaction and nature of glycans. Microarray is a recent development in the field of analytical approach for the detection of different biomolecules, and this technique has revolutionized the area of proteomics, genomics, and glycomics studies by introducing high-throughput analysis (McWilliam et al. 2011). This technique was first introduced with publication in year 1995 (Schena et al. 1995), and subsequently its area of application spread widely from analyzing small molecules to complex ones along with their interaction and

linkages with other molecules (Park et al. 2008). All microarray techniques are based on the same principle which involves the immobilization of samples onto a membrane-like plate by formation of covalent or non-covalent bond (Rillahan and Paulson 2011). Plant polysaccharides are mostly immobilized non-covalently onto the surface by using adsorption phenomenon and the surfaces used are mostly nitrocellulose membranes (Willats et al. 2002). In case of oligosaccharides, adsorption does not facilitate the non-covalent immobilization to surface, so they are mostly bound covalently with the help of chemical linkers (Feizi 2000). Recently carbohydrate microarray has been applied to study important complexes that are involved in the glycosyltransferase activity in plants (Shipp et al. 2008). In this technique multiple probe based detection is done. Lectin microarray is the recent development in the field of glycomic analysis. Lectin is a protein that is known to bind with a number of other carbohydrates, helping in identification of glycomes. These proteins are strongly bound to the surface in the form of discrete support. A fluorescent dye is used that will give fluorescence if any interaction between lectin and carbohydrates occurs.

Generally carbohydrate microarrays can be divided into two basic methodologies. One is extracted glycan arrays that are applied for the determination of relative abundance of specific carbohydrates in a large amount of sample. It is mostly used for the analysis of carbohydrates present in different parts like leaves, roots, flowers, and stem. Under different conditions, the content of carbohydrates varies significantly, leading to a plant's adaptation to extant abnormal stresses. In some of the studies, this methodology also led to the discovery of novel carbohydrates in a number of plant species that have been previously believed to be present in only limited families of plants (Sørensen et al. 2008). Another approach in microarray is defined glycan arrays that makes use of the known carbohydrates, and these known molecules are used to determine antibody or enzymatic interactions.

7 Future Prospects

Plant glycomics is progressing continuously at a great pace in many aspects. With advancements in technological era, revolutionary improvements in glycome analysis and methodologies have made it even more rapid. The fact that new analytical techniques are being introduced or many are about to be applied in this field will make it easier to analyze even small samples with much more accuracy and specificity. In future, a lab-on-chip is expected that will be able to analyze nano-quantities (Nelson et al. 2001) of sample involving nano-separation with high resolution. Within few years, developments will lead to characterization of unknown glycoproteins and glycolipids that are important structural components of plants in an easy and specific way. In case of recent analytical methods like microarrays, it has been possible to discover some of the novel carbohydrates that were not reported previously. Advancements in detectors used in HPLC and CE would result in accurate and precise results, leading to low errors count. Due to the presence of a number of complex

glycoconjugates, improvements are necessary to analyze these complexes and also to find out the interactions and linkages present in plants. Under different environmental conditions the glycan level of plants varies depending upon the type of change and the part of plant that is affected. Understanding this content variation will make it possible to find out the adaptive behavior of plants under different stress conditions and will also lead to the determination of physiological impact of carbohydrates on plants. Glycomic approach in future will also lead to a better understanding of marine plants and it will reveal some of the valuable aspects of these plants that have been conserved for millions of years. The development of a number of bioinformatics tools that are related to carbohydrates will make data comparison and retrieval easier, hence making our glycomic approach fast and also resulting in storage of vast data. Plant carbohydrate databases are essential for the global scientific community to access the recent developments and newly characterized glycans, and they will make current glycomics more rapid and precise with the use of software, predicting interactions of carbohydrates with other non-carbohydrate molecules.

References

Antosiewicz DM, Purugganan MM, Polisensky DH, Braam J (1997) Cellular localization of *Arabidopsis* xyloglucanendotransglycosylase-related proteins during development and after wind stimulation. Plant Physiol 115:1319–1328

Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1473:4–8

Arioli T et al (1998) Molecular analysis of cellulose biosynthesis in *Arabidopsis*. Science 279:717–720

Aspinall GO (1980) Chemistry of cell wall polysaccharides. In: Preiss J (ed) The biochemistry of plants (a comprehensive treatise). Carbohydrates: structure and function, vol 3. Academic, New York, NY, pp 473–500

Atmodjo MA, Hao Z, Mohnen D (2013) Evolving views of pectin biosynthesis. Annu Rev Plant Biol 64:747–779

Austin JR, Frost E, Vidi PA, Kessler F, Staehelin LA (2006) Plastoglobules are lipoprotein subcompartments of the chloroplast that are permanently coupled to thylakoid membranes and contain biosynthetic enzymes. Plant Cell Online 18(7):1693–1703

Baker EA (1982) Chemistry and morphology of plant epicuticular waxes. In: Cutler DF, Alvin KL, Price CE (eds) The plant cuticle. Academic, London, pp 139–165

Baldwin L, Domon JM, Klimek JF, Fournet F, Sellier H, Gillet F, Pelloux J, Lejeune-Hénaut I, Carpita NC (2014) Catherine Rayon structural alteration of cell wall pectins accompanies pea developmentin response to cold. Phytochemistry 104:37–47

Bashline L, Du J, Gu Y (2011) The trafficking and behavior of cellulose synthase and a glimpse of potential cellulose synthesis regulators. Front Biol 6:377–383

Bashline L, Li S, Anderson CT, Lei L, Gu Y (2013) The endocytosis of cellulose synthase in Arabidopsis is dependent on mu2, a clathrin mediated endocytosis adaptin. Plant Physiol 163(1):150–160

Baskin TI, Gu Y (2012) Making parallel lines meet: transferring information from microtubules to extracellular matrix. Cell Adh Migr 6:404–408

Basu D, Liang Y, Liu X, Himmeldirk K, Faik A, Kieliszewski M et al (2013) Functional identification of a hydroxyproline-o-galactosyltransferase specific for arabinogalactan protein biosynthesis in Arabidopsis. J Biol Chem 288:10132–10143

Blevins DG (1998) Boron in plant structure and function. Annu Rev Plant Physiol Plant Mol Biol 49:481–500

Block MA, Dorne AJ, Joyard J, Douce R (1983) Preparation and characterization of membrane fractions enriched in outer and inner envelope membranes from spinach chloroplasts. II. Biochemical characterization. J Biol Chem 258(21):13281–13286

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Ann Rev Plant Biol 54(1):519–549

Bonza MC, De Michelis MI (2011) The plant Ca2+-ATPase repertoire: biochemical features and physiological functions. Plant Biol (Stuttgart) 13:421–430

Bootten TJ, Harris PJ, Melton LD, Newman RH (2011) Using solid-state C-13 NMR spectroscopy to study the molecular organisation of primary plant cell walls. Plant Cell Wall Meth Protoc 715:179–196

Bourquin V et al (2002) Xyloglucan endotransglycosylases have a function during the formation of secondary cell walls of vascular tissues. Plant Cell 14:3073–3088

Brabham C, Debolt S (2012) Chemical genetics to examine cellulose biosynthesis. Front Plant Sci 3:309

Braybrook SA, Peaucelle A (2013) Mechano-chemical aspects of organ formation in *Arabidopsis thaliana*: the relationship between auxin and pectin. PLoS One 8, e57813

Brett C, Waldron K (1990) In: Black M, Chapman J (eds) Physiology and biochemistry of plant cell walls. Unwin Hyman, London

Briggs JB, Keck RG, Ma S, Lau W, Jones AJ (2009) An analytical system for the characterization of highly heterogeneous mixtures of N-linked oligosaccharides. Anal Biochem 389:40–51

Bringmann M, Li E, Sampathkumar A, Kocabek T, Hauser MT, Persson S (2012) POM-POM2/cellulose synthase interacting is essential for the functional association of cellulose synthase and microtubules in Arabidopsis. Plant Cell 24:163–177

Brockhausen I, Schutzbach J, Kuhns W (1998) Glycoproteins and their relationship to human disease. Acta Anat (Basel) 161:36–78

Brummell DA, Harpster MH, Dunsmuir P (1999) Differential expression of expansin gene family members during growth and ripening of tomato fruit. Plant Mol Biol 39:161–169

Buchanan BB, Gruissem W, Jones RL (2000) Biochemistry and molecular biology of plants. American Society of Plant Biologists, Rockville, MD. ISBN 0-943088-37-2

Burén S, Ortega-Villasante C, Blanco-Rivero A et al (2011) Importance of post-translational modifications for functionality of a chloroplast localized carbonic anhydrase (CAH1) in Arabidopsis thaliana. PLoS One 6:1–15

Burn JE, Hocart CH, Birch RJ, Cork AC, Williamson RE (2002) Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. Plant Physiol 129:797–807

Burton RA, Shirley NJ, King BJ, Harvey AJ, Fincher GB (2004) The *Ces*Agene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol 134:224–236

Cação SM, Leite TF, Budzinski IG, Santos TB, Scholz MB, Carpentieri-Pipolo V, Domingues DS, Vieira LG, Pereira LF (2012) Gene expression and enzymatic activity of pectin methylesterase during fruit development and ripening in Coffea arabica L. Genet Mol Res 11:3186–3197

Camejo D, Martí MC, Jiménez A, Cabrera JC, Olmos E, Sevilla F (2011) Effect of oligogalacturonides on root length, extracellular alkalinization and O2− accumulation in alfalfa. J Plant Physiol 168:566–575

Cao J (2012) The pectin lyases in Arabidopsis thaliana: evolution, selection and expression proiles. PLoS One 7, e46944

Cao Y, Junling L, Li Y, Guohua C, Guo H, Ruibo H, Guang Q, Yingzhen K, Chunxiang F, Gongke Z (2014) Cell wall polysaccharide distribution in Miscanthus lutarioriparius stem using immuno-detection. Plant Cell Rep 33:643–653

Carpita NC, Gibeaut DM (1993) Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. Plant J 3:1–30

Carroll A, Specht CD (2011) Understanding plant cellulose synthases through a comprehensive investigation of the cellulose synthase family sequences. Front Plant Sci 2:5

Chabannes M et al (2001) *In situ* analysis of lignins in transgenic tobacco reveals a differential impact of individual transformations on the spatial patterns of lignin deposition at the cellular and subcellular levels. Plant J 28(3):271–282. doi:10.1046/j.1365-313X.2001.01159.x

Chebli Y, Kaneda M, Zerzour R, Geitmann A (2012) The cell wall of the Arabidopsis pollen tube—spatial distribution, recycling, and network formation of polysaccharides. Plant Physiol 160:1940–1955

Chen P, Baker AG, Novotny MV (1997) Anal Biochem 244:144

Cho HT, Cosgrove DJ (2000) Altered expression of expansin modulates leaf growth and pedicel abscission in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 97:9783–9788

Cho HT, Cosgrove DJ (2002) Regulation of root hair initiation and expansin gene expression in *Arabidopsis*. Plant Cell 14:3237–3253

Cho HT, Kende H (1997) Expression of expansin genes is correlated with growth in deepwater rice. Plant Cell 9:1661–1671

Choi D, Lee Y, Cho HT, Kende H (2003) Regulation of expansin gene expression affects growth and development in transgenic rice plants. Plant Cell 15:1386–1398

Cosgrove DJ (1989) Characterization of long-term extension of isolated cell walls from growing cucumber hypocotyls. Planta 177:121–130

Crawford RL (1981) Lignin biodegradation and transformation. John Wiley and Sons, New York, NY. ISBN 0-471-05743-6

Dennis JW, Granovsky M, Warren CE (1999) Protein glycosylation in development and disease. Bioessays 21:412–421

Dhugga KS et al (2004) Guar seed®-mannan synthase is a member of the cellulose synthase super gene family. Science 303:363–366

Dilokpimol A, Geshi N (2014) Arabidopsis thaliana glucuronosyltransferase in family GT14. Plant Signal Behav 9, e28891. doi:10.4161/psb.28891

Dilokpimol A, Poulsen CP, Vereb G, Kaneko S, Schulz A, Geshi N (2014) Galactosyltransferases from Arabidopsis thaliana in the biosynthesis of type II arabinogalactan: molecular interaction enhances enzyme activity. BMC Plant Biol 14:90

Dube DH, Bertozzi CR (2005) Glycans in cancer and inflammation—potential for therapeutics and diagnostics. Nat Rev Drug Discov 4:477–488

Dumont M, Arnaud L, Sophie B, Marie C, Aline V, Kiefer-Meyer MC, Jerome P, Patrice L, Jean-Claude M (2014) The cell wall pectic polymer rhamnogalacturonan-II is required for proper pollen tube elongation: implications of a putative sialyltransferase-like protein. Ann Bot 114:1177–1188

Dvorackova E, Marie S, Hrdlicka E (2014) Carbohydrate analysis: from sample preparation to HPLC on different stationaryphases coupled with evaporative light-scattering detection. J Sep Sci 37:323–337

Edidin M (2003) The state of lipid rafts: from model membrane to cells. Annu Rev Biophys Biomol Struct 32:257–283

Erikkson O, Lindgren BO (1977) About the linkage between lignin and hemicelluloses in wood. Sven Papperstidn 80(2):59–63

Fagard M et al (2000) *PROCUSTE1* encodes a cellulose synthase required for normal cell elongation specifically in roots and dark-grown hypocotyls of *Arabidopsis*. Plant Cell 12:2409–2424

Fanutti C, Gidley MJ, Reid JSG (1993) Action of a pure xyloglucanendo-transglycosylase (formerly called xyloglucan-specific endo-1,4-β-D-glucanase) from the cotyledons of germinated nasturtium seeds. Plant J 3:691–700

Farkas V, Sulova Z, Stratilova E, Hanna R, Maclachlan G (1992) Cleavage of xyloglucan by nasturtium seed xyloglucanase and transglycosylation to xyloglucan subunit oligosaccharides. Arch Biochem Biophys 298:365–370

Feizi T (2000) Progress in deciphering the information content of the 'glycome'—a crescendo in the closing years of the millennium. Glycoconj J 17:553–565

Fengel D, Wegener Wood G (1989) Chemistry, ultrastructure and reactions. Walter de Gruyter, Berlin, p 613

Ferrari S, Savatin D, Sicilia F, Gramegna G, Cervone F, de Lorenzo G (2013) Oligogalacturonides: plant damage-associated molecular patterns and regulators of growth and development. Front Plant Sci 4:1–9

Freudenberg K, Neish AC (1968) Constitution and biosynthesis of lignin. In: Springer GF, Kleinzeller A (eds) Molecular biology biochemistry and biophysics. Springer, New York, NY, p 129

Fry SC (1988) The growing plant cell wall: chemical and metabolic analysis. Longman Scientific & Technical, Harlow, pp 103–185

Fry SC (1989) Cellulases, hemicelluloses and auxin-stimulated growth: a possible relationship. Physiol Plant 75:532–536

Fry SC (1998) Oxidative scission of plant cell wall polysaccharides by ascorbate-induced hydroxyl radicals. Biochem J 332:507–515

Fry SC et al (1992) Xyloglucanendotransglycosylase, a new wall-loosening enzyme activity from plants. Biochem J 282:821–828

Fry SC, Miller JG, Dumville JC (2002) A proposed role for copper ions in cell wall loosening. Plant Soil 247:57–67

Fu Y, Yin H, Wanga W, Wang M, Zhang H, Zhao X et al (2011) β-1,3-glucan with different degree of polymerization induced different defense responses in tobacco. Carbohydr Polym 86:774–782

Fujita M, Lechner B, Barton DA, Overall RL, Wasteneys GO (2012) The missing link: do cortical microtubules define plasma membrane nanodomains that modulate cellulose biosynthesis? Protoplasma 249(Suppl 1):S59–S67

Fujita M, Himmelspach R, Ward J, Whittington A, Hasenbein N, Liu C, Truong TT, Galway ME, Mansfield SD, Hocart CH, Wasteneys GO (2013) The anisotropy1 D604N mutation in the Arabidopsis cellulose synthase1 catalytic domain reduces cell wall crystallinity and the velocity of cellulose synthase complexes. Plant Physiol 162:74–85

Gabius H-J (2011) Glycobiomarkers by glycoproteomics and glycan profiling (glycomics): emergence of functionality. Biochem Soc Trans 39:399–405

Gardner S, Burrell MM, Fry SC (2002) Screening of Arabidopsis thaliana stems for variation in cell wall polysaccharides. Phytochemistry 60:241–254

Gaspar Y, Johnson KL, McKenna JA, Bacic A, Schultz CJ (2001) Plant Mol Biol 47(1):161

Geshi N, Johansen JN, Dilokpimol A, Rolland A, Belcram K, Verger S et al (2013) A galactosyltransferase acting on arabinogalactan protein glycans is essential for embryo development in Arabidopsis. Plant J 76:128–137

Gillard BK, Thurmon LT, Marcus DM (1993) Variable subcellular localization of glycosphingolipids. Glycobiology 3:57–67

Gille S, Pauly M (2012) O-Acetylation of plant cell wall polysaccharides. Front Plant Sci 3:1–7

Godin B, Lamaudiere S, Agneessens R, Schmit T, Goffart J-P, Stilmant D, Gerin PA, Delcarte J (2013) Ind Crops Prod 48:1–12

Gomez L, Jordan MO, Adamowicz S, Leiser H, Pages L (2003) Du prélèvement au dosage: réflexions sur les problèmes posés par la mesure des glucides non structuraux chez les végétaux ligneux. Cah Agric 12:369

Gou J, Miller LM, Hou G, Yu X, Chen X, Liu C (2012) Acetylesterase mediated deacetylation of pectin impairs cell elongation, pollen germination, and plant reproduction. Plant Cell 24:50–65

Gough C, Cullimore J (2011) Lipo-chitooligosaccharide signaling in endosymbiotic plant-microbe interactions. Mol Plant Microbe Interact 24(8):867–878

Guttman A, Pritchett T (1995) Capillary gel electrophoresis separation of high-mannose type oligosaccharides derivatized by 1-aminopyrene-3,6,8-trisulfonic acid. Electrophoresis 16:1906–1911

Guttman A, Starr C (1995) Capillary and slab gel electrophoresis profiling of oligosaccharides. Electrophoresis 16:993–997

Hager A, Menzel H, Krauss A (1971) Versuche und hypothesezurprimaerwirkung des auxinsbeimstreckungswachstum. Planta 100:47–75

Hahn MG, Darvill AG, Albersheim P (1981) Plant Physiol 68(5):1161. doi:10.1104/pp.68.5.1161
Haltiwanger RS, Lowe JB (2004) Role of glycosylation in development. Annu Rev Biochem 73:491–537
Harpaz-Saad S, Western TL, Kieber JJ (2012) The FEI2-SOS5 pathway and CELLULOSE SYNTHASE 5 are required for cellulose biosynthesis in the Arabidopsis seed coat and affect pectin mucilage structure. Plant Signal Behav 7:285–288
Harris DM, Corbin K, Wang T, Gutierrez R, Bertolo AL, Petti C, Smilgies DM, Estevez JM, Bonetta D, Urbanowicz BR, Ehrhardt DW, Somerville CR, Rose JK, Hong M, Debolt S (2012) Cellulose microfibril crystallinity is reduced by mutating C-terminal transmembrane region residues CESA1A903V and CESA3T942I Of cellulose synthase. Proc Natl Acad Sci U S A 109:4098–4103
Hayashi T (1989) Xyloglucans in the primary cell wall. Annu Rev Plant Physiol Plant Mol Biol 40:139–168
Hernandez JL, Gonzalez-Castro MJ, Alba IN, de la Garcia Cruz C (1998) Performance liquid chromatographic determination of mono- and oligosaccharides in vegetables with evaporative light-scattering detection and refractive index detection. J Chromatogr Sci 36:292–298
Hoffstetter-Kuhn S, Paulus A, Gassmann E, Widmer WH (1991) Influence of borate complexation on the electrophoretic behavior of carbohydrates in capillary electrophoresis. Anal Chem 63:1541
Holzl G, Dormann P (2007) Structure and function of glycoglycerolipids in plants and bacteria. Prog Lipid Res 46:225–243
Hon DNS (1994) Cellulose: a random walk along its historical path. Cellulose 1:1–25
Houel S, Hilliard M, Yu YQ, MacLoughlin N, Martin SM, Rudd PM, Williams JP, Chen W (2014) Anal Chem 86:576–584
Hubbard CS, Ivatt RJ (1981) Annu Rev Biochem 50:555–583
Ikegami T, Horie K, Saad N, Oliver Fiehn KH, Tanaka N (2008) Anal Bioanal Chem 391:2533–2542
Immerzeel P, Eppink MM, de Vries SC, Schols HA, Voragen AGJ (2006) Physiol Plant 128(1):18
Ishii TT (1995) Mokuzai Gakkaishi 41(7):669
Ishii T, Matsunaga T, Pellerin P, O'Neill MA, Darvill A, Albersheim P (1999) The plant cell wall polysaccharide rhamnogalacturonan II self-assembles into a covalently cross-linked dimer. J BiolChem 274(19):13098. doi:10.1074/jbc.274.19.13098
Itano N, Kimata K (2002) Mammalian hyaluronan synthases. IUBMB Life 54:195–199
Jayo RG, Thaysen-Andersen M, Lindenburg PW, Haselberg R, Hankemeier T, Ramautar R, Chen DD (2014) Simple capillary electrophoresis-mass spectrometry method for complex glycan analysis using a flow-through microvial interface. Anal Chem 86:6479–6486
Kailemia MJ, Ruhaak LR, Lebrilla CB, Amste IJ (2014) Oligosaccharide analysis by mass spectrometry: a review of recent developments. Anal Chem 86:196–212
Kelly AA, Dormann P (2004) Green light for galactolipid trafficking. Curr Opin Plant Biol 7:262–269
Kimura S et al (1999) Immunogold labeling of rosette terminal cellulose-synthesizing complexes in the vascular plant *Vigna angularis*. Plant Cell 11:2075–2085
Kita Y, Miura Y, Furukawa J, Nakano M, Shinohara Y, Ohno M, Takimoto A, Nishimura S (2007) Quantitative glycomics of human whole serum glycoproteins based on the standardized protocol for liberating N-glycans. Mol Cell Proteomics 6:1437–1445
Kjellbom P, Larsson C (1984) Preparation and polypeptide composition of chlorophyll-free plasma membranes from leaves of light-grown spinach and barley. Physiol Plant 62:501–509
Knoch E, Dilokpimol A, Geshi N (2014) Arabinogalactan proteins: focus on carbohydrate active enzymes. Front Plant Sci 5:198
Kohler M, Haerdi W, Christen P, Veuthey JL (1997) Trends Anal Chem 16:475–484
Koller A, O'Neil MA, Darvill AG, Albersheim P (1991) Phytochemistry 30:3903
Krzeminski M (2011) Human galectin-3 (Mac-2 antigen):defining molecular switches of affinity to natural glycoproteins, structural and dynamic aspects of glycan binding by flexible ligand

docking and putative regulatory sequences in the proximal promoter sequence. Biochim Biophys Acta 1810:150–161
Kuang H, Xia Y, Liang J, Yang B, Wang Q, Sun Y (2011) Carbohydr Polym 84:1258–1266
Kussmann M, Nordhoff E, Rahbek-Nielsen H, Haebel S, Rossel-Larsen M, Jakobsen L, Gobom J, Mirgorodskaya E, Kroll-Kristensen A, Palm L, Roepstroff P (1997) J Mass Spectrom 32:593
Lamport DTA, Kieliszewksi MJ, Chen Y, Cannon MC (2011) Role of the extension superfamily in primary cell wall architecture. Plant Physiol 156:11–19
Lane DR et al (2001) Temperature-sensitive alleles of RSW2 link the KORRIGAN endo-1,4-β-glucanase to cellulose synthesis and cytokinesis in *Arabidopsis*. Plant Physiol 126:278–288
Lee Y, Kende H (2001) Expression of β-expansins is correlated with internodal elongation in deepwater rice. Plant Physiol 127:645–654
Lee KJD, Cornuault V, Manfield IW, Ralet M-C, Knox JP (2013) Multi-scale spatial heterogeneity of pectic rhamnogalacturonan I (RG–I) structural features in tobacco seed endosperm cell walls. Plant J 75:1018–1027
Lerouge P, O'Neill MA, Darvill AG, Albersheim P (1993) Carbohydr Res 243(2):359
Liepman AH, Wilkerson CG, Keegstra K (2005) Expression of cellulose synthase-like (*Csl*) genes in insect cells reveals that *CslA* family members encode mannan synthases. Proc Natl Acad Sci U S A 102:2221–2226
Liszkay A, Schopfer P (2003) Plasma membrane-generated superoxide anion radicals and peroxidase-generated hydroxyl radicals may be involved in the growth of coleoptiles. Free Radic Res 37:26–27
Liszkay A, Kenk B, Schopfer P (2003) Evidence for the involvement of cell wall peroxidase in the generation of hydroxyl radicals mediating extension growth. Planta 217:658–667
Liu J, Shirota O, Novotny MV (1992) Anal Chem 64:973
Structure and morphology of cellulose by Serge Pérez and William Mackie, CERMAV-CNRS, 2001. Chapter IV
Marga F, Grandbois M, Cosgrove DJ, Baskin TI (2005) Cell wall extension results in the coordinate separation of parallel microfibrils: evidence from scanning electron microscopy and atomic force microscopy. Plant J 43:181–190
Marth JD, Grewal PK (2008) Mammalian glycosylation in immunity. Nat Rev Immunol 8:874–887
Matsui A et al (2005) AtXTH27 plays an essential role in cell wall modification during the development of tracheary elements. Plant J 42:525–534
McBride HM, Neuspiel M, Wasiak S (2006) Mitochondria: more than just a powerhouse. Curr Biol 16(14):551–560
McCarthy TW, Der JP, Honaas LA, de Pamphilis CW, Anderson CT (2014) Phylogenetic analysis of pectin-related gene families in *Physcomitrella patens* and nine other plant species yields evolutionary insights into cell walls. BMC Plant Biol 14:79
McMillan JD (1993) Pretreatment of lignocellulosic biomass. In: Himmel ME, Baker JO, Overend RP (eds) Enzymatic conversion of biomass for fuel production. American Chemical Society, Washington, DC, pp 292–323
McQueen-Mason S, Cosgrove DJ (1994) Disruption of hydrogen bonding between plant cell wall polymers by proteins that induce wall extension. Proc Natl Acad Sci U S A 91:6574–6578
McQueen-Mason SJ, Cosgrove DJ (1995) Expansin mode of action on cell walls. Analysis of wall hydrolysis, stress relaxation, and binding. Plant Physiol 107:87–100
McQueen-Mason S, Durachko DM, Cosgrove DJ (1992) Two endogenous proteins that induce cell wall expansion in plants. Plant Cell 4:1425–1433
McWilliam I, Chong Kwan M, Hall D (2011) Inkjet printing for the production of protein microarrays. Methods Mol Biol 785:345–361
Merali Z, Ho JD, Samuel RAC, Le Gall G, Elliston A, Waldron KA (2013) Characterization of cell wall components of wheat straw following hydrothermal pretreatment and fractionation. Bioresour Technol 131:226–234
Miriam A, Elysa JRO, Jan HNS, Anne MCE, Tijs K (2011) Golgi body motility in the plant cell cortex correlates with actin cytoskeleton organization. Plant Cell Physiol 52(10):1844–1855

Mohnen D (1999) In: Barton D, Nakanishi K, Meth-Cohn O (eds) Comprehensive natural products chemistry. Elsevier, Dordrecht, pp 497–527
Mollet J, Leroux C, Dardelle F, Lehner A (2013) Cell wall composition, biosynthesis and remodeling during pollen tube growth. Plants 2:107–147
Moore JP, Fangel JU, Willats WG, Vivier MA (2014) Pectic-b(1,4)-galactan, extensin and arabinogalactan–protein epitopes differentiate ripening stages in wine and table grape cell walls. Ann Bot 114(6):1279–1294. doi:10.1093/aob/mcu053
Müller K, Levesque-Tremblay G, Bartels S, Weitbrecht K, Wormit A, Usadel B, Haughn G, Kermode AR (2013) Demethylesterification of cell wall pectins in Arabidopsis plays a role in seed germination. Plant Physiol 161:305–316
Nakamura A, Furuta H, Maeda H, Takao T, Nagamatsu Y (2002) Analysis of the molecular construction of xylogalacturonan isolated from soluble soybean polysaccharides. Biosci Biotechnol Biochem 66:1155–1158
Nelson BP, Grimsrud TE, Liles MR, Goodman RM (2001) Corn RM. Surface plasmon resonance imaging measurements of DNA and RNA hybridization adsorption onto DNA microarrays analytical chemistry 73:1–7
Nicol F et al (1998) A plasma membrane-bound putative endo-1,4-β-D-glucanase is required for normal wall assembly and cell elongation in *Arabidopsis*. EMBO J 17:5563–5576
Nishitani K, Tominaga T (1992) Endo-xyloglucantransferase, a novel class of glycosyltransferase that catalyzes transfer of a segment of xyloglucan molecule to another xyloglucan molecule. J Biol Chem 267:21058–21064
O'Neill MA, Eberhard S, Albersheim P, Darvill AG (2001) Science 294:846
Obro J, Harholt J, Scheller HV, Orfila C (2004) Phytochemistry 65:1429–1438
Ochoa-Villarreal M, Vargas-Arispuro I, Islas-Osuna MA, González-Aguilar G, Martínez-Téllez MA (2011) Pectin-derived oligosaccharides increase color and anthocyanin content in flame seedless grapes. J Sci Food Agric 91:1928–1930
Ochoa-Villarreal M, Aispuro-Hernández E, Vargas-Arispuro I, Martínez-Téllez MÁ (2012) Plant cell wall polymers: function, structure and biological activity of their derivatives. In: De Souza Gomes A (ed) Polymerization. InTech, Osaka, pp 64–89
Ohtsubo K, Marth JD (2006) Glycosylation in cellular mechanisms of health and disease. Cell 126:855–867
Pabst M, Fischl RM, Brecker L, Morelle W, Fauland A, Köfeler H, Altmann F, Léonard R (2013) Rhamnogalacturonan II structure shows variation in the side chains monosaccharide composition and methylation status within and across different plant species. Plant J 76: 61–72
Packer NH, Harrison MJ (1998) Glycobiology and proteomics: is mass spectrometry the holy grail? Electropherosis 19:1872–1882
Pagel K (2013) Ion mobility-mass spectrometry of complex carbohydrates: collision cross sections of sodiated n-linked glycans. Anal Chem 85:5138–5145
Palacio S, Maestro M, Montserrat-Martı G (2007) Environ Exp Bot 59:34
Palin R, Geitmann A (2012) The role of pectin in plant morphogenesis. Biosystem 109:397–402
Park S, Lee M-R, Shin I (2008) Carbohydrate microarrays as powerful tools in studies of carbohydrate-mediated biological processes. Chem Commun 2008:4389–4399
Payen A (1838) Me´moiresur la composition du tissupropre des planteset du ligneux. C R Hebd Seances Acad Sci 7:1052–1056
Pearl IW (1967) The chemistry of lignin pectins – a new hypothetical model. Marcel Dekker Inc, New York, NY, p 339
Peng L, Kawagoe Y, Hogan P, Delmer D (2002) Sitosterol-β-glucoside as primer for cellulose synthesis in plants. Science 295:147–150
Pien S, Wyrzykowska J, McQueen-Mason S, Smart C, Fleming A (2001) Local expression of expansin induces the entire process of leaf development and modifies leaf shape. Proc Natl Acad Sci U S A 98:11812–11817
Pornsak S (2003) Chemistry of pectin and its pharmaceutical uses: a review. Silpakorn Univ Int J 3(1–2):206

Purugganan MM, Braam J, Fry SC (1997) The *Arabidopsis* TCH4 xyloglucanendotransglycosylase. Substrate specificity, pH optimum, and cold tolerance. Plant Physiol 115:181–190
Raessler M, Wissuwa B, Breul A, Unger W, Grimm T (2008) J Agric Food Chem 56:7649
Raven P (2005) Biology of plants. Freeman and Co, Madison, NY, p 54
Rayle DL, Cleland RE (1970) Enhancement of wall loosening and elongation by acid solutions. Plant Physiol 46:250–253
Redgwell RJ, Fry SC (1993) Xyloglucanendotransglycosylase activity increases during kiwifruit (*Actinidia deliciosa*) ripening (implications for fruit softening). Plant Physiol 103: 1399–1406
Reinhardt D, Wittwer F, Mandel T, Kuhlemeier C (1998) Localized upregulation of a new expansin gene predicts the site of leaf formation in the tomato meristem. Plant Cell 10:1427–1437
Reusch D, Haberger M, Kailich T, Heidenreich AK, Kampe M, Bulau P, Wuhrer M (2014) High-throughput glycosylation analysis of therapeutic immunoglobulin G by capillary gel electrophoresis using a DNA analyzer. MAbs 6:185–196
Richmond TA, Somerville CR (2000) The cellulose synthase superfamily. Plant Physiol 124:495–498
Ridley BL, O'Neill MA, Mohnen D (2001a) Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. Phytochemistry 57:929. doi:10.1016/S0031-9422(01)00113-3
Ridley BL, O'Neill MA, Mohnen D (2001b) Complex carbohydrate research center and department of biochemistry and molecular biology. Phytochemistry 57:929. doi:10.1016/S0031-9422(01)00113-3
Rillahan CD, Paulson JC (2011) Glycan microarrays for decoding the glycome. Annu Rev Biochem 80:797–823
Roongsattham P, Morcillo F, Jantasuriyarat C et al (2012) Temporal and spatial expression of polygalacturonase gene family members reveals divergent regulation during fleshy fruit ripening and abscission in the monocot species oil palm. BMC Plant Biol 12:1–15
Rose JK, Braam J, Fry SC, Nishitani K (2002) The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature. Plant Cell Physiol 43:1421–1435
Roxana G, Jayo, Morten Thaysen-Andersen, Petrus W. Lindenburg (2014)
Scheible WR, Pauly M (2004) Glycosyltransferases and cell wall biosynthesis: novel players and insights. Curr Opin Plant Biol 7:285–295
Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470
Schmidt HW, Schönherr J (1982) Development of plant cuticles - occurrence and role of non-ester bonds in cutin of Clivia miniata Reg. leaves. Planta 156:380–384
Schopfer P, Liszkay A, Bechtold M, Frahry G, Wagner A (2002) Evidence that hydroxyl radicals mediate auxin-induced extension growth. Planta 214:821–828
Schuster-Wolff-Buhring R, Michel R, Hinrichs J (2011) Dairy Sci Technol 91:27–37
Sénéchal F, Wattier C, Rustérucci C, Pelloux J (2014) Homogalacturonan-modifying enzymes: structure, expression, and roles in plants. J Exp Bot 18:5125–5160
Shanmugavelan P, Kim SY, Kim JB, Kim HW, Cho SM, Kim SN, Kim SY, Cho YS, Kim HR (2013) Evaluation of sugar content and composition in commonly consumed Korean vegetables, fruits, cereals, seed plants, and leaves by HPLC-ELSD. Carbohydr Res 380:112–117
Sharma U, Bhandari P, Kumar N, Singh B (2010) Chromatographia 71:633–638
Sharon N (2006) Carbohydrates as future anti-adhesion drugs for infectious diseases. Biochim Biophys Acta 1760:527–537
Shibuya N, Iwasaki T (1985) Structural features of rice bran hemicellulose. Phytochemistry 24:285–289
Shipp M, Nadella R, Gao H, Farkas V, Sigrist H, Faik A (2008) Glyco-array technology for efficient monitoring of plant cell wall glycosyltransferase activities. Glycoconj J 25:49–58
Silbert JE, Sugumaran G (2002) Biosynthesis of chondroitin/dermatan sulfate. IUBMB Life 54:177–186
Smith D (1967) Crop Sci 7:62

Snovida SI, Perreault H (2007) A 2,5-dihydroxybenzoic acid/*N*,*N*-dimethylaniline matrix for the analysis of oligosaccharides by matrix-assisted laser desorption/ionization mass spectrometry. Rapid Commun Mass Spectrom 21:3711–3715

Sørensen I, Pettolino FA, Wilson SM, Doblin MS, Johansen B, Bacic A, Willats WGT (2008) Mixed linkage (1 3), (1 4)- β-D-glucan is not unique to the Poales and is an abundant component of *Equisetum arvense* cell walls. Plant J 54:510–521

Steele NM et al (2001) Ten isoenzymes of xyloglucanendotransglycosylase from plant cell walls select and cleave the donor substrate stochastically. Biochem J 355:671–679

Sugahara K, Kitagawa H (2002) Heparin and heparan sulfate biosynthesis. IUBMB Life 54:163–175

Suzuki S (2014) Highly sensitive methods using liquid chromatography and capillary electrophoresis for quantitative analysis of glycoprotein glycans. Chromatography 35:1–22

Taiz L (1984) Annu Rev Plant Physiol 35:585

Takeda T et al (2002) Suppression and acceleration of cell elongation by integration of xyloglucans in pea stem segments. Proc Natl Acad Sci U S A 99:9055–9060

Talbott LD, Ray PM (1992) Molecular size and separability features of pea cell wall polysaccharides. Implications for models of primary wall structure. Plant Physiol 92:357–368

Taniguchi N, Miyoshi E, Gu J, Honke K, Matsumoto A (2006) Decoding sugar functions by identifying target glycoproteins. Curr Opin Struct Biol 16:561–566

Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR (2003) Interactions among three distinct CesA proteins essential for cellulose synthesis. Proc Natl Acad Sci U S A 100:1450–1455

Teace MA, Fogel ML (2007) Org Geochem 38:458

Terao A, Hyodo H, Satoh S, Iwai H (2013) Changes in the distribution of cell wall polysaccharides in early fruit pericarp and ovule, from fruit set to early fruit development, in tomato (Solanum lycopersicum). J Plant Res 126:719–728

Thompson JE, Fry SC (1997) Trimming and solubilization of xyloglucan after deposition in the walls of cultured rose cells. J Exp Bot 48:297–305

Thompson JE, Fry SC (2001) Restructuring of wall-bound xyloglucan by transglycosylation in living plant cells. Plant J 26:23–34

Townsend RR (1993) Quantitative monosaccharide analysis of glycoproteins. In: Horvath C, Ettre LS (eds) Chromatography in biotechnology. American Chemical Society, Washington, DC, pp 86–101

Trones R, Andersen T, Hunnes I, Greibrokk T (1998) J Chromatogr A 814:55–61

Underwood W (2012) The plant cell wall: a dynamic barrier against pathogen invasion. Front Plant Sci 3:85

Vallarino JG, Osorio S (2012) Signaling role of oligogalacturonides derived during cell wall degradation. Plant Signal Behav 7:1447–1449

Van de Vis JW (1994) Characterization and mode of action of enzymes degrading galactan structures of arabinogalactans. Doctoral thesis. Wageningen University, Wageningen

Van Der Wal A, Leveau JH (2011) Modelling sugar diffusion across plant leaf cuticles: the effect of free water on substrate availability to phyllosphere bacteria. Environ Microbiol 13:792–797

Vanderschaeghe D, Festjens N, Delanghe J, Callewaert N (2010) Glycome profiling using modern glycomics technology: technical aspects and applications. Biol Chem 391:149–161

Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3:97–130

Varki A et al (1999) Essentials of glycobiology. Cold Spring Harbor Laboratory Press, New York, NY

Velasquez SM, Ricardi MM, Dorosz JG, Fernandez PV, Nadra AD, Pol-Fachin L, Egelund J, Gille S, Harholt J, Ciancia M (2011) O-Glycosylated cell wall proteins are essential in root hair growth. Science 332:1401–1403

Veytsman BA, Cosgrove DJ (1998) A model of cell wall expansion based on thermodynamics of polymer networks. Biophys J 75:2240–2250

Vriezen WH, De Graaf B, Mariani C, Voesenek LACJ (2000) Submergence induces expansin gene expression in flooding tolerant Rumex palustris and not in flooding intolerant R. acetosa. Planta 210:956–963

Wang L, Wang W, Wang Y-Q et al (2013a) Arabidopsis galacturonosyltransferase (GAUT) 13 and GAUT14 have redundant functions in pollen tube growth. Mol Plant 6:1131–1148

Wang H, Songshan S, Xuelan G, Chao Z, Guodong W, Hongwei W, Bin B, Hongwei F, Wuxia Z, Jinyou D, Shunchun W (2013b) Homogalacturonans from preinfused green tea: structural characterization and anticomplementary activity of their sulfated derivatives. J Agric Food Chem 61:10971–10980

Wang H, Guodong W, Fei L, Gautam B, Manoj J, Annie BSW, Songshan S, Hui L, Hongwei F, Xuelan G, Shunchun W (2014) Characterization of two homogalacturonan pectins with immunomodulatory activity from green tea. Int J Mol Sci 15:9963–9978

Wei Y, Ding M (2000) J Chromatogr A 904:113–117

Wei Y, Hendrix DL, Nieman R (1996) J Agric Food Chem 44:3214–3218

Willats WG, Rasmussen SE, Kristensen T, Mikkelsen JD, Knox JP (2002) Sugar-coated microarrays: a novel slide surface for the high-throughput analysis of glycans. Proteomics 2:1666–1671

Wolf S, Hématy K, Höfte H (2012) Growth control and cell wall signaling in plants. Annu Rev Plant Biol 63:381–407

Wu Y, Thorne ET, Sharp RE, Cosgrove DJ (2001) Modification of expansin transcript levels in the maize primary root at low water potentials. Plant Physiol 126:1471–1479

Xiang L, Li Y, Rolland F, Van Den Ende W (2011) Neutral invertase, hexokinase and mitochondrial ROS homeostasis: emerging links between sugar metabolism, sugar signaling and ascorbate synthesis. Plant Signal Behav 6:1567–1573

Yamada K, Hyodo S, Matsuno YK, Kinoshita M, Maruyama SZ, Osaka YS, Casal E, Lee YC, Kakehi K (2007) Rapid and sensitive analysis of mucin-type glycans using an in-line flow glycan-releasing apparatus. Anal Biochem 371:52–61

Yang Z-L, Liu H-J, Wang X-R, Zeng Q-Y (2013) Molecular evolution and expression divergence of the Populus polygalacturonase supergene family shed light on the evolution of increasingly complex organs in plants. New Phytol 197:1353–1365

Yapo BM (2011) Pectic substances: from simple pectic polysaccharides to complex pectins—a new hypothetical model. Carbohydr Polym 86(2):373–385

Yokoyama R, Nishitani K (2001) A comprehensive expression analysis of all members of a gene family encoding cell-wall enzymes allowed us to predict *cis*-regulatory regions involved in cell-wall construction in specific organs of *Arabidopsis*. Plant Cell Physiol 42:1025–1033

Zaia J (2010) Mass spectrometry and glycomics. Omics 14:401–418

Zenoni S et al (2004) Downregulation of the *Petunia hybrida* α-expansin gene *PhEXP1* reduces the amount of crystalline cellulose in cell walls and leads to phenotypic changes in petal limbs. Plant Cell 16:295–308

Zhang Y, Aurélie G, Michel Z, Benoît V, Elisabeth J, Cécile A (2011) Combining various strategies to increase the coverage of the plant cell wall glycoproteome. Phytochemistry 72:1109–1123

Zhao X, Moates GK, Wellner N, Collins SRA, Coleman MJ, Waldron KW (2014) Chemical characterisation and analysis of the cell wall polysaccharides of duckweed (Lemna minor). Carbohydr Polym 111:410–418

Technological Platforms to Study Plant Lipidomics

Fakiha Afzal, Mehreen Naz, Gohar Ayub, Maria Majeed, Shizza Fatima, Rubia Zain, Sundus Hafeez, Momina Masud, and Alvina Gul

Contents

Abstract The emergence and rapid growth of "omics" has led to a radical change in the viewpoint of life sciences research. Plant metabolomics is a progressing field of study in plant biology where lipidomics is one of the subunits of metabolomics, covering the entire lipidome of the plant body. Previously, mass spectrometry has been used to study plant lipidome and is presently coupled with various chromatographic

F. Afzal • G. Ayub • M. Majeed • S. Fatima • R. Zain • S. Hafeez • M. Masud • A. Gul (✉)
Atta-ur-Rahman School of Applied Biosciences, National University of Sciences & Technology (NUST), Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

M. Naz
Department of Bioinformatics and Biotechnology, International Islamic University (IIU), Islamabad, Pakistan

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_20

techniques to produce more accurate results. Different environmental conditions and stresses contribute to the varying lipids profiling of plants. Numerous environmental stresses trigger lipid-facilitated signaling such as pathogen attack, temperature change, salinity, and drought. N-acylethanolamine, oxylipins, lysophospholipid, phosphatidic acid, inositol phosphate, fatty acid, sphingolipid, diacylglycerol, and N-acylethanolamine have all been suggested to have a signaling role. This chapter reviews various analytical techniques for studying plant lipids. Latest research carried out on lipids variations due to different environmental stresses have also been focused upon in the chapter.

Keywords Metabolomics • Lipidomics • Mass spectrometry • Chromatography

1 Introduction

Lipids play very crucial roles in cells, tissues, organs, and organisms, thus studying lipids, their extensive networks and pathways at cellular level including characterization of the molecular species of lipids is very important to understand complex life buildup and that is all covered in "lipidomics" (Zhaoa et al. 2014). It is quite a new addition to the research made in biological world. Lipidome is a subclass of metabolome (Fig. 1) and describes an organism's entire lipid profile (Raterink et al. 2014). Besides other essential functions in plant body, lipids not only provide stability as membrane part but also serve as signaling and energy storage molecules (Wymann and Schneiter 2008). Numerous environmental stresses trigger lipid-facilitated signaling such as temperature change, salinity, drought, and pathogen attack. N-acylethanolamine, oxylipins, lysophospholipid, phosphatidic acid, inositol phosphate, fatty acid, sphingolipid, diacylglycerol, and N-acylethanolamine have all been suggested to have a signaling role (Okazaki and Saito 2014).

1.1 What Is Plant Lipidomics?

Plant lipidomics is often related to genotype with both spatial and temporal lipid characterization within a plant body. Lipid profiling of plants which have been subjected to reverse or forward genetic manipulation gives a deep insight of manipulated genes and enzymes by comparing it with lipid profile of wild plants (Welti et al. 2002). General analysis of metabolome which is the complement of all metabolites within a cell is very challenging due to heterogeneous chemical nature and complex structure of molecules (Oliver et al. 1998). There is a plenty of research available on polar compounds such as amino acids and sugars

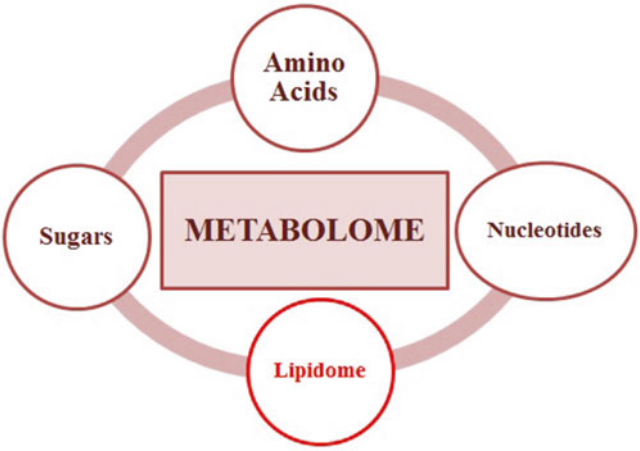


Fig. 1 Subunits of metabolomics

(subclasses of metabolome), but very less data is available on nonpolar molecules. Thus, there is a dire need of latest technologies and techniques to explore these nonpolar water-insoluble molecules. As these molecules are highly abundant, complex, and chemically diverse, there is no sole analytical platform being developed to detect and quantify them altogether in a single run (Oldiges et al. 2007). Resultantly before analysis, advancement of various methods of sample extraction and fractionation allows a rough division of simple and homogeneous metabolite segments (Vuckovic et al. 2010). One of the important components of this crude extract is hydrophobic lipids. Keeping in view the primary importance of lipids, introduction of a novel branch, lipidomics, in the area of metabolomics has greatly improved the knowledge about plant lipids within the past few years (Wenk 2010a, b). The emerging classification of lipids comprises diverse chemical structures with a varying range of physicochemical properties. Because of this diversity, different methods and strategies are being applied to quantification of lipids (Brügger 2014).

The steps involved in the characterization of complex membrane lipids are:

1. First, on the basis of distinct morphology lipids are isolated.
2. Secondly, multistep process is employed to extract lipids from the membranes. Sometimes these lipids are non-covalently attached to non-protein component of membrane. Primary function of these lipids is anchorage and studying them can be an additional aspect of lipidomics.
3. Fractionation of the extracted lipids is done afterwards which requires multistep sophisticated chromatographic techniques. It aids in quantification and fractionation of discrete molecular components (Wolf and Qunin 2008).

2 Classification of Lipids

Not only lipids are integral part of biological membranes but also regarded as latent pool for precursors of signaling second messengers. Studying these types of lipids also comes in lipidomics. The basic property of lipids is their water insolubility but still this property varies in different types of lipids, i.e., some of them are completely water insoluble (highly nonpolar), e.g., triacylglycerol while some are water soluble to some extent (slightly polar) e.g., glycosylated sphingolipids (Merrill et al. 2009; Kuksis 2007; Wenk 2010a, b). Number of lipid species ranges from 100s to 1000s in eukaryotes. Each type is distinct and exhibits its very own complex structure. Therefore, in order to classify them there was a need for a proper nomenclature. Fahy et al. (2005) devised a classification system for lipids named as "LIPID MAPS" which was updated in 2009. A new precise definition was also proposed signifying lipids as hydrophobic or amphipathic small molecules having both nonpolar and polar end rather than the old one based on water insolubility (Fahy et al. 2005, 2009). Thus, according to the new definition, the following subclasses to lipids have been assigned:

1. Polyketides.
2. Sterols.
3. Fatty acids.
4. Sphingolipids.
5. Glycerolipids.
6. Saccharolipids.
7. Glycerophospholipids.
8. Prenols (Fahy et al. 2005).

3 Techniques Involved to Study Lipidomics

Up till now many analytical technologies have come into being, but it impossible to identify and quantify all the metabolites (or lipids) present in a metabolome at the same time in single experimental run (Wenk 2010a, b). Many chromatographic techniques have been employed since now to separate and analyze lipids (Bausch 1993), such as 1D (one dimensional) chromatography, 2D (two dimensional) chromatography, TLC (thin layer chromatography), HPLC (high performance liquid chromatography) with different imaging techniques (Ivanova et al. 2009; Touchstone 1995a, b; Picchioni et al. 1996). Although all these techniques proved to be quite helpful in lipid characterization, they were not ideal for large scale lipid quantification (Blanksby and Mitchell 2010). Therefore, mass spectrometry is the most predominant method recently used along with chromatographic separation techniques but chromatographic technique is not used always (Harkewicz and Dennis 2011).

3.1 Lipid Detection by Mass Spectrometry

Mass spectrometry was discovered by Sir J.J. Thomson in which he used mass-to-charge ratios (m/z) to separate different elements, and now it is frequently used in the fields of proteomics and metabolomics including lipidomics. A large number of developments have been in the field of lipidomics due to MS. Accurate mass detection is quite costly and requires high resolution. There is a range of instruments used in lipidomics for ion detection, but tandem mass spectrometry (MS/MS) has an advantage over others due to higher resolution and accuracy.

Presently two fundamental approaches are used for mass spectrometry-based lipidomics:

1. CLASS: "Comprehensive lipidomics analysis by separation simplification" (CLASS) includes chromatography for separation of different lipids based on LIPID MAPS Consortium prior mass spectrophotometry. Afterwards optimization of mass spectrometer is done to let it implement in a lipid class-specific manner (Nakanishi et al. 2009).
2. Shotgun Lipidomics: Excluding chromatography in this approach, it only relies upon direct infusion of lipids of all classes together into mass spectrometer to separate and analyze them while employing different ion source polarities (to form positive or negative ions) and infusing ionization solution additives, which serve to provide a kind of lipid class-specific favored analysis (Han et al. 2005).

Both approaches have their own benefits and merits, but CLASS is not reported in many studies. Shotgun lipidomics includes triple quadrupole (QqQ) or quadrupole time-of-flight (qTOF) mass spectrometers for crude lipid extracts (Nygren et al. 2011). Basically, MS can be categorized into three different events: analyte ionization, mass-dependent separation of ion, and ion detection (Roberts et al. 2008).

3.1.1 Analyte Ionization

Sample is ionized through bombardment of high energy electrons to give a positive ion. Mass spectrometers always work with positive ions. The primary methods employed within lipidomics for analyte ionization are electron impact (EI) ionization, electrospray ionization (ESI) and matrix assisted laser desorption/ionization (MALDI). Both ESI and MALDI are known as "soft ionization techniques" as EI requires a volatile analyte while ESI and MALDI do not require a volatile analyte (Roberts et al. 2008). Each of them is discussed below.

Electron Impact (EI) Ionization

Electron impact is frequently coupled with gas chromatography as it requires a volatile analyte. First of all, high energy electrons are bombarded to the vapored analyte, thus causing fragmentation of the analyte. It leads to the production of

reproducible arrangement of signals (Downard 2004). These signals are further utilized to recognize different substances by probing spectral databases. The main disadvantage of EI is that it cannot be used to analyze complex mixtures and reliability of interpretation is decreased. Also identification of related or similar species is quite challenging. Therefore, GC is accompanied with EI which has amazing resolving power and reducing the chance of co-elution of different species of analyte (Köfeler et al. 2012). Further roles of GC/MS are explained in detail in this chapter below.

Electrospray Ionization (ESI)

ESI in lipidomics was first proposed by Han and Gross (1994). An extremely charged needle is used to spray the analyte by elution. This is actually done by evaporation of solvent caused by heating charged spray droplets. As a result of this, the analyte is entered into MS unit in the form of ions (Kebarle and Ho 1997). ESI is often coupled with liquid chromatography giving marvelous results. In 1997, static nano electrospray ionization source lacking syringe pump was introduced and proved to be successful for lipid study with much higher ionization than syringe pump (Bruegger et al. 1997). This novel method, though utilized the same basic principle, is currently known as "shotgun lipidomics" and has greatly improved the knowledge about lipids during the last 15 years. Another expansion of shotgun lipidomics is "multidimensional mass spectrometry based shotgun lipidomics (MDMS-SL)." It is also coupled with Nanomate® system, "Multidimensional mass spectrometry based shotgun lipidomics (MDMS-SL)," making it more efficient, automated, fast, and robust and covering various classes of lipids (Han and Gross 2005a, b).

Matrix Assisted Laser Desorption/Ionization (MALDI)

MALDI is also a soft ionization technique requiring co-crystallized molecules along with a soft organic compound known as matrix. This matrix can be of variable types depending upon the sample type, but the most common matrix used in lipid analysis is dihydroxy benzoic acid (DHB) (Petkovic et al. 2001). For desorption and ionization of crystalline sample (analyte), a laser beam is used. MALDI is not as much frequently used as ESI in lipid analysis but is very successful and has widespread applications (Schiller et al. 2007; Knochenmuss 2006).

Advantages and Shortcomings of Soft Ionization Techniques (ESI and MALDI) Over EI

The following are advantages of ESI and MALDI over EI, making them more favored to study lipidomics.

1. Formation of chemical derivatives is necessary for GC/MS, but soft ionization techniques do not require it.
2. Although electron impact (EI) do thorough fragmentation of analyte, still it is very difficult to analyze it later. In contrast, fragmentation of analyte is very nominal in soft ionization techniques but reliable and accurate.

MALDI/ESI can suffer from ion suppression which is the major drawback of these techniques. Ion suppression is generated as a result of competition between analytes for charge during the ionization process. This challenge can be overcome by different methods or by coupling with different chromatographic techniques prior to MS analysis (Knochenmuss 2004; Chico et al. 2012).

3.1.2 Mass-Dependent Separation of Ion

The separation can be accomplished by various types of mass spectrometer such as time-of-flight (TOF), quadrupoles, magnetic and/or electric sectors, or Fourier transform ion cyclotron resonance (Downard 2004).

3.1.3 Ion Detection

The final step is ion detection by a sophisticated ion detector identifying different species.

3.2 Sample Preparation for Mass Spectrometry

Many procedures have been devised by scientists as pretreatment techniques for mass spectrometry-based metabolomics (Raterink et al. 2014). The following are the general steps involved in sample preparation.

1. First cells or tissues are cultured in optimum media.
2. Then probes for the cells are designed accordingly.
3. Sonication is then done in order to release lipids which are stored in cell wall or tissue matrix to produce a uniform homogenate (Ejsing et al. 2009).
4. Afterwards a mixture of internal standards is added to quantitate lipids in the sample. An internal standard can either be deuterium-labeled analogs or odd carbon chain molecular species of similar lipids. Whatever standard is use one should keep in mind that organism under study should not be able to synthesize those (Moore et al. 2007).
5. Crude extract is obtained which is further subjected to either CLASS or shotgun according to desired study. The procedure of mass spectrometry is outlined in Fig. 2.

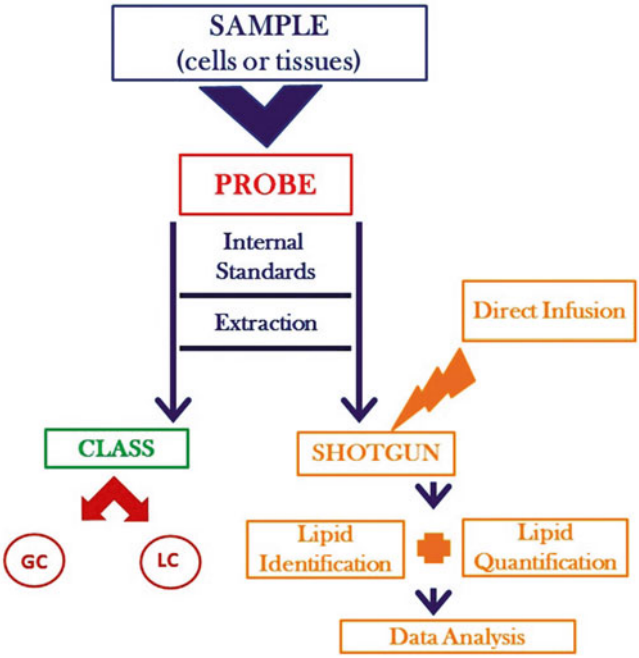


Fig. 2 Major steps for sample preparation of mass spectrometry. Type of mass spectrometer, ionization mode, additives used for ionization and mass spectrometer monitoring mode are variables of this procedures

3.3 Tandem Mass Spectrometry

There are a lot of recent developments made in the use of tandem mass spectrometry (MS/MS) to study the emerging field of lipidomics. It is not only useful for the characterization of lipids by providing detailed information regarding structure but also help in selective determination of lipids residing in a complex mixture. It is also useful in studying novel lipids and their targeted lipidomic analysis (Zehethofer and Pinto 2008). For complex analytes and isomeric lipids, tandem mass spectrometry is coupled with efficient chromatographic techniques such as GC, LC, or HPLC, thus making lipid analysis accurate and reliable.

Tandem mass spectrometry has wide number of applications in plant lipidomics using a triple quadrupole mass spectrometer which is commonly used in the exploration of numerous polar glycerolipids and sphingolipids by exploiting highly specific functions including precursor ion and neutral loss scanning or multiple reaction monitoring according to well-characterized mass fragmentation patterns (Okazaki et al. 2013; Markham and Jaworski 2007). These methodologies provided deep insights into the metabolism and genetics of plant body (Xiao et al. 2010; Peters et al. 2010). According to a recent study it has also been discovered that very long-

chain fatty acids are involved in polar auxin transport and developmental patterning in Arabidopsis (Roudier et al. 2010). Plant synthesize variety of lipids which are yet not classified and characterized by tandem MS. General metabolomics studies cannot be applied to study them hence improved analytical approaches are still required to understand the diversity of plant lipidomics (Okazaki et al. 2013).

3.4 Chromatography Coupled with MS Based Lipidomics

Complex samples which are hard to characterize solely by MS are then analyzed by any chromatographic techniques coupled with the best compatible type of MS. It utilizes the mass resolving and mass determining capabilities of mass spectrometry with highly sensitive and specific nature of chromatography (Kern et al. 2014). The following are the few applications of chromatography coupled to MS.

3.4.1 Thin Layered Chromatography–Mass Spectrometry (TLC/MS)

Thin layered chromatography remained technique of choice for many years but now it is not frequently used. TLC separates lipids on the basis of their class difference (Touchstone 1995a, b). Coupled with MS, the separated lipids are measured using MALDI/MS or ESI/MS. Resolution of TLC is not that much accurate as compared to LC but yet is appropriate for phospholipids (Hu et al. 2009).

3.4.2 Liquid Chromatography–Mass Spectrometry (LC/MS)

In LC, sample is directly infused into ESI source and a lot of advancements have been in lipidomics using LC/MS (Han et al. 2005; Schwudke et al. 2006; Koulman et al. 2007). There are several studies available utilizing LC coupled with analytical procedures. In a recent comparative study by Okazaki et al. 2013, liquid chromatography ion trap time-of-flight mass spectrometry (LC-IT-TOF-MS) was successfully applied to study lipidome of Arabidopsis in a mutant and wild type. It was revealed that there were differences in membrane composition and in lipophilic secondary metabolites when comparison was made in lipidomes of both. Hydrophilic interaction chromatography (HILIC) was also successfully applied for class separation in the same study (Okazaki et al. 2013). The major drawback of LC is ion suppression which not only affects accuracy but also leads to false negative results. Ion suppression can be overcome by HPLC.

Therefore, ESI is sometimes coupled with HPLC/MS for efficient results in lipidomics and to avoid ion suppression. One major advantage of this coupling is that it exploits both the effectual separation power of HPLC and superb selectivity by mass spectrometry. This coupling is done for studying complex lipid classes which cannot be separated by other analytical technique. These lipid classes

include some glycolipids, glycerolipids, and glycerophospholipids. HPLC is much more selective with increased specificity as compared to direct infusion systems. Besides these benefits, quantitation becomes complicated (Köfeler et al. 2012; Whitehouse et al. 1995).

3.4.3 Gas Chromatography–Mass Spectrometry

GC was invented by Martin and Synge in 1941 in which a mobile phase is a gas instead of a liquid (Martin and Synge 1941). Later it was coupled with MS. In the field of lipidomics, GC/MS has proved to be a very efficient technique for separation, identification, and quantification of lipids and has helped to understand lipid metabolism. Sometimes it is accompanied with HPLC or TLC in which separation is done prior GC/MS to deal with complex biological samples (Lehmann et al. 1992).

An important step of GC/MS is chemical derivatization which is done to reduce the unwanted absorption effects. As for GC/MS, analyte is to be thermally stable and volatile, chemical derivatization increases the polarity of polar compounds. In order to aid identification, derivatization is used to remove polar functional groups and is used to generate derivatives. Fatty acid methyl ester (FAME) method is the most common of type of derivatization in which complex lipids are hydrolyzed followed by methylation (Roberts et al. 2008).

3.4.4 Ultraperformance Liquid Chromatography–Mass Spectrometry (UPLC–MS)

Lipidomic research has been greatly facilitated by recent advances in ultraperformance liquid chromatography–mass spectrometry (UPLC–MS) and involved in lipid extraction, lipid identification, and data analysis supporting applications from qualitative and quantitative assessment of multiple lipid species (Zhaoa et al. 2014).

4 Analyses of Complex Plant Lipids Using Different Techniques

Plant metabolism varies in a response to developmental, environmental, and stress-induced physiological changes and it can be studied by high-throughput lipid profiling. Various techniques as discussed above are used to discover functions of lipids in a plant body exposed to different kinds of stresses and further in vivo evaluation of change in the plant lipidome and metabolism caused due to stress. Genes involved in the regulation of lipid metabolism can also be studied. Food quality of plants can also be well studied by analyzing plant lipidome (Welti et al. 2007). The following studies are few examples of latest lipidomics research showing how vital it is.

- Sterols are an abundant lipid class in the extraplastidic membranes of plant cells. In addition to free sterols, plants contain different conjugated sterols, i.e., sterol esters, sterol glucosides, and acylated sterol glucosides. Sterol lipids can be measured by gas chromatography after separation via thin-layer chromatography (Wewer and Dörmann 2014).
- Spruce (*Picea* spp.) and other conifers employ terpenoid-based oleoresin as part of their defense against herbivores and pathogens. This terpenoid-based oleoresin was characterized and quantified using chromatography (LC)-tandem mass spectrometry (MS/MS) method (Nagel et al. 2014).
- A sensitive method based on electrospray ionization tandem mass spectrometry was used to profile glycerolipids in *Pyropia haitanensis* suggesting that agarooligosaccharides caused changes of lipids in chloroplasts and plasma membrane (Wang et al. 2014).
- Changes in the composition of triacylglycerols (TAGs) of three varieties of flaxseeds (H52, O116, and P129) during development were investigated using nonaqueous reversed phase liquid chromatography coupled to atmospheric pressure photoionization-mass spectrometry (LC-APPI-MS), and it was suggested that the most active period of triacylglycerol synthesis was between 7 and 35 days after flowering (Herchi et al. 2012).
- In another recent study, liquid chromatography with quadrupole time-of-flight mass spectrometry was used to perform a metabolomics study of cured tobacco owing to its efficient separation and detection of semi-polar metabolites and developed a method that can be applied to metabolomics analysis of plant materials.

5 Studies on Stress-Induced Alterations in Plant Lipidome

Lipid-mediated signaling occurs in response to various environmental stresses. Membranes are the first things in plant body to face external stresses such as heat, cold, wounding, or phosphorus deficiency. This leads to change in lipid profile of membrane in response to temperature stress. Studying such stress-induced changes in plant lipidome is the primary goal of plant scientist. The following are highlights of latest work done on the analysis of complex plant lipidome alterations in response to various stresses.

- Whenever a plant is exposed to cold (non-freezing temperature) numbers of phospholipids and unsaturation of fatty acids increases to improve membrane fluidity and to decrease the tendency of membrane freezing. It happens in order to maintain the integrity of membrane too (Uemura et al. 1995; Thomashow 1999). It has been reported that profiling methods are been employed to plants for the identification of Cold-Regulated Primary Metabolites Using Gas Chromatography Coupled to Mass Spectrometry (Dethloff et al. 2014). Several studies on different plants were done in order to analyze these lipid changes such as Welti et al. in 2002 performed

experiments to show precise lipid changes in Arabidopsis when exposed to cold stress. Lipid profiling showed that there was an increase in those species which contained two polyunsaturated acyl groups. Those species were:

- Digalactosyldiacylglycerol (GalGalDG).
- Phosphatidylglycerol (GPGro)
- Monogalactosyldiacylglycerol (GalDG)
- Phosphatidylethanolamine (GPEtn)

Simultaneously saturation at the specific points in the following lipids decreases:

- Phosphatidylcholine (GPCho, 18:1/18:2 and 18:0/18:3)
- 36:2—Stearoyl (18:0)/18:2
- Di-oleoyl (18:1)

Contrariwise, phosphatidylinositol (GPIns) contents remained unchanged when subjected to cold stress. It can be deduced from the above findings that during cold assimilation production of lysophospholipids and GPA is increased by phospholipases (PLA and PLD). Lysophospholipids and GPA are signaling molecules and they play a vital role in regulation of cytoskeletal organization and functions of ion channels. Hence both regulatory and structural roles of lipids during acclimatization and endurance to cold stress are proved (Welti et al. 2002; Wang et al. 2006). Recently Vu et al. (2014) have carried out lipidomic analysis using electrospray ionization triple quadrupole mass spectrometry was employed to monitor lipid changes that occur during cold and freezing stress in Arabidopsis and found obvious lipid remodeling.

- Most vulnerable organelle's membrane to temperature stress is of chloroplast which is a player to photosynthesis (Weis and Berry 1988). Burke et al. in 2000 studied thermo tolerance in plants by employing forward genetics approach. As a result of this study they were able to isolate and characterize a wide range of mutants (atts) of *Arabidopsis thaliana* that die in short-term exposure high temperature (38.5 °C). In contrast to the mutant ones, wild plant can build thermo tolerance at this temperature and survive even at higher temperature (Burke et al. 2000). It was later confirmed that loss of acquired thermo tolerance was due to the conformational change in DGD1 due to single amino acid substitution. DGD1 is responsible for developing thermotolerance in plant and damaging its integrity resulted in the loss of its function (Chen et al. 2006).
- Oxylipins are polar lipids present in the plastids of *Arabidopsis thaliana*. Pathways involved in the biosynthesis of oxylipins are not known yet (Hisamatsu et al. 2005). Many analytical techniques have been combined to study the role of oxylipins in mechanical stress such as wounding. ESI/TOF/MS was used to identify the fatty acyl chains and 17 oxylipins were found in the wounded plants including some polar species (Buseman et al. 2006).
- Phosphorous is a vital macronutrient and is essential for proper growth and development of plants. Plant growth is stunted in limited phosphorous and along with this galactolipids typically in the form of GalGalDG starts to accumulate in plant body and lower levels of phospholipids were found (Kusano et al. 2014; Li et al. 2006).

6 Conclusion and Future Prospective

Lipids unlike other biological components of plant body (such as proteins, sugars, or nucleic acids) are not made up of small building blocks and are not genetically encoded but certain enzymes produce and metabolize them as a result of catabolism or anabolism under strict physiological control thus making it difficult to classify them. Lipids (glycerophospholipids, sterols, and sphingolipids) are part of biological membranes and thus are responsible to maintain integrity of cell. They also act as signaling molecules. To study major lipids in complex samples (such as tissues and cell extracts) along with several biological species (including yeast and mammals), mass spectrometry based methods are now available for both qualitative and quantitative analysis. Latest technologies including nuclear magnetic resonance (NMR) and X-crystallography give more accurate structural information. Forthcoming promises include technical improvements stemming from cell isolation, sample fractionation and preparation, standardization and cross-validation and automation. Additionally, a broader coverage of biochemical lipidomics from integration with imaging, databases and further addition of biological species is part of the package that the near future will provide. Similarly, interdisciplinary programs can be anticipated to continue integrating biochemical lipidomics with chemical biology, proteomics, and genomics to span the entire information flow encoded in biological systems. These efforts will contribute to a better understanding of natural variation in lipids and will most optimally lead to tailored applications in life sciences, industrial settings, and medicine.

References

Bausch JN (1993) Lipid analysis. Curr Opin Biotechnol 4:57–62

Blanksby SJ, Mitchell TW (2010) Advances in mass spectrometry for lipidomics. Annu Rev Anal Chem 3:433–465

Bruegger B, Erben G, Sandhoff R, Wieland FT, Lehmann WD (1997) Quantitative analysis of biological membrane lipids at the low pico mole level by nano-electrospray ionization tandem mass spectrometry. Proc Natl Acad Sci U S A 94:2339–2344

Brügger B (2014) Lipidomics: analysis of the lipid composition of cells and subcellular organelles by electrospray ionization mass spectrometry. Annu Rev Biochem 83:79–98

Burke JJ, O'Mahony PJ, Oliver MJ (2000) Isolation of *Arabidopsis* mutants lacking components of acquired thermotolerance. Plant Physiol 123:575–588

Buseman CM, Tamura P, Sparks AA, Baughman EJ, Maatta S, Zhao J, Roth MR, Esch SW, Shah J, Williams TD, Welti R (2006) Wounding stimulates the accumulation of glycerolipids containing oxophytodienoic acid and dinor-oxophytodienoic acid in Arabidopsis leaves. Plant Physiol 142:28–39

Chen J, Burke JJ, Xin Z, Xu C, Velten J (2006) Characterization of the Arabidopsis thermosensitive mutant atts02 reveals an important role for galactolipids in thermotolerance. Plant Cell Enviorn 29:1437–1448

Chico J, Holthoon FV, Zuidema T (2012) Ion suppression study for tetracyclines in feed. Chromatogr Res Int 2012:Article ID 135854. doi:10.1155/2012/135854

Dethloff F, Erban A, Orf I, Alpers J, Fehrle I, Beine-Golovchuk O, Schmidt S, Schwachtje J (2014)
Downard K (2004) Mass spectrometry: a foundation course. Royal Society of Chemistry, Cambridge
Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K (2009) Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. Proc Natl Acad Sci U S A 106:2136–2141
Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AHJ, Murphy RC, Raetz CR, Russell DW, Seyama Y, Shaw W, Shimizu T, Spener F, VanMeer G, Vannieuwenhze MS, White SH, Witztum JL, Dennis EA (2005) A comprehensive classification system for lipids. J Lipid Res 46:839–861
Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, Spener F, VanMeer G, Wakelam MJ, Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50:9–14
Han XL, Gross RW (1994) Electrospray ionization mass spectroscopic analysis of human erythrocyte plasma membrane phospholipids. Proc Natl Acad Sci U S A 91:10635–10639
Han X, Gross RW (2005a) Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. Mass Spectrom Rev 24:367–412
Han XL, Gross RW (2005b) Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes. Expert Rev Proteomics 2:253–264
Han X, Yang J, Cheng H, Yang K, Abendschein DR, Gross RW (2005) Shotgun lipidomics identifies cardiolipin depletion in diabetic myocardium linking altered substrate utilization with mitochondrial dysfunction. Biochemistry 44(50):16684–16694
Harkewicz R, Dennis EA (2011) Applications of mass spectrometry to lipids and membranes. Annu Rev Biochem 80:301–325
Herchi W, Trabelsi H, Salah HB, Zhao YY, Boukhchina S, Kallel H, Curtis JM (2012) Changes in the triacylglycerol content of flaxseeds during development using liquid chromatography-atmospheric pressure photoionization-mass spectrometry (LC-APPI-MS). Afr J Biotechnol 11(4):904–911
Hisamatsu Y, Goto N, Sekiguchi M, Hasegawa K, Shigemori H (2005) Oxylipins arabidopsides C and D from Arabidopsis thaliana. J Nat Prod 68:600–603
Hu C, Heijden RVD, Wang M, Greef JVD, Hankemeier T, Xu G (2009) Analytical strategies in lipidomics and applications in disease biomarker discovery. J Chromatogr B 877:2836–2846
Ivanova PT, Milne SB, Myers DS, Brown HA (2009) Lipidomics: a mass spectrometry based systems level analysis of cellular lipids. Curr Opin Chem Biol 13:526–531
Kebarle FP, Ho Y (1997) Electrospray ionization mass spectrometry. In: Cole RB (ed) Fundamentals instrumentation and applications. John Wiley & Sons, New York, NY
Kern W, Mende R, Denefeld B, Sackewitz M, Chelius D (2014) Ion-pair reversed-phase high performance liquid chromatography method for the quantification of isoaspartic acid in a monoclonal antibody. J Chromatogr B 995(26-33)
Knochenmuss R (2004) Photoionization pathways and free electrons in UV-MALDI. Anal Chem 76(11):3179–3184
Knochenmuss R (2006) Ion formation mechanisms in UV-MALDI. Analyst 131(9):966–986
Köfeler HC, Fauland A, Rechberger GN, Trötzmüller M (2012) Mass spectrometry based lipidomics: an overview of technological platforms. Metabolites 2:19–38
Koulman A, Tapper BA, Fraser K, Cao M, Lane GA, Rasmussen S (2007) High-throughput direct-infusion ion trap mass spectrometry: a new method for metabolomics. Rapid Commun Mass Spectrom 21(3):421–428
Kuksis A (2007) Lipidomics in triacylglycerol and cholesteryl ester oxidation. Front Biosci 12:3203–3246
Kusano M, Yang Z, Okazaki Y, Nakabayashi R, Fukushima A, Saito K (2014) Using metabolomics approaches to explore chemical diversity in rice. Mol Plant pii:ssu125. doi:10.1093/mp/ssu125
Lehmann WD, Stephan M, Fürstenberger G (1992) Profiling assay for lipoxygenase products of linoleic and arachidonic acid by gas chromatography-mass spectrometry. Anal Biochem 204(1):158–170

Li M, Welti R, Wang X (2006) Quantitative profiling of Arabidopsis polar glycerolipids in response to phosphorus starvation: Roles of PLDzeta1 and PLDzeta2 in phosphatidylcholine hydrolysis and digalactosyldiacylglycerol accumulation in phosphorus-starved plants. Plant Physiol. doi:10.1104/pp.106.085647

Markham JE, Jaworski JG (2007) Rapid measurement of sphingolipids from Arabidopsis thaliana by reversed-phase high performance liquid chromatography coupled to electrospray ionization tandem mass spectrometry. Rapid Commun Mass Spectrom 21(7):1304–1314

Martin AJP, Synge RLM (1941) A new form of chromatogram employing two liquid phases. Biochem J 35:1358–1368

Merrill AH, Stokes J, Momin TH, Park A, Portz H, Kelly BJ, Wang S, Sullards EMC, WangM D (2009) Sphingolipidomics: a valuable tool for understanding the roles of sphingolipids in biology and disease. J Lipid Res 50:97–102

Moore JD, Caufield WV, Shaw WA (2007) Quantitation and standardization of lipid internal standards for mass spectrometry. Methods Enzymol 432:351–367

Nagel R, Berasategui A, Paetz C, Gershenzon J, Schmidt A (2014) Overexpression of an isoprenyl diphosphate synthase in spruce leads to unexpected terpene diversion products that function in plant defense. Plant Physiol 164(2):555–569

Nakanishi H, Ogiso H, Taguchi R (2009) Qualitative and quantitative analyses of phospholipids by LC-MS for lipidomics. Methods Mol Biol 579:287–313

Nygren H, Seppanen-Laakso T, Castillo S, Hyotylainen T, Oresic M (2011) Liquid chromatography mass spectrometry (LC-MS) based lipidomics for studies of body fluids and tissues. Methods Mol Biol 708:247–257

Okazaki Y, Saito K (2014) Roles of lipids as signaling molecules and mitigators during stress response in plants. Plant J 79(4):584–596

Okazaki Y, Kamide Y, Hirai MY, Saito K (2013) Plant lipidomics based on hydrophilic interaction chromatography coupled to ion trap time-of-flight mass spectrometry. Metabolomics 9:121–131

Oldiges M, Lutz S, Pflug S, Schroer K, Stein N, Wiendahl C (2007) Metabolomics: current state and evolving methodologies and tools. Appl Microbiol Biotechnol 76:495–511

Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16:373–378

Peters C, Li M, Narasimhan R, Roth M, Welti R, Wang X (2010) Nonspecific phospholipase C NPC4 promotes responses to abscisic acid and tolerance to hyperosmotic stress in Arabidopsis. Plant Cell 22(8):2642–2659

Petkovic M, Schiller J, Müller M, Benard S, Reichl S, Arnold K, Arnhold J (2001) Detection of individual phospholipids in lipid mixtures by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: phosphatidylcholine prevents the detection of further species. Anal Biochem 289(2):202–216

Picchioni GA, Watada AE, Whitaker BD (1996) Quantitative high performance liquid chromatography analysis of plant phospholipids and glycolipids using light scattering detection. Lipids 31:217–221

Raterink RJ, Lindenburg PW, Vreeken RJ, Ramautar R, Hankemeier T (2014) Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. Trends Anal Chem 61:157–167

Roberts LD, McCombie G, Titman CM, Griffin JL (2008) A matter of fat: An introduction to lipidomic profiling methods. J Chromatogr 871:174–181

Roudier F, Gissot L, Beaudoin F, Haslam R, Michaelson L, Marion J (2010) Very-long-chain fatty acids are involved in polar auxin transport and developmental patterning in Arabidopsis. Plant Cell 22(2):364–375

Schiller J, Suss R, Fuchs B, Muller M, Zschornig O, Arnold K (2007) MALDI-TOF MS in lipidomics. Front Biosci 1(12):256825–256879

Schwudke D, Oegema J, Burton L, Entchev E, Hannich JT, Ejsing CS, Kurzchalia T, Shevchenko A (2006) Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. Anal Chem 78(2):585–595

Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. Annu Rev Plant Physiol Plant Mol Biol 50:571–599
Touchstone JC (1995a) Thin layer chromatographic procedures for lipid separation. J Chromatogr B Biomed Appl 671:169–195
Touchstone JC (1995b) Thin-layer chromatographic procedures for lipid separation. J Chromatogr 671:169
Uemura M, Joseph RA, Steponkus PL (1995) Effect of cold acclimation on the lipid composition of the inner and outer membrane of the chloroplast envelope isolated from rye leaves. Plant Physiol 109:15–30
Vu HS, Shiva S, Hall AS, Welti R (2014) A lipidomic approach to identify cold-induced changes in Arabidopsis membrane lipid composition. Methods Mol Biol 1166:199–215
Vuckovic D, Zhang X, Cudjoe E, Pawliszyn J (2010) Solid-phase micro extraction in bioanalysis: new devices and directions. J Chromatogr A 1217:4041–4060
Wang X, Devaiah SP, Zhang W, Welti R (2006) Signaling functions of phosphatidic acid. Prog Lipid Res 45:250–278
Wang X, Su X, Luo Q, Xu J, Chen J, Yan X, Chen H (2014) Profiles of glycerolipids in *Pyropia haitanensis* and their changes responding to agaro-oligosaccharides. J Appl Phycol 26(6):2397–2404
Weis E, Berry JA (1988) Plants and high temperature stress. Symp Soc Exp Biol 42:329–346
Welti R, Li W, Li M, Sang Y, Biesiada H, Zhou HE, Rajashekar CB, Williams TD, Wang X (2002) Profiling membrane lipids in plant stress responses. Role of phospholipase D-alpha in freezing-induced lipid changes in Arabidopsis. J Biol Chem 277:31994–32002
Welti R, Shah J, Li W, Li M, Chen J, Burke JJ, Fauconnier ML, Chapman K, Chye ML, Wang X (2007) Plant lipidomics: discerning biological function by profiling plant complex lipids using mass spectrometry. Front Biosci 12:2494–2506
Wenk MR (2010a) Lipidomics: new tools and applications. Cell 143:888–895
Wenk MR (2010b) Lipidomics: new tools and applications. Cell 142(6):888–895
Wewer V, Dörmann P (2014) Determination of sterol lipids in plant tissues by gas chromatography and q-tof mass spectrometry. Methods Mol Biol 1153:115–133
Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB (1995) Electrospray interface for liquid chromatographs and mass spectrometers. Anal Chem 57:675–679
Wolf C, Qunin PJ (2008) Lipidomics: practical aspects and applications. Prog Lipid Res 47:15–36
Wymann MP, Schneiter R (2008) Lipid signalling in disease. Nat Rev Mol Cell Biol 9:162–176
Xiao S, Gao W, Chen QF, Chan SW, Zheng SX, Ma J (2010) Overexpression of Arabidopsis acyl-CoA binding protein ACBP3 promotes starvation-induced and age-dependent leaf senescence. Plant Cell 22(5):1463–1482
Zehethofer N, Pinto DM (2008) Recent developments in tandem mass spectrometry for lipidomic analysis. Anal Chim Acta 627:62–70
Zhaoa YY, Wuc SP, Liub S, Zhang Y, Lind RC (2014) Ultra-performance liquid chromatography–mass spectrometry as a sensitive and powerful technology in lipidomic applications. Chem Biol Interact 220:181–192

Plant Interactomics Under Salt and Drought Stress

Atif Shafique, Zeeshan Ali, Abdul Mohaimen Talha, Muneeb Haider Aftab, Alvina Gul, and Khalid Rehman Hakeem

Contents

A. Shafique • Z. Ali • A.M. Talha • M.H. Aftab • A. Gul (✉)
Atta-Ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad, Pakistan
e-mail: alvina_gul@yahoo.com

K.R. Hakeem (✉)
Faculty of Forestry, Universiti Putra Malaysia, Serdang-43400, Selangor, Malaysia
e-mail: kur.hakeem@gmail.com

K.R. Hakeem et al. (eds.), *Plant Omics: Trends and Applications*,
DOI 10.1007/978-3-319-31703-8_21

Abstract Different abiotic stresses are responsible for low yield in crops. The same apparently goes for wheat; it is affected by both biotic and abiotic stresses. Among abiotic stresses, salinity and drought are the top ones. Plants have evolved elaborate mechanisms and ways to counter these threats. Studies in Arabidopsis have uncovered genes responsible for combating such stresses. For example, LEA protein gene is expressed under these conditions. Moreover, ERF plays the same role as LEA. Besides this different transcription factors are involved in stress tolerance which bind to different elements like C-repeat/DRE (CRT) and cause transcription of these stress tolerance genes under stresses like drought. A number of signaling transduction pathways are responsible for regulation of plant responses in stress. It is possible that TaCRT might be involved in direct or indirect activation of specific signal transduction pathways, which results in an enhanced metabolism. This enhanced metabolism enables cells to protect them from drought stress injury. Salt overly sensitive protein family (SOS1, SOS2, SOS3, and SOS4) is involved in salt tolerance. SOS1 gene is a genetic arrangement for salt tolerance because of Na^+/H^+ antiporter which allows plants to reproduce and multiply under salt stress. In addition to the SOS1 gene, pyrroline carboxylate synthetase (P5cS) and betaine aldehyde dehydrogenase (BADH) genes are also expressed under salt stress. A number of genes are involved in circulation of Na^+ under stress, like TaHKT1;5 in Mahuti wheat.

Keywords CRT • Dehydrin • ERF • TaCor410 • Abiotic stress • LEA • DBFS • TaAIDFs • TaCRT • PEG • PEBV • GFP WRC • MDR • WUE • Calreticulin • ABA • DRE • ABRE • rd29A • AtSOS • Calcineurin • Na^+/H^+ antiporter • TaHKT1;5 • P5cS

1 Introduction

Interactomics is the science of physical interactions between proteins and other cellular entities. These interactions can be protein–protein interactions, between same or different types of proteins, or they can be protein–DNA interactomics. A network of such interconnections is called an interactome. Proteins are involved in most cellular functions including, cell division, motility, apoptosis, differentiation, energy production, immunity, and many more. In most cases, proteins group up to form a functional complex. The amount and variety of protein–protein interactions that occur in a biological system are so complex that it can only be represented as a complicated graph. This graph functions as a map which shows the possible paths a protein may take and the molecules it may encounter. This set of interactions a cell may possess is called the interactome of that cell.

Being at the functional end of the "central dogma" of biology, protein–protein interactions are at the center of almost every activity which takes place in a cell, therefore playing a role in gene function directly, ranging from as basic and complex a function as signal transduction to higher functions like plant defense and organ formation (Tong et al. 2001). Molecular level responsibilities of protein–

protein interactions include roles in posttranslational modifications, cytoskeleton assembly, and activation and deactivation of transporter molecules. All these functionally important roles make these interactions an essential part of processes at large which include defense mechanisms, physiological behaviors, and also developmental processes (Auerbach et al. 2002).

Carrying out signal transductions is one of the major roles of protein–protein interactions as most of the molecules involved in signaling processes are protein in nature. Regulation of communication is also a part of the said function so as to orchestrate the transductions in such a manner that the result is the targeted function—not only the signaling between the molecules of importance but also the dynamics with which these interactions take place. Kinases and phosphatases are general components of the signal transduction machinery. Larger members of these protein families are employed by higher plants to cater to their needs due to complicated defense and stress related pathways. Rice protein kinase models have been used to elucidate the function of kinases (Broder et al. 1998). The models provided an insight into the molecular and biological functions played by these kinases in the defense mechanisms of rice plant (Kerppola 2008).

Systems biology approaches are being used today to reveal all the information that can be extracted from any experimental data. With this very aim, global gene expression and quantitative trait loci were integrated with protein–protein interaction network to investigate the response to abiotic and biotic stresses in rice. The study resulted in the discovery of disease resistance genes and their homologs in both monocots and dicots with functional similarities. Thus, this gives the power to predict gene functions in a broader range (Forler et al. 2003).

Van Leene et al. (2008) investigated *Arabidopsis* cell cycle proteins and their interactions using a high-throughput AP-MS. Out of the 42 interactions that were found, 28 were interactions that had not been seen before. The results also mapped core proteins involved in cell cycle and their important regulatory mechanisms (Lehner et al. 2006). Maturation of RNA and the turnover number involves polyadenylation of RNA. Interaction networks developed to study the machinery involved in the mRNA polyadenylation revealed many new genes involved in the process. In addition to the networks, the interaction mode was validated in wet lab using Y2H and AP-MS. Gene expression profiling of the networks uncovered subcomplexes that may be involved in tissue-specific and developmental stage-specific RNA processing (Dortay et al. 2008).

Protein interactomics is at the crossroads of a number of disciplines, namely genomics, bioinformatics, computational modeling, and cell and molecular biology. Being somewhat of an interdisciplinary science, it uses approaches that are a combination of all these areas of study, and an elaborate discussion of these approaches used is beyond the scope of this work. However, briefly, these approaches can be divided into two sets, namely bioinformatics and experimental, or in other words, dry lab and wet lab approaches (Ivanov et al. 2011).

The experimental data already collected forms the basis on which the bioinformatics approaches mainly work. Using genomic information, evolutionary relationships, protein structure and domains, predictions are done about how might a certain protein interact with other proteins based on the available data. These computational models help finding new possible interactions and how they affect other processes in the cell.

Experimental approaches involve real-life wet lab experiments which help validate the computational models as well as collect new data for new and improved models to be predicted and designed. This has caused an increase in wet lab experiments over the bioinformatics approaches, but in no way it can be assumed that the computational approaches have lost their significance. Unless both approaches are employed side by side, a quick progress will be impossible (Ivanov et al. 2011).

The name molecular fishing has been appropriately given to studies involving protein–protein interactions which employ "bait" and "prey" proteins. The experimenter chooses a bait protein and analyzes which other proteins he can "catch" with this protein. The caught proteins are the prey proteins. This reveals as to with which other proteins the bait protein can and has interactions in its environment. These molecular fishing approaches are categorized into two based on (1) genomic information and (2) biochemical identification (Serebriiskii et al. 2001).

2 Gene–Protein Interatomics Under Abiotic Stress

Different abiotic stresses like salinity and drought are the most important hurdles in plant growth and development, which lead to low yield (Wang et al. 2003). Stresses such as abiotic stresses can cause annihilation of proteins and its products in the cell (Sachs et al. 1980). Drought is the major challenge which wreaks havoc in the agronomic world (Chaves et al. 2003; Somerville and Briscoe 2001). The tension created by low water availability can devastate a crop (Ramanjulu and Bartels 2002).

Plants have found a loophole in all these scenarios as they tend to make certain modifications in cells or physiology. This enables them to lose less water and thus save them from devastation. Moreover, an elaborate mechanism is altering the expression few or more important genes. By this they can resist the onslaught by these abiotic stresses. There is a common way by which we identify and characterize certain genes that express themselves under these conditions. The advances in genomics have made the way to such tactics easy.

Through this we can now understand the underlying mechanisms of such genes and see their expression in plants under drought stress (Yamaguchi-Shinozaki and Shinozaki 2000; Bray 1997; Ingram and Bartels 1996). Studies on Arabidopsis have uncovered many genes that are involved in stress tolerance, especially in drought stress (Yamaguchi-Shinozaki and Shinozaki 2000; Shinozaki et al. 2003).

3 Dehydrins and Related Proteins

There are a group of certain proteins that may be involved in stress tolerance. It includes LEA or late embryogenesis-abundant proteins, heat shock proteins, and also osmoprotectants. Moreover, there are others like different biological pathways related proteins, synthesis proteins, enzymes responsible for detoxification, antifreeze proteins, inhibitors, and senescence-related proteins. An important area is characterization and identification of LEA proteins (Ingram and Bartels 1996).

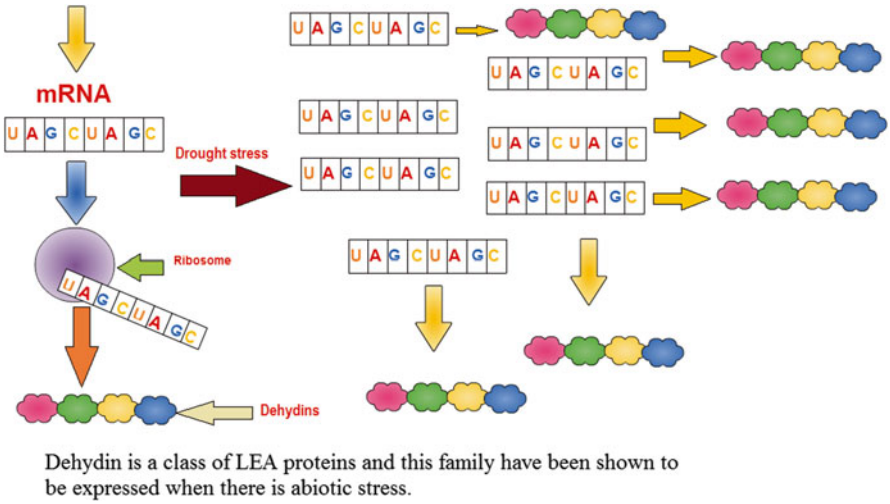


Fig. 1 Expression of dehydrin increases as plant experiences drought

LEA proteins were first identified in plant seeds (Dure et al. 1981), based on similarity of sequences and conserved domains. At present these proteins have been categorized into nine groups (Bies-Etheve et al. 2008; Wise 2003). Among these groups the most investigated is dehydrins (Yuxiu et al. 2007; Veeranagamallaiah et al. 2011). These are also produced during the late stages of embryogenesis when the water content of the seed decreases and when the seed is in maturation phase (Mtwisha et al. 1998). These proteins possess hydrophilic properties, and therefore, they tend to associate with water molecules, which stabilizes membranes and other macromolecules (Wang et al. 2003).

Some research indicates that these classes of proteins are involved in saving cellular molecules and other vital structures from being destroyed. They do so by sequestering ions and replacing hydrogen bonding in proteins, which then tends to change structure and provide resistance (Bray 2002). Some studies show that dehydrins are linked with macromolecules like nucleoproteins which reside in the nucleus and also with endomembranes located in the cytoplasm (Schneider et al. 1993). They are also associated with the plasma membrane of the cell (Fig. 1).

4 ERF, Dehydrin, and TaCor410

The ERF proteins are recognized to associate with just GCC-box. But two of these ERF have been found to bind with CRT and GCC box; these are from pepper and wheat plants (Xu et al. 2007; Yi et al. 2004). On the other hand dehydrins are a class of LEA proteins, and this family has been shown to be expressed when there is abiotic stress (Close et al. 1989). Normally they are expressed in mature embryos and endosperm, but the expression increases rapidly as the plant experiences stresses like drought, cold, and salinity (Mundy et al. 1990) (Fig. 2).

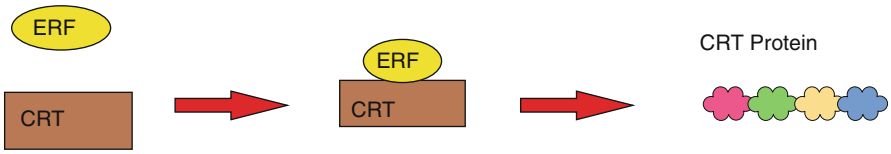


Fig. 2 ERF when binds to CRT creates CRT protein

The Cor410 gene which was first identified as a LEA protein is shown to accumulate in the roots, leaves, and crown tissues of freeze-tolerant plants (Danyluk et al. 1994), and it is now demonstrated that expression levels of this gene is high in wheat plants under cold stress. It has been shown that this protein accumulated near the plasma membrane, the area that experiences dehydration the most and hence the high activity of the protein (Danyluk et al. 1989). Now a finding suggests that the TaCor410 protein is involved in dehydration stress which can be important in drought conditions as well (Houde et al. 2004).

5 Transcription Factors, DBFS, and TaAIDFs Interactions with CRT/DRE

Transcription factors play a vital role in stress conditions; they are involved in direct regulations or indirect, altering gene expression at different levels (Bray 2004). DREB/CBFs are well-known transcription factors that are expressed and function very well under stress like cold and drought (Thomashow 2001; Stockinger et al. 1997). At present they are the most studied transcription factors in abiotic stress reaction. They activate certain genes that are responsible to control metabolism and osmoprotection (Lata and Prasad 2011; Hussain et al. 2011).

The proteins created under such stresses bind to C-repeat/DRE (CRT), leading to transcription of desired genes (Stockinger et al. 1997), in an ABA-independent way. It has been seen that these factors counter stress (Agarwal and Jha 2010). The DREB proteins which are also known as CRT or C repeat binding factors help restrict the effect of drought and cold; they do so by binding to the CRT elements (Maruyama et al. 2004; Sakuma et al. 2006).

6 What Is CRT/DRE?

CRT is a protein that is expressed in all eukaryotes that have been subjected to study (Michalak et al. 1999; Coppolino and Dedhar 1998). It is an abundant calcium-binding protein. It was first discovered in rabbit skeletal muscle endoplasmic

reticulum (Ostwald and MacLennan 1974; Opas et al. 1996). Later a cDNA clone of the protein was isolated and sequenced (Fliegel et al. 1989); besides this it was also found in nuclear envelope and in the spindle machinery of cells undergoing cell division (Denecke et al. 1995). It was also found on the cell surface and the plasmodesmata (Gardai et al. 2005), it is expected that CRT is involved in normal functions of the cell (Chen et al. 2005; Laporte et al. 2003).

A comprehensive investigation of mammalian CRT has showed many important functions which include Ca^{2+} regulation and its dependent pathways. Moreover, its properties enable it to chaperone and fold many proteins. It is also involved in receptor-mediated expression of genes; it regulates cell adhesion, but most importantly it plays a crucial role in the immune system of the cell and apoptosis (Krause and Michalak 1997; Denecke et al. 1995; Coppolino et al. 1997; Gelebart et al. 2005; Guo et al. 2002) (Fig. 3).

Wheat is a very complex and a central food crop in the world. The main challenge to its ultimate yield is abiotic and biotic stresses. And drought is the stress that mainly decreases the yield of this important crop. To improve the yield of this crop we need to understand the molecular bases and mechanisms behind all these stresses and tolerance in plants if any. A CRT-like sequence was reported (GI 56606826) in wheat (Xia-yun et al. 2008).

7 Cor410, Dehydrin, and CRT

Expression of gene Cor410b which is a dehydrin gene is increased when there is drought and other stresses like wounding and cold. By using deletion analysis it was found out that its promoter binds with seven transcription factors. It was detected by utilizing a yeast hybrid system of barley and wheat. The cDNA libraries were of

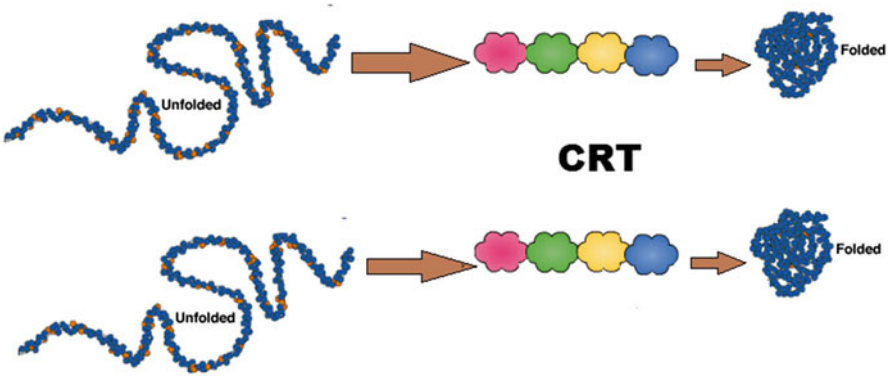


Fig. 3 CRT helping an unfolded protein transform into a folded one

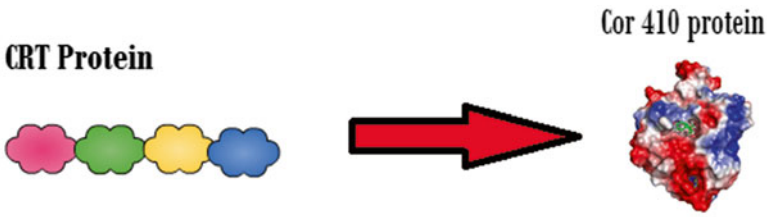


Fig. 4 An example of synthesis of CR410, a protein which acts against drought

DREB class. While in case of rest of the six transcription factors which beong to ERF family, they further belonged to three subfamilies (Fig. 4).

Further analysis showed that these factors bind to CRT elements (GCCGAC). Out of these seven, three could bind to the sequence of ethylene-responsive box called as GCC box with the sequence of (GCCGCC). They can bind to both ethylene and CRT elements. TaERF4 members did not attach to either DRE elements or GCC box. It was suggested that both ERF and DREB can regulate the expression of Cor410b, and this can be done with the help of a single element called CRT. The subfamilies of TaERF4 are positive regulators of the gene explained above (Brown et al. 2003; Guo and Ecker 2004; Shinshi 1995; Park et al. 2001; Schumacher et al. 2002).

8 DBFs and TaAIDFs Interactions with CRT

DBFs are constituents of AP2/ERF family; this is a superfamily and has a vital role in tolerance of abiotic stress conditions in plants. Xu et al. (2008) isolated three homologs of these families in wheat plant. This was done by screening a drought-resistant c-DNA library. It was named TaAIDFs. It was found out that the N terminal region of TaAIDFs controls nuclear localization. This protein has the ability to bind to CRT elements in the cell; both in vivo and in vitro results show its binding properties. The promoter triggers its activation under different stress conditions (Xu et al. 2008).

The study was further confirmed when the responses were checked against different abiotic stresses like drought and salt. Moreover, the activation was also seen under low temperature conditions. TaAIDFa started to activate CRT related genes and this improved resistance under stress and also osmotic tolerance in *A. thaliana*. It was suggested that this TaAIDFa might code a DRE/CRT element that binds with it, and it must be involved in other multiple stress-evading pathways generated by plants (Xu et al. 2008).

9 TaCRT Gene Analysis in Wheat Drought Stress

TaCRT gene was examined by northern blot. No expression was reported in drought-stressed wheat seedlings, and PEG stress was reported to induce TaCRT gene when expression levels of TaCRT gene was examined under varying PEG concentrations. It was observed that under PEG-induced stress the level of TaCRT gene expression was greatly enhanced. This indicated that cells elevate the concentration of their transcription elements under drought stress (Xia-yun et al. 2008).

10 Variation in Crt1 Expression Under Stress

Varying expression of isoforms of CRT1 gene was also observed in *Arabidopsis thaliana* under external stimuli. CRT1 and CRT2 showed a relatively slow induction of gene expression under salinity or tunicamycin. The other isoform CRT3 showed a rapid response to drought stress. CRT2 was found to be associated with both tunicamycin and dithiothreitol in Arabidopsis. Although the expression of CRT varies with time and space, it can be postulated that expression of CRT gene increases under abiotic stresses like salinity and drought that leads to enhanced expression of TaCRT. This CRT upregulation may be considered as a protection mechanism that plants acquire to tolerate unfavorable osmotic conditions (Persson et al. 2003).

11 Expression Profile of TaCRT

In order to investigate the expression profile of TaCRT in vitro, tobacco plant was induced to overexpress the TaCRT. For this purpose, plant expression vectors of virus origin were used. PEBV is one of the commonly used viral plant expression vectors. Virus based plant vectors are preferred in the analysis studies because of their convenient engineering. These vectors do not require stable transformation and also have short interval of time between phenotypic analysis and cloning (Lindboet et al. 2001).

Expression analysis was also done with PEBV. They were used because of their good efficiency in tomato and *N. benthamiana*. In *N. benthamiana*, PEBV is also considered as an expression vector with GFP as a reporter gene. *N. benthamiana* being one of the highly considered plant hosts in developmental and genetic studies was used in this experimental genetic and developmental analysis. The plants were transformed via Agrobacterium method. For further confirmation of transformation, RT-PCR and PCR and western blot were used. RWC, MDR, WRA, and WUE were used as physiological indicators for drought tolerance evaluation and resistance in crops. Higher WRA, WUE, RWC and lower MDR in plants show greater tolerance and resistance to drought stress.

12 Overexpression of TaCRT

When plants were transformed with overexpressed TaCRT, no change in seed production and physical appearance and time of flowering was observed as compared to control and wild type plant under normal soil conditions. This transgenic plant overexpressing the TaCRT showed delayed as well as less wilting as compared to control and wild type with GFP as a reporter gene. Furthermore, high WRC and WUE and low MDR were observed in these transgenic plants with overexpressed TaCRT as compared to control and wild plants.

Analysis at physiological and phenotypic levels indicated better growth in plants with higher TaCRT expression under drought stress conditions. These results were found to be in consensus with the report indicating high Ca^{2+} storage in plants with overexpressed calcium-binding domain of CRT. It was also observed in the very report that Arabidopsis plants transformed with maize C-domain showed substantial resistance to heavy metal, water, and salinity stress (Wyatt et al. 2002). So this showed that overexpression of TaCRT is related to drought resistance.

13 Molecular Mechanism of TaCRT Drought Resistance

As far as the molecular mechanism of TaCRT drought resistance is concerned, it is something yet to be determined. However, the unfolding and misfolding of TaCRT proteins have been observed in endoplasmic reticulum under environmental, physiological, and developmental stress (Borisjuk et al. 1998). So in this concern this misfolding and unfolding of proteins can be prevented by introducing molecular chaperones and overexpressing the enzymes responsible for protein folding located in ER lumen. CRT is one of the important proteins among those chaperones (Gelebart et al. 2005) (Fig. 5).

14 Overexpression of TaCRT and Protein Folding

The correct folding of proteins can be maintained by overexpression of TaCRT under stress conditions. This overexpression leads to an increased amount of chaperones in the cell. This overexpression is also associated with Ca^{2+} exchange ability of ER. This increased exchange ability will lead to new homeostasis and eventually high ER stress (Jin et al. 2005).

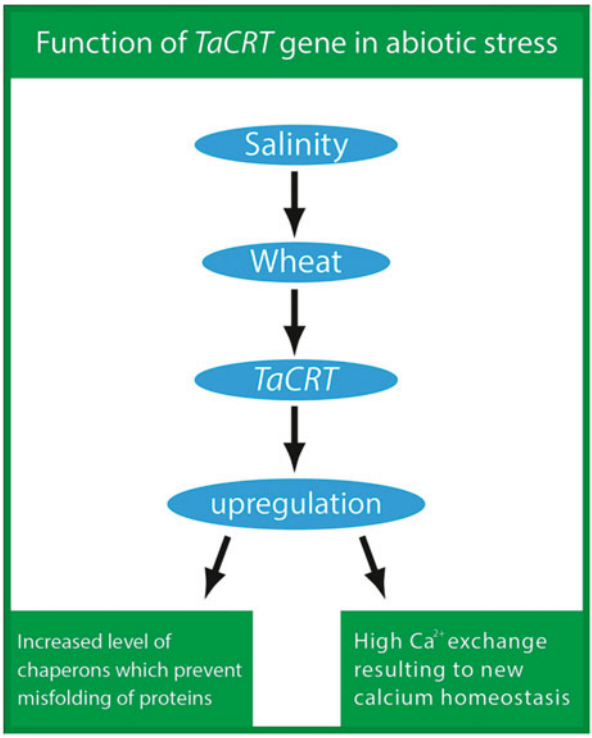


Fig. 5 Flowchart showing the regulation of CRT gene and its effect

15 TaCRT Gene and Drought-Specific Signaling Pathways

As the matter of fact, several signaling transduction pathways are responsible for regulation of plant responses against stress. It is possible that TaCRT might be involved in direct or indirect activation of specific signal transduction pathways, which would result in an enhanced metabolism. This enhanced metabolism enables the cells to protect them from drought stress injuries. So it can be conceived that calreticulin in wheat can be one of the positive regulators which are responsible for plant responses against drought. They promote these responses by $Ca2^{+}$ homeostasis and signaling modulation. These modulations basically ensure calreticulin availability and modulate gene expression. TaCRT overexpression was not found to have any adverse effect in adult plants (Xiong et al. 2008).

16 Calreticulin BrCRT1 Overexpression

Plant growth inhibition was observed in Chinese cabbage when Calreticulin 1 BrCRT1 was overexpressed. Enhanced organogenesis was observed in transgenic tobacco overexpressing BrCRT1 gene. This growth inhibition and stunted seedlings were found to be associated with the large amount of CRT which was expressed in these transgenic plants (Jin et al. 2005).

17 Drought-Inducible Genes and Their Functional Analysis

Expression of several genes can be induced by salinity, cold, and water stresses. Effects of ABA on expression induction of such inducible genes were investigated in two Arabidopsis mutants. One was a ABA-deficient mutant and the other a ABA-insensitive mutant. It was observed that several genes were induced under above-mentioned stresses in both ABA-deficient and ABA-insensitive mutants. This proposes that expression of such genes is independent of ABA under stress, but they showed response to exogenous ABA (Yamaguchi-Shinozaki and Shinozaki 1994, Bartels 1996; Shinozaki). Few drought-inducible genes are cor47 (rdl7), rd29A (lti78 or cor78), and cor6.6 (kin2). The most extensively analyzed gene rd 29Allti781cor78. The expression of this gene was studied under drought stress (Yamaguchi-Shinozaki and Shinozaki 1994).

18 Regulatory Systems of Drought Related Genes

The expression of genes in stress is always regulated by some sort of systems. Such type of regulatory systems also exists for drought and cold stress genes. Two regulatory systems have been found to be associated with expression of genes under cold and drought stress.

- ABA-independent system.
- ABA-dependent system.

TACCGACAT, a 9-base-pair conserved sequence, also termed as DRE, is thought to be essential in rd29A regulation under salinity, low temperature, and drought stress.

The most important thing is that this gene has nothing to do with ABRE (ABA-responsive element). The ABRE however is present in the promoter of rd29A gene which means its function is associated with ABA-responsive expression. Many drought and cold inductive genes were found to have DRE related motifs in their promoter region (Yamaguchi-Shinozaki and Shinozaki 1994). So these results indicate the involvement of DRE-related motifs including the core motif of C-domain containing CCGAC in drought and cold response. These responses are ABA-independent (Yamaguchi-Shinozaki and Shinozaki 1994) (Fig. 6).

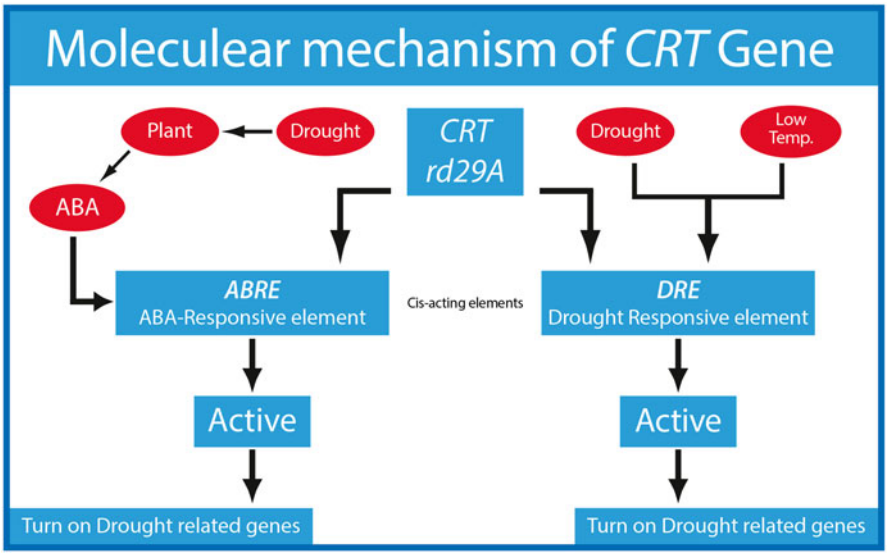


Fig. 6 Mechanism of CRT and how it triggers responses of DRE and ABRE

19 DRE Interacting Protein Factors

Protein factors are very much important for the expression and regulation of a gene. These protein factors basically interact with genes and carry out their regulation. Protein factors interacting with DRE were detected in water-stressed plants of Arabidopsis (Yamaguchi-Shinozaki and Shinozaki 1994). Several c-DNA for C repeat binding domains and DRE have been independently cloned by using yeast two hybrid screening method (Stockinger et al. 1997, Abe et al., unpublished data). A conserved DNA binding motif was observed in all DRE or C repeat-binding proteins. EREBP and AP2 proteins are also found to have such conserved regions. EREBP and AP2 proteins are also responsible for ethylene inductive gene expression and floral morphogenesis (Nakashima et al. 1997).

20 Drought Insensitive and Their Proteins

Several genes have been reported which are insensitive to cold and drought stress. This insensitivity suggests that another protein regulatory pathway might be present in the dehydration stress response. Different ethiol protease and Clp protease regulatory subunits (Nakashima et al. 1997) are encoded by the Rd21 and rd 19 genes. The promoter analysis of these important genes is yet to be determined.

21 Role of AtSOS1 Gene in Abiotic Stress

The mixture of different stresses like osmotic, oxidative, and ionic stresses results in salt stress. There is a huge decrease in the productivity of the crops due to salt stress. Salinity in soil and water poses a great threat by inhibiting agricultural crop productivity in the region which receives no rain or rain below the average level. It has been estimated that 25 % (one-third) of the total arable land is affected due to salinity. This 25 % of the land also includes 15 % of Iran's land. There are two types of stresses: biotic and abiotic, salt stress is the major abiotic stress globally. Na^+ is involved in different functions inside the cell which includes photosynthesis, enzyme activity in cytosol, different metabolism inside the cell, and potassium nutrition, and when there is a high concentration of Na^+ inside the cell, it has undesirable effects on cell functions (Gorham 1990; Schachtman and Munns 1992). When there is salt stress, it has been reported that there is decrease in the number of leaves (Kamkar et al. 2004; Amthor and McCree 1990). Due to the coalition of hydrogen peroxide and superoxide mediated oxidations, salt stress results in inhibition of photosynthesis (Hernandez et al. 1995). Plants acclimatize themselves to adverse conditions by evolving various immune processes for their survival and progress. Plants basically use three strategies to overcome and rectify high concentrations of sodium: Na^+ discharge, discrimination of Na^+ in the vacuole, and prevention of Na^+ inflow.

Making crops salt tolerant or increasing salt tolerance in crops is the main propagation target in areas where salinity and soil containing sodium restrict crop production and areas where there is constant variation in the salinity of soil, i.e., river mouths (estuaries) and tidal lands (intertidal zones) (de Leeuw et al. 1991).

22 Role of Calcineurin and Its Related Transcription Factors

There are some antiporters present such as Na^+/H^+ antiporter which promote the interchange of H^+ and Na^+ to the other side of the membrane. In plants, these antiporters have been designated in the plasma membrane and the vacuolar membrane. The primary duty of these antiporters is to eliminate Na^+ from the cytoplasm of the cell. So they help in warding off poisonous integral cellular assembling of Na^+ (Blumwald et al. 2000; Hasegawa et al. 2000). The signal pathway for the salt stress requires Ca^{++} increase. When Ca^{++} increases, it is felt by CBL4 protein known as calcineurin B-like protein. It is also known as SOS3. When Ca^{++} increases in the cytosol, it results in the formation of a complex of CBL4 and CIPK24 (CBL-interacting protein kinase). The Na^+/H^+ antiporter which is bound to the membrane is activated when phosphorylation of CBL4 and CIPK24 complex occurs and the complex reaches the plasma membrane through myristoyl fatty acid chain which is bound to CBL4. We cloned the AtSOS1 gene from salt overly sensitive mutants of Arabidopsis (Shi et al. 2000). We can also monitor the AtSOS1 function when we enhance salt tolerance in plants grown in laboratory.

Salt overly sensitive 1 commonly known as SOS1 is a genetic arrangement. It is required for salt tolerance. It was recognized firstly as a genetic arrangement in Arabidopsis plant for stress tolerance (Wu et al. 1996). This gene encodes for a antiporter, i.e., Na^+/H^+ antiporter, which is involved in reproduction and multiplication of plants under salt stress conditions. By analyzing the genetics of plants which are sensitive to salt a number of allelomorphs of AtSOS1 gene have been recognized. These allelomorphs were in controlled the sensitivity of plants in environments that contain high Na^+ (Koopman 2005; Wyn Jones and Pollard 1983; Zhu 2002; Koopman and Gort 2004).

23 Role of SOS1 Gene in Salinity

The signaling pathway of the salt overly sensitive gene was obsessive to have a crucial function in salt tolerance, which is responsible for maintaining ion homeostasis. This SOS pathway which is activated by Ca^{++} actually modulates the Na^+ and K^+ homeostasis. It has been clearly revealed that the salt overly sensitive protein family arbitrates salt tolerance either directly or indirectly. There are several types of salt overly sensitive genes: SOS1, SOS2, SOS3 (Ca^{2+} binding protein), and SOS4. SOS1 and SOS2 (a serine/threonine protein kinase that is activated by SOS3), and SOS4 are involved in the regulation and transport of Na^+. The Na^+/H^+ antiporter is encoded by SOS1 gene. This antiporter contains 12 transmembrane domains which are present in the N-terminal, whereas the C-terminal contains a tail which is hydrophilic. SOS2 and SOS3 complex is involved in the regulation of activity of SOS1. SOS1 gene which is the member of SOS family of proteins has 23 exons and 22 introns, and it produces a well-known protein which contains approximately 1146 residues of amino acids and molecular mass of this protein is 127 kDa. When we analyzed Arabidopsis plants which we breed in our laboratory (transgenic) that encodes for AtSOS1 gene and also a promoter GUS, they exhibited expression in parenchyma and epidermal cells at the tip of roots, stems and leaves respectively. Forty genes have been located which encode Na^+/H^+ antiporters in *Arabidopsis thaliana*. As a result of alternative splicing at the first intron of the SOS4 gene which encodes a transcript of 13 exons, it produces two vaticinating proteins which differ by 34 amino acids. Both these proteins have been identified as cytoplasmic pyridoxal kinases (PL). The function of these kinases is producing PL-5-phosphate. It has a dual function; it is a cofactor of an enzyme and also acts as a ligand for ion transporters. Whenever there is salt stress, SOS4 gene will be expressed (Shi et al. 2000). SOS4 helps in salt tolerance, but it is not involved in the pathways of SOS2, SOS3, and SOS4. There are two transcripts of SOS4 that have been identified: one transcript is long (1.4 kb) and one is short (1.3 kb). The short one is present abundantly compared to the long one. The short transcript is detected in stem, root, leaf, flower, and silique, with the highest abundance in leaf, whereas the long transcript is abundant in silique, flower, and root. It is present in minute quantities in stem, but not seen in leaf.

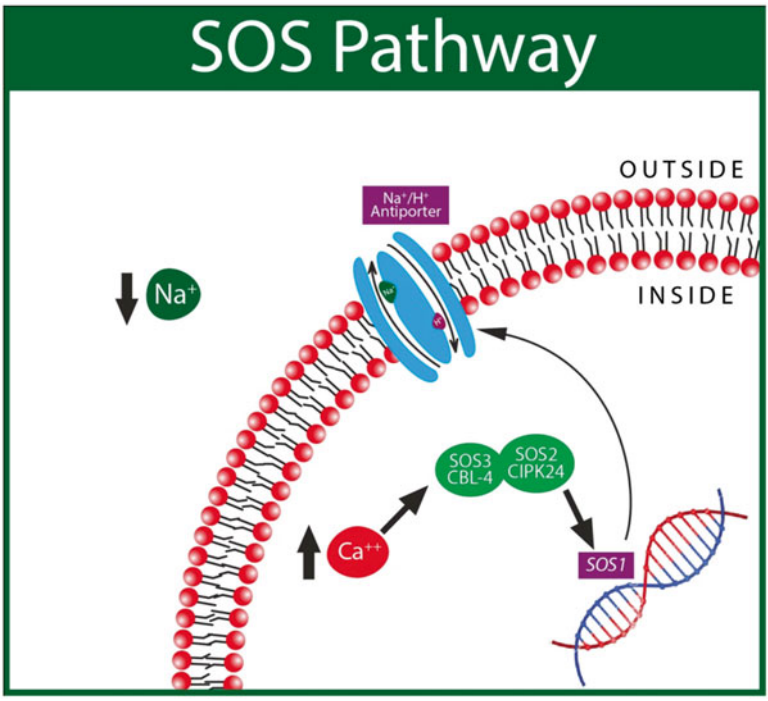


Fig. 7 Model for the salt overly sensitive (SOS) regulatory pathway. Salt-induced increases in cytoplasmic Ca^{++} are sensed by SOS3 (CBL-4). Ca^{++} together with SOS3 activates SOS2 (CIPK24). Activated SOS2 phosphorylates and stimulates SOS1, a plasma membrane localized Na^+/H^+ exchanger. It regulates ion homeostasis during salt stress

NaCl induces salt stress in soil (Blumwald et al. 2000; Hasegawa et al. 2000; Niu et al. 1995). Different experiments have been performed on different wheat plants like Alamut, Mahuti, *T. boeticum*, and *Aegilops crassa* to check the salinity effect on them at different levels. Mahuti (Iranian salt-tolerant) and Alamut (salt-tolerant) are wheat cultivars, whereas *T. boeticum* and *Aegilops crassa* are wild wheat plants (Ghavami et al. 2006) (Fig. 7).

24 Regulation of AtSOS1

Regulation of AtSOS1 was studied by expression analysis. The expression patterns of AtSOS1 and AtSOS4 genes were taken from different wheat plants by treating them different concentrations of salts and also different NaCl treatment periods. The expression of AtSOS1 gene is different in each cultivar when treated with different NaCl concentrations. The higher the salt concentration, the higher the AtSOS1 gene expression.

The wild wheat *A. crassa* expresses these genes, i.e., SOS1 and SOS4, at a high level, which indicates that SOS is involved in immunity against salt stress in the leaves of *A. crassa* wild wheat plant. Mahuti and Alamut show different patterns of AtSOS1 transcript. In Mahuti, initially there is decrease in the transcript level but it ascends after some time. This difference in the transcript is due to the transfer of Na^+ from root to shoot. The transcript level of AtSOS1 increases in Mahuti leaves, because there is increase in the concentration of Na^+ and Cl^- (Munns 2005). It has been reported that plants enhance the activity or level of expression of the transporter when these plants adapt the salt conditions which are new to them but when the level of sufficient ions achieved under saline conditions they are maintained by passive mechanisms (Niu et al. 1995). There is also upregulation and downregulation of some genes in Mahuti after salt stress (Ghavami et al. 2006).

25 Conclusions

The research quoted above demonstrates that CRT causes the transcription of certain genes that counter abiotic and harsh factors alike. The interaction of CRT is of vital importance to the cell. Different proteins bind to different regions of the DNA and such proteins are created from this interaction. This information clearly indicates that CRT and other proteins help counter drought and other stresses as well, and prompt interactions with various transcription factors and various genes.

Intreractomics not merely relates to CRT function but here we discuss the function of CRT genes which help to interact with other proteins, transcription factors, and DNA binding regions, enabling the cell to counter some stress of abiotic nature. Moreover, some proteins interact with other molecules in the cell to regulate different functions and counter stress. LEA is one such protein that serves as an osmoprotectant and also as a heat shock protein. Interactions of other proteins are also observed as they are related to synthesis of other proteins, enzymes which are utilized by cells for detoxification and antifreeze as well.

The facts demonstrated that CRT genes effect the cell's signaling pathways by regulating enzymes and other metabolic proteins. This also suggests that different CRTs are involved in different regulatory pathways. This shows functional diversity among CRT genes. All these facts demonstrated that TaCRT gene has a vital role in plant drought resistance. These drought related proteins are regulated by various protein factors. ABA also has a vital role in drought associated responses of genes. DRE and C-repeat binding domains are the critical sites for interaction of proteins at the promoter level.

It seems that when there is Na^+ stress, AtSOS1 gene is not the only component that contributes to salt resistance in wheat. Expression of genes encoding proteins like pyrroline carboxylate synthetase (P5cS) and betaine aldehyde dehydrogenase (BADH) (osmotic responsive) are also involved in the process. TaHKT1;5 gene expression is involved in the recirculation of Na^+ to roots from shoots in Mahuti as compared to AtSOS1 gene expression. For Na^+ transfer, Na^+ is loaded into xylem in

roots and for storage in the mesophyll vacuoles in leaves. It is controlled by SOS1 gene activity under average stress. But when Na^+ stress is acute, SOS1 protein unloads Na^+ from the root xylem. This helps prevent destruction of leaves by Na^+ segregation in the vacuoles of leaf cells.

26 Future Perspectives

In this review we focus on the interaction of different genes with their transcriptional factors as well as their role in different environmental stresses. This study can be used to identify different genes that are involved in different stresses. Genes that are of vital importance in salt resistance and other biotic and abiotic stresses are discussed. We can manipulate the expression of these genes by using the information about their regulation by different transcriptional factors like CRT and other regulatory proteins. These manipulations can be used to enhance their role in stress resistance in wheat and other plant crops. This study can also be used for understanding of genetic variations in different genes like TaCRT and SOS gene family, in various abiotic stresses like drought. We can also utilize the certain drought-insensitive gene mentioned in this study, for induction of drought resistance by genetic transfer.

References

Agarwal PK, Jha B (2010) Transcription factors in plants and ABA dependent and independent abiotic stress signaling. Biologia Plantarum 54:201–212

Amthor JS, McCree KJ (1990) Carbon balance of stressed plants: a conceptual model for integrating research results. In: Ascher RG, Cumming JR (eds) Stress responses in plants: adaptation and acclimation mechanisms. Wiley-Liss, Inc., Wilmington, DE, pp 1–15

Auerbach D, Thaminy S, Hottiger MO, Stagljar I (2002) The post-genomic era of interactive proteomics: facts and perspectives. Proteomics 2:611–623

Bartels J (1996) The molecular basis of dehydration tolerance in plants. Annu Rev Plant Physiol Plant Mol Biol 47:377–403

Bies-Ethève N, Gaubier-Comella P, Debures A, Lasserre E, Jobet E, Raynal M, Cooke R, Delseny M (2008) Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in Arabidopsis thaliana. Plant Mol Biol 67:107–124

Blumwald E, Aharon GS, Apse MP (2000) Sodium transport in plant cells. Biochim Biophys Acta 1465:140–151

Borisjuk N, Sitailo L, Adler K, Malysheva L, Tewes A, Borisjuk L, Manteuffel R (1998) Calreticulin expression in plant cells: developmental regulation, tissue specificity and intracellular distribution. Planta 206:504–514

Bray EA (1997) Plant responses to water deficit. Trends Plant Sci 2:48–54

Bray EA (2002) Abscisic acid regulation of gene expression during water-deficit stress in the era of the Arabidopsis genome. Plant Cell Environ 25:153–161

Bray GA (2004) Medical consequences of obesity. J Clin Endocrinol Metab 89(6):2583–2589

Broder YC, Katz S, Aronheim A (1998) The ras recruitment system, a novel approach to the study of protein-protein interactions. Curr Biol 8:1121–1124

Brown SL, Nesse RM, Vinokur AD, Smith DM (2003) Providing social support may be more beneficial than receiving it: results from a prospective study of mortality. Psychol Sci 14:320–327

Chaves MM, Maroco JP, Pereira JS (2003) Understanding plant responses to drought—from genes to the whole plant. Function Plant Biol 30:239–264

Chen H, Kunnimalaiyaan M, Van Gompel JJ (2005) Medullary thyroid cancer: the functions of raf-1 and human achaete-scute homologue-1. Thyroid I5:511–521

Close TJ, Kortt AA, Chandler PM (1989) A cDNA-based comparison of dehydration-induced proteins (dehydrins) in barley and corn. Plant Mol Biol 13(1):95–108

Coppolino MG, Dedhar S (1998) Calreticulin. Int J Biochem Cell Biol 30(5):553–558

Coppolino M, Woodside MJ, Demaurex N, Grinstein S, St-Arnaud R, Dedhar S (1997) Calreticulin is essential for integrin-mediated calcium signaling and cell adhesion. Nature 386:843–847

Danyluk J, Houde M, Rassart E, Sarhan F (1994) Differential expression of a gene encoding an acidic dehydrin in chilling sensitive and freezing tolerant graminae species. FEBS Lett 344:20–24

De Leeuw J, van den Dool A, de Munck W, Nieuwenhuize J, Beeftink WG (1991) Factors influencing the soil salinity regime along an intertidal gradient. Estuar Coast Shelf Sci 32(1):87–97

Denecke J, Carlsson LE, Vidal S, Höglund A-S, Ek B, Van Zeijl MJ, Sinjorgo KMC, Palva ET (1995) The tobacco homolog of mammalian calreticulin is present in protein complexes in vivo. Plant Cell 7:391–406

Dortay H, Gruhn N, Pfeifer A, Schwerdtner M, Schmulling T, Heyl A (2008) Toward an interaction map of the two-component signaling pathway of Arabidopsis thaliana. J Proteome Res 2008(7):3649–3660

Dure L, Greenway SC, Galau GA (1981) Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by in vitro and in vivoprotein synthesis. Biochemistry 20:4162–4168

Fliegel L, Burns K, MacLennan DH, Reithmeier RA, Michalak M (1989) Molecular cloning of the high affinity calcium-binding protein (calreticulin) of skeletal muscle sarcoplasmic reticulum. J Biol Chem 264(36):21522–21528

Forler D, Kocher T, Rode M, Gentzel M, Izaurralde E, Wilm M (2003) An efficient protein complex purification method for functional proteomics in higher eukaryotes. Nat Biotechnol 21:89–92

Gardai SJ, Mcphillips KA, Frasch SC, Janssen WJ, Starefeldt A, Murphy-Ullrich JE, Bratton DL, Oldenborg PA, Michalak M, Henson PM (2005) Cell-surface calreticulin initiates clearance of viable or apoptotic cells through trans-activation of LRP on the phagocyte. Cell 123(2):321–334

Gelebart P, Opas M, Michalak M (2005) Calreticulin, a Ca^{2+}-binding chaperone of the endoplasmic reticulum. Int J Biochem Cell Biol 37:260–266

Ghavami F, Malboobi MA, Ghannadha MR, Lohrasebi T, Yazdi-Samadi B, Mozaffari J et al (2006) Up-regulation of succinate dehydrogenase and prophobilinogen deaminase in response to high salt concentration in wheat. Iran J Agric Sci 36(6):1437–1444

Gorham J (1990) Salt tolerance in the Triticeae: K?Na discrimination in synthetic hexaploid wheats. J Exp Bot 41:623–627

Guo H, Ecker JR (2004) The ethylene signaling pathway: new insights. Curr Opin Plant Biol 7:40–49

Guo Z-H, Chen Y-Y, Li D-Z (2002) Phylogenetic studies on theThamnocalamus group and its allies (Gramineae: Bambusoideae) based on ITS sequence data. Mol Phylogenetics Evol 22:20–30

Hasegawa PM, Bressan RA, Zhu JK, Bohnert HJ (2000) Plant cellular and molecular responses to high salinity. Annu Rev Plant Physiol Plant Mol Biol 51:463–499

Hernandez JA, Olmos E, Corpas FJ, Sevilla F, Del-Rio LA (1995) Salt-induced oxidative stress in chloroplasts of pea plants. Plant Sci 105:151–167

Hideaki S (1995) Identification of an ethylene-responsive region in the promoter of a tobacco class I chitinase gene. Plant Mol Biol 27(5):923–932
Houde A, Kademi A, Leblanc D (2004) Lipases and their industrial applications: an overview. App Biochem Biotechnol 118:155–170
Hussain I, Riazullah R, Khurram M, Naseemullah A, Baseer FA, Khan MR, Khattak M, Zahoor JK, Khan N (2011) Phytochemical analysis of selected medicinal plants. Afr J Biotechnol 10(38):7487–7492
Ingram and Bartels (1996) The molecular basis of dehydration tolerance in plants. Annu Rev Plant Physiol Plant Mol Biol. 47:377–403
Ivanov AS, Zgoda VG, Archakov AI (2011) Technologies of protein interactomics: a review. Russ J Bioorg Chem 37:4–16
Jin ZL, Hong JK, Yang KA, Koo JC, Choi YJ, Chung WS, Yun DJ, Lee SY, Cho MJ, Lim CO (2005) Over-expression of Chinese cabbage calreticulin 1, BrCRT1, enhances shoot and root regeneration, but retards plant growth in transgenic tobacco. Transgenic Res 14:619–626
Kamkar B, Kafi M, Nassiri-Mahallati M (2004) Determination of the most sensitive developmental period of wheat (Triticum aestivum) to salt stress to optimize saline water utilization. Australian agronomy conference, 12th AAC, 4th ICSC
Kerppola TK (2008) Bio-molecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells. Annu Rev Biophys 37:465–487
Koopman WJM (2005) Phylogenetic signal in AFLP data sets. Syst Biol 54:197–217
Koopman WJM, Gort G (2004) Significance tests and weighted values for AFLP similarities based on Arabidopsis in silico AFLP fragment length distributions. Genetics 167:1915–1928
Krause KH, Michalak M (1997) Calreticulin. Cell 88(4):439–443
Laporte C, Vetter G, Loudes A-M et al (2003) Involvement of the secretory pathway and the cytoskeleton in intracellular targeting and tubule assembly of grapevine fanleaf virus movement protein in tobacco BY-2 cells. Plant Cell 15(9):2058–2075. doi:10.1105/tpc.013896
Lata C, Prasad M (2011) Role of DREBs in regulation of abiotic stress responses in plants. J Exp Bot err210
Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet 38:896–903
Lindboet JA, Fitzmaurice WP, Della-Cioppa G (2001) Virus mediated reprogramming of gene expression in plants. Curr Opin Plant Biol 4:181–185
Maruyama K, Sakuma Y, Kasuga M, Ito Y, Seki M, Goda H, Shimada Y, Yoshida S, Shinozaki K, Yamaguchi-Shinozaki K (2004) Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. Plant J 38:982–993
Mundy P, Sigman M, Kasari C (1990) A longitudinal study of joint attention and language development in autistic children. J Autism Dev Disorders 20:115–128. doi:10.1007/BF02206861
Munns R (2005) Genes and salt tolerance: bringing them together. New Phytol 167:645–663
Nakashima K, Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K (1997) A nuclear gene, erd1, encoding a chloroplast-targeted Clp protease regulatory subunit homolog is not only induced by water stress but also developmentally up-regulated during senescence in *Arabidopsis thaliana*. Plant J 12:851–861
Niu X, Bressan RA, Hasegawa PM, Pardo JM (1995) Ion homeostasis in NaCl stress environments. Plant Physiol 109:735–742
Ostwald TJ, Maclennan DH (1974) Isolation of a high affinity calcium-binding protein from sarcoplasmic reticulum. J Biol Chem 249(3):974–979
Park H, Suzuki T, Lennarz WJ (2001) Identification of proteins that interact with mammalian peptide: Nglycanase and implicate this hydrolase in the proteasome-dependent pathway for protein degradation. Proc Natl Acad Sci USA 98(20):11163–11168
Persson S, Rosenquist M, Svensson K, Galvão R, Boss WF, Sommarin M (2003) Phylogenetic analysis and expression studies reveal two distinctive groups of Calreticulin isoform in higher plants. Plant Physiol 133:1385–1396, Science 24:23–58

Ramanjulu S, Bartels D (2002) Drought- and desiccation-induced modulation of gene expression in plants. Plant Cell Environ 25:141–151
Sachs MM, Freeling M, Okimoto R (1980) The anaerobic proteins of maize. Cell 20:761–767
Sakuma Y, Maruyama K, Osakabe Y, Qin F, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2006) Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression. Plant Cell 18:1292–1309
Schachtman DP, Munns R (1992) Sodium accumulation in leaves of Triticum species that differ in salt tolerance. Aus J Plant Physiol 19:331–340
Schneider K, Wells B, Schmelzer E, Salamini F, Bartels D (1993) Desiccation leads to the rapid accumulation of both cytosolic and chloroplastic proteins in the resurrection plant Craterostigma plantagineum Hochst. Planta 189:120–131
Schumacher MA, Bashor CJ, Song MH, Otsu K, Zhu S, Parry RJ, Ullman B, Brennan RG (2002) The structural mechanism of GTP stabilized oligomerization and catalytic activation of the Toxoplasma gondii uracil phosphoribosyltransferase. Proc Natl Acad Sci USA 99(1):78–83
Serebriiskii IG, Mitina OV, Chernoff J, Golemis EA (2001) Two-hybrid dual bait system to discriminate specificity of protein interactions in small GTPases. Methods Enzymol 332:277–300
Shi H, Ishitani M, Kim C, Zhu J-K (2000) The *Arabidopsis thaliana* salt tolerance gene *SOS1* encodes a putative Na+/H+ anti-porter. Proc Natl Acad Sci USA 97(12):6896–6901. doi:10.1073/pnas.120170197
Shinozaki K, Yamaguchi-Shinozaki K, Seki M (2003) Regulatory network of gene expression in the drought and cold stress responses. Curr Opin Plant Biol 6:410–417
Somerville C, Briscoe J (2001) Genetic engineering and water. Science 292:2217
Stockinger E, Gilmour SJ, Thomashow MF (1997) Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci U S A 94:1035–1040
Thomashow MF (2001) So what's new in the field of plant cold acclimation? Lots! Plant Physiol 125:89–93
Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294:2364–2368
Van Leene J, Witters E, Inze D, De Jaeger G (2008) Boosting tandem affinity purification of plant protein complexes. Trends Plant Sci 13:517–520
Veeranagamallaiah G, Prasanthi J, Reddy KE, Merum P, Mtwisha L, Brandt W, McCready S, Lindsey GG (1998) HSP 12 is a LEA-like protein in Saccharomyces cerevisiae. Plant Mol Biol 37:513–521
Wang Y, Shirogane T, Liu D, Harper JW, Elledge SJ (2003) Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling. Cell 112(5):697–709
Wise MJ (2003) Leaping to conclusions: a computational reanalysis of late embryogenesis abundant proteins and their possible roles. BMC Bioinformatics 4:52
Wu S-J, Ding L, Zhu J-K (1996) *SOS1*, a genetic locus essential for salt tolerance and potassium acquisition. Plant Cell 8(4):617–627. doi:10.1105/tpc.8.4.617
Wyatt SE, Tsou PL, Robertson D (2002) Expression of the high capacity calcium-binding domain of calreticulin increases bioavailable calcium stores in plants. Transgenic Res 11:1–10
Wyn Jones RG, Pollard A (1983) Encyclopedia of plant physiology. Springer, Berlin
Xia-yun J, Xul C, Jing R, Li R, Mao X, Wang J, Chang X (2008) Molecular cloning and characterization of wheat calreticulin (CRT) gene involved in drought-stressed responses. J Exp Bot 4:739–751
Xiong J-W, Yu Q, Zhang J, Mably JD (2008) An acyltransferase controls the generation of hematopoietic and endothelial lineages in zebrafish. Circ Res 102:1057–1064
Xu ZS, Xia LQ, Chen M, Cheng XG, Zhang RY, Li LC, Zhao YX, Lu Y, Ni ZY, Liu L, Qiu ZG, Ma YZ (2007) Isolation and molecular characterization of the Triticum aestivum L. ethylene-responsive factor 1 (TaERF1) that increases multiple stress tolerance. Plant Mol Biol 65(6):719–732

Xu ZS, Ni Z-Y, Liu L, Nie LN, Li LC, Chen M, Ma YZ (2008) Characterization of the TaAIDFa gene encoding a CRT/DRE-binding factor responsive to drought, high-salt, and cold stress in wheat. Mol Genet Genomics 280:497–508

Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. Plant Cell 6:251–264

Yamaguchi-Shinozaki K, Shinozaki K (2000) A stress-inducible gene for 9-cis-epoxycarotenoid dioxygenase involved in abscisic acid biosynthesis under water stress in drought-tolerant cowpea. Plant Physiol 123:553–562

Yi SY, Kim JH, Joung YH, Lee S, Kim WT, Yu SH, Choi D (2004) The pepper transcription factor CaPF1 confers pathogen and freezing tolerance in Arabidopsis. Plant Physiol 136(1):2862–2874

Zhang Y, Wang Z, Xu J (2007) Molecular mechanism of dehydrin in response to environmental stress in plant. Prog Nat Sci 17:237–246

Zhu JK (2002) Salt and drought stress signal transduction in plants. Annu Rev Plant Biol 53(1):247–273