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

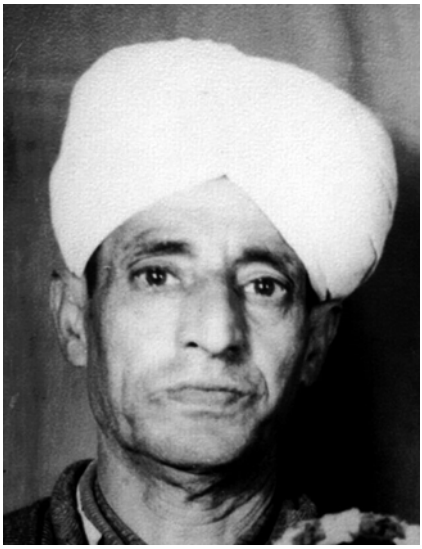
Güzin Tombuloğlu
Pathology Laboratory Techniques Program
Vocational School of Medical Sciences
Fatih University
Buyukcekmece, Istanbul, Turkey

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