Compendium of Plant Genomes

Series Editor: Chittaranjan Kole

Zhihong Gao *Editor*

The Prunus mume Genome

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, Raja Ramanna Fellow, Government of India, ICAR-National Research Center on Plant Biotechnology, Pusa, New Delhi, India

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant Arabidopsis thaliana in 2000, whole genomes of about 100 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

Interested in editing a volume on a crop or model plant?

Please contact Dr. Kole, Series Editor, at ckoleorg@gmail.com

More information about this series at http://www.springer.com/series/11805

Zhihong Gao
Editor

The Prunus mume Genome

Springer

Editor
Zhihong Gao
College of Horticulture
Nanjing Agricultural University
Nanjing, China

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-030-10796-3 ISBN 978-3-030-10797-0 (eBook)
https://doi.org/10.1007/978-3-030-10797-0

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

This book series is dedicated to my wife Phullara, and our children Sourav, and Devleena

Chittaranjan Kole

Preface to the Series

Genome sequencing has emerged as the leading discipline in the plant sciences coinciding with the start of the new century. For much of the twentieth century, plant geneticists were only successful in delineating putative chromosomal location, function, and changes in genes indirectly through the use of a number of "markers" physically linked to them. These included visible or morphological, cytological, protein, and molecular or DNA markers. Among them, the first DNA marker, the RFLPs, introduced a revolutionary change in plant genetics and breeding in the mid-1980s, mainly because of their infinite number and thus potential to cover maximum chromosomal regions, phenotypic neutrality, absence of epistasis, and codominant nature. An array of other hybridization-based markers, PCR-based markers, and markers based on both facilitated construction of genetic linkage maps, mapping of genes controlling simply inherited traits, and even gene clusters (QTLs) controlling polygenic traits in a large number of model and crop plants. During this period, a number of new mapping populations beyond F_2 were utilized and a number of computer programs were developed for map construction, mapping of genes, and for mapping of polygenic clusters or QTLs. Molecular markers were also used in the studies of evolution and phylogenetic relationship, genetic diversity, DNA fingerprinting, and map-based cloning. Markers tightly linked to the genes were used in crop improvement employing the so-called marker-assisted selection. These strategies of molecular genetic mapping and molecular breeding made a spectacular impact during the last one and a half decades of the twentieth century. But still they remained "indirect" approaches for elucidation and utilization of plant genomes since much of the chromosomes remained unknown and the complete chemical depiction of them was yet to be unraveled.

Physical mapping of genomes was the obvious consequence that facilitated the development of the "genomic resources" including BAC and YAC libraries to develop physical maps in some plant genomes. Subsequently, integrated genetic–physical maps were also developed in many plants. This led to the concept of structural genomics. Later on, emphasis was laid on EST and transcriptome analysis to decipher the function of the active gene sequences leading to another concept defined as functional genomics. The advent of techniques of bacteriophage gene and DNA sequencing in the 1970s was extended to facilitate sequencing of these genomic resources in the last decade of the twentieth century.

As expected, sequencing of chromosomal regions would have led to too much data to store, characterize, and utilize with the-then available computer software could handle. But the development of information technology made the life of biologists easier by leading to a swift and sweet marriage of biology and informatics, and a new subject was born—bioinformatics.

Thus, the evolution of the concepts, strategies, and tools of sequencing and bioinformatics reinforced the subject of genomics—structural and functional. Today, genome sequencing has traveled much beyond biology and involves biophysics, biochemistry, and bioinformatics!

Thanks to the efforts of both public and private agencies, genome sequencing strategies are evolving very fast, leading to cheaper, quicker, and automated techniques right from clone-by-clone and whole-genome shotgun approaches to a succession of second-generation sequencing methods. The development of software of different generations facilitated this genome sequencing. At the same time, newer concepts and strategies were emerging to handle sequencing of the complex genomes, particularly the polyploids.

It became a reality to chemically—and so directly—define plant genomes, popularly called whole-genome sequencing or simply genome sequencing.

The history of plant genome sequencing will always cite the sequencing of the genome of the model plant Arabidopsis thaliana in 2000 that was followed by sequencing the genome of the crop and model plant rice in 2002. Since then, the number of sequenced genomes of higher plants has been increasing exponentially, mainly due to the development of cheaper and quicker genomic techniques and, most importantly, the development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this preface for the first volume of the new series “Compendium of Plant Genomes,” a net search tells me that complete or nearly complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and 3 basal plants is accomplished, the majority of which are in the public domain. This means that we nowadays know many of our model and crop plants chemically, i.e., directly, and we may depict them and utilize them precisely better than ever. Genome sequencing has covered all groups of crop plants. Hence, information on the precise depiction of plant genomes and the scope of their utilization are growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated Web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful both to students and teaching faculties. Most importantly, research scientists involved in genomics research will have access to systematic deliberations on the plant genomes of their interest. Elucidation of plant genomes is of interest not only for the geneticists and breeders, but also for practitioners of an array of plant science disciplines, such as taxonomy, evolution, cytology, physiology, pathology, entomology, nematology, crop production, biochemistry, and obviously bioinformatics. It must be mentioned that information regarding each plant genome is ever-growing. The contents of the volumes of this compendium are, therefore, focusing on the basic aspects of the genomes and their utility. They include information on the academic and/or economic importance of the plants, description of their genomes from a molecular genetic and cytogenetic point of view, and the genomic resources developed. Detailed deliberations focus on the background history of the national and international genome initiatives, public and private partners involved, strategies and genomic resources and tools utilized, enumeration on the sequences and their assembly, repetitive sequences, gene annotation, and genome duplication. In addition, synteny with other sequences, comparison of gene families, and, most importantly, the potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

I must confess that as the series editor, it has been a daunting task for me to work on such a huge and broad knowledge base that spans so many diverse plant species. However, pioneering scientists with lifetime experience and expertise on the particular crops did excellent jobs editing the respective volumes. I myself have been a small science worker on plant genomes since the mid-1980s and that provided me the opportunity to personally know several stalwarts of plant genomics from all over the globe. Most, if not all, of the volume editors are my longtime friends and colleagues. It has been highly comfortable and enriching for me to work with them on this book series. To be honest, while working on this series I have been and will remain a student first, a science worker second, and a series editor last. And I must express my gratitude to the volume editors and the chapter authors for providing me the opportunity to work with them on this compendium.

I also wish to mention here my thanks and gratitude to the Springer staff, Dr. Christina Eckey and Dr. Jutta Lindenborn in particular, for all their constant and cordial support right from the inception of the idea.

I always had to set aside additional hours to edit books beside my professional and personal commitments—hours I could and should have given to my wife, Phullara, and our kids, Sourav, and Devleena. I must mention that they not only allowed me the freedom to take away those hours from them but also offered their support in the editing job itself. I am really not sure whether my dedication of this compendium to them will suffice to do justice to their sacrifices for the interest of science and the science community.

Kalyani, India Chittaranjan Kole

Preface

Prunus mume (*Prunus mume* Sieb. et Zucc), commonly known as Japanese apricot, is an ornamental and fruit tree plant. It originated in China, where it was domesticated more than 3000 years ago, and its genome is one of the first genomes among the *Prunus* subfamily of the Rosaceae family to be sequenced. The genome of *Prunus mume* was fully sequenced in 2012. A 280-Mb genome was assembled combining 101-fold next-generation sequencing and optical mapping data. Scaffolds of 83.9% were further anchored to eight chromosomes, with the genetic map being constructed by restriction-site-associated DNA sequencing. Combining the *P. mume* genome with available data, scientists succeeded in reconstructing the nine ancestral chromosomes of the Rosaceae family, as well as in depicting the chromosome fusion, fission and duplication history in three major subfamilies. The transcriptome of various tissues and genome-wide analysis revealed the characteristics of *P. mume*, including the regulation of early blooming in endo-dormancy, the immune response against bacterial infection and the biosynthesis of the flower scent. The *P. mume* genome sequence increases our understanding of Rosaceae evolution and provides important data for the improvement of fruit trees in this family.

This book aims at reviewing of whole-genome sequencing, including the nuclear genome, the chloroplast genome, the functional genome, the molecular biology, molecular markers, epigenetics and the genetic relationship with other Rosaceae species.

11 scientists from two countries have authored 14 chapters of this book to illustrate the strategies of whole-genome sequencing and advanced breeding techniques. Experts from China and Japan, involved in *Prunus mume* research, investigated the different aspects of *Prunus mume* genetics and how genome sequencing could affect biological explanation, breeding and production in *Prunus mume*.

The book will be a guide for those who are interested in gene discovery, comparative genomics as well as molecular and advanced breeding techniques. It will be particularly useful for scientists, breeders and students involved in research related to the development of the citrus industry for updating the amount of knowledge generated in recent years.

In Chap. 1, the production, origin and the present distribution, as well as the economic importance and medical value of *Prunus mume*, are described. The botanical description is covered in Chap. 2, which provides essential information about genetics and molecular mapping. In Chap. 4, the focus is on taxonomy and germplasm. The highlight of this book is Chap. 5, which introduces whole-genome sequencing as well as gene annotation and genome evolution. Molecular mapping, systems biology, small RNA, transcriptomics and the molecular and developmental biology on self-incompatibility and pistil abortion are described in the other chapters. In addition, genome sequence-based marker development, with the knowledge of the genome, will be discussed.

Nanjing, China
February 2019

Zhihong Gao

Contents

Contributors

Wenxing Chen Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Zhihong Gao College of Horticulture, Nanjing Agricultural University, Nanjing, People's Republic of China

Shahid Iqbal College of Horticulture, Nanjing Agricultural University, Nanjing, People's Republic of China

Yuto Kitamura Japanese Apricot Laboratory, Wakayama Fruit Experimental Station, Wakayama, Japan

Wenjie Luo College of Horticulture, Nanjing Agricultural University, Nanjing, People's Republic of China

Xiaopeng Ni College of Horticulture, Nanjing Agricultural University, Nanjing, People's Republic of China

Ting Shi College of Horticulture, Nanjing Agricultural University, Nanjing, People's Republic of China

Lidan Sun Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Engineering Research Center of Landscape Environment of Ministry of Education, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, School of Landscape Architecture, Beijing Forestry University, Beijing, China

Ryutaro Tao Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Hisayo Yamane Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Qixiang Zhang Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Engineering Research Center of Landscape Environment of Ministry of Education, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, School of Landscape Architecture, Beijing Forestry University, Beijing, China

Abbreviations

2-DE	Two-dimensional gel electrophoresis
ABA	Abscisic acid
ACC	1-aminocyclopropane-1-carboxylic acid
AFLP	Amplified fragment length polymorphism
AGL24	*AGAMOUS-LIKE24*
AP2/ERF	APETALA 2/ethylene-responsive element binding factor
AS-PCR	Allele-specific polymerase chain reaction
At	*Arabidopsis thaliana*
BAC	Bacterial artificial chromosomes
BD	Blooming date
BEAT	Benzyl alcohol acetyltransferase
BR	Bud break
CBF	C-repeat binding transcription factor
cDNA	Complementary DNA
CEN	Centroradialis
CO	CONSTANS
cpDNA	Chloroplast DNA
CR	Chilling requirement
CRL	Chilling requirement of leaf buds
CS	Cell structure
CUL1s	Cullin1-likes
DAM	Dormancy-associated MADS-box transcription factors
DCL3	*dicer-like 3*
DGE	Digital gene expression
DUBs	Deubiquitinating enzymes
EM	Energy metabolism
EST	Expressed sequence tag
FAB1	Fatty acid biosynthesis 1
FAD5	Fatty acid desaturase 5
FB	Flower buds
FT	Flowering Locus T
Fv	*Fragaria vesca*
GA	Gibberellin
GBS	Genotyping-by-sequencing
GI	General inhibitor
GO	Gene ontology

GPX	Glutathione peroxidase
GS	Genomic selection
GSI	Gametophytic self-incompatibility
GWAS	Genome-wide association study
HR	Heat requirement
HRL	Heat requirement for bud break of leaf buds
HVR	Hypervariable region
IF	Imperfect
IPCC	Intergovernmental Panel on Climate Change
IR	Inverted repeat
IRMP	International Rosaceae mapping project
JADB	Japanese apricot dormant bud EST database
LAMP	Loop-mediated isothermal amplification
LC–MS/MS	Liquid chromatography and tandem mass spectrometry
LD	Leafing date
LEA	Late embryogenesis-abundant
LFY	LEAFY
LG	Linkage group
LRR	Leucine repeat sequence
LSC	Large single copy region
LTP3	Lipid transfer protein 3
LTR	Long terminal repeat
MAB	Marker-assisted breeding
MALDI-TOF/TOF	Matrix-assisted laser desorption/ionization time of flight/time of flight
MAS	Marker-assisted selection
Md	*Malus* × *domestics*
miRNA	Micro RNA
MOS	Mirror orientation selection
mRNA	Messenger RNA
NBS	Nuclear binding site
NGS	Next-generation sequencing
OR	Oxidation–reduction
Os	*Oryza sativa*
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PEs	Paired ends
PF	Perfect
PM	Protein metabolism
Pm	*Prunus mume*
PME	Pectin methyl esterase
PPME1	Pectin lyase-like superfamily protein
Pt	*Populus trichocarpa*
PVE	Phenotypic variance explained
QTLs	Quantitative trait locus
R	Resistance
RACE	Rapid amplification of the cDNA ends
RAD	Restriction-site-associated DNA

RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
RGA	Resistance gene analogues
RGS	Reference genome sequence
RNase	Ribonuclease
RNA-Seq	Transcriptome sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription-polymerase chain reaction
SC	Self-compatible
SD	Stress and defence
SF	Self-fruitfulness
SFB	S haplotype-specific F-box protein gene
SFBB	S-locus F-box brothers
SI	Self-incompatibility
SLAF-seq	Specific locus amplified fragment sequencing
SLF	S-locus F-box
SMRM	Single-molecule restriction map
SMRT	Single-molecule real time
SNP	Single-nucleotide polymorphism
snRNA	Small nuclear RNA
SOC1	Suppressor of overexpression of CONSTANS 1
SSC	Small single copy region
SSH	Suppression subtractive hybridization
SSR	Simple sequence repeat
SU	Self-unfruitfulness
SVP	*Short Vegetative Phase*
TCP2	Teosinte Branched/Cycloidea/PCF
TE	Transposable element
TFL1	Terminal Flower1
tRNA	Transfer RNA
VB	Vegetative buds
Vv	*Vitis vinifera*
WGD	Whole-genome duplication
WGM	Whole-genome mapping
WGS	Whole-genome sequencing
XTH2	Xyloglucan endotransglucosylase/hydrolase 2