Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Amanda R. De La Torre *Editor*

The Pine Genomes

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, President, International Climate Resilient Crop Genomics Consortium (ICRCGC), President, International Phytomedomics & Nutriomics Consortium (IPNC) and President, Genome India International (GII), Kolkata, 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.

More information about this series at https://link.springer.com/bookseries/11805

Amanda R. De La Torre
Editor

The Pine Genomes

Springer

Editor
Amanda R. De La Torre
Northern Arizona University
Flagstaff, AZ, USA

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-030-93389-0 ISBN 978-3-030-93390-6 (eBook)
https://doi.org/10.1007/978-3-030-93390-6

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 F2 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 crops and model plants, eight model plants, 15 crop progenitors and relatives, and three 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 to both 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, particularly Dr. Christina Eckey and Dr. Jutta Lindenborn, for the earlier set of volumes and presently Ing. Zuzana Bernhart for all their timely help and support.

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.

New Delhi, India Chittaranjan Kole

Preface

Pines (*Pinus*) are the world's most economically important forest tree species. With more than 100 species, pines are also the most abundant extant group of Gymnosperms. Pines are naturally distributed in the Northern hemisphere, where they inhabit pure or mixed-species forests or are planted for commercial uses. Some species such as *Pinus radiata* are also planted as commercial species in the Southern hemisphere. Efforts to understand their complex biology, functions and evolution were limited by their non-model system attributes (e.g., long generation times, slow growth, difficulty to clone or vegetative propagate) and huge genome sizes (20–40 Gbp) with high percentages (>70%) of repeat sequences, mostly transposable elements. In the last five years, improved and more accessible sequencing and bioinformatic tools have allowed significant changes in the study of the genomics and transcriptomics of pines. Since 2014, four species (*Pinus taeda*, *Pinus lambertiana*, *Pinus pinaster* and *Pinus radiata*) have been sequenced, and numerous transcriptomic resources have been developed.

This book is the first comprehensive compilation of the most up-to-date research in the genomics, transcriptomics and breeding of pine species across Europe, North America and Australia. The twelve chapters in this book aim to cover different aspects in genomic and transcriptomic research mainly focusing on the species with sequenced genomes but also in other pines of ecological and economical importance. In the Chap. 1, recent advances in whole-genome sequencing, transcriptome sequencing and target enrichment of nuclear genes for North American pine species are described. In the absence of chromosome-level reference genomes, studies on the genome architecture have been based on the presence of genetic and linkage maps. Genetic mapping and comparative mapping approaches are reviewed with an emphasis on *P. taeda* in Chap. 2.

Transposable elements are major components of pines and gymnosperm genomes. Although initial studies have revealed important information on their structure, classification and genome organization, many questions remain regarding their role in adaptive responses and genome x environment interactions in long-generation species such as pines. Chapter 3 provides a comprehensive review on the latest discoveries and future research perspectives for the study of transposable elements in plants, with an emphasis on pine species.

Rapid changes in the climate due to increases in temperature and altered precipitation regimes pose a significant challenge for natural and planted populations of pines. Our ability to predict future responses to environmental changes will only come from a thorough understanding of the genomic and transcriptomic basis of abiotic stress. In this book, associations between genotypes and environmental variables are tested across the *P. lambertiana* species' natural distribution (Chap. 4). Plastic responses to low water availability analyzed with transcriptome analysis for *P. pinaster* are reviewed in Chaps. 5 and 9.

Whole-genome sequencing and the development of a transcriptome atlas in *P. pinaster* are fully covered in Chap. 5. In addition, this chapter provides a comprehensive review of some of the most important research in *P. pinaster*, including recent findings in molecular breeding, transgenesis, comparative genomics, gene expression regulation, biotic and abiotic stress and genetic architecture and variation in the species. Perspectives about the impact of these recent discoveries and future research approaches are also discussed.

Most phenotypic traits of commercial importance in pines and plants in general have complex genomic architecture, meaning that a large number of genes are usually involved. Chapter 6 summarizes recent studies on complex traits in *P. taeda*, while Chaps. 7 and 8 focus on the genomics of disease resistance against fungal pathogens causing three major diseases in North American pine species: white pine blister rust, pitch canker and fusiform rust.

Transcriptomic approaches in European pines such as *P. pinaster* and *P. sylvestris* are covered in Chaps. 9 and 10. While Chap. 9 focuses on the transcriptomic, proteomics, metabolomic and genetic transformation used in functional genomic studies in *P. pinaster*; Chap. 10 focuses on the transcriptional and genomic responses to radiation stress in *P. sylvestris* populations from the Chernobyl exclusion zone after the Chernobyl nuclear power accident in the late '80s.

Given the economic importance of many widely distributed pine species, there is wide interest on improving breeding efficiency by shortening breeding cycles in long-generation pines. A big limitation, however, is the little knowledge about the genomics of complex traits in conifer species. The genomic selection was, therefore, developed as a potential solution. Genomic selection has the potential to shorten breeding cycles when compared with conventional (pedigree-based) breeding, reduce the cost of phenotyping and also does not require the identification of causal genes (as in marker-assisted breeding). Chapter 11 reviews recent research advances in genomic selection in *P. sylvestris* in Sweden and *P. radiata* in New Zealand.

Finally, the book concludes by discussing the future needs and applications in pine genomics by proposing a collaborative international advisory committee that organizes and prioritizes species to be sequenced and publicly accessed by the scientific community (Chap. 12). All the recent genomic and transcriptomic resources and studies described in this book have paved the way for understanding the complex biology of this very important group of plants and will help future management, conservation and breeding efforts.

Flagstaff, Arizona, USA
June 2021

Amanda R. De La Torre

Contents

Contributors

Sara Abrahamsson SKOGFORSK (The Forestry Research Institute of Sweden), Sävar, SE, Sweden

Ana Alves BioISI - Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, PT, Portugal

Ricardo Alía Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain

Isabel Arrillaga Biotechnology and Biomedicine (BiotecMed) Institute and Plant Biology Department, University of Valencia, Valencia, ES, Spain

Concepción Ávila Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

Laurent Bouffier INRAE, Univ. Bordeaux, BIOGECO, Cestas, FR, France

Jeremy T. Brawner Department of Plant Pathology, University of Florida, Gainesville, FL, USA

Katharina B. Budde Department of Forest Genetics and Forest Tree Breeding, Buesgen Institute, Georg-August University of Göttingen, Göttingen, Germany

José Antonio Cabezas Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Ainhoa Calleja-Rodríguez SKOGFORSK (The Forestry Research Institute of Sweden), Sävar, SE, Sweden

Francisco R. Cantón Dpto. Biología Molecular y Bioquímica. Facultad de Ciencias Campus de Teatinos S/N, Universidad de Málaga, Málaga, ES, Spain

Vanessa Castro-Rodríguez Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

Rafael A. Cañas Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, ES, Spain

María Teresa Cervera Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Francisco M. Cánovas Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, ES, Spain

John M. Davis School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL, USA

Amanda R. De La Torre School of Forestry, Northern Arizona University, Flagstaff, Arizona, USA

Fernando N. de la Torre Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, Spain

Nuria de María Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Marina de Miguel INRAE, Univ. Bordeaux, BIOGECO, Cestas, FR, France;
EGFV, Univ. Bordeaux, Bordeaux Sciences Agro, INRAE, ISVV, Villenave d’Ornon, France

Gustavo T. Duarte Max Plank Institute of Molecular Plant Physiology, Potsdam-Golm, Germany;
Belgian Nuclear Research Centre (SCK CEN), Biosphere Impact Studies, Mol, Belgium

Heidi Dungey Scion (New Zealand Forest Research Institute), Whakarewarewa, Rotorua, New Zealand

Carmen Díaz-Sala Departamento de Ciencias de la Vida (Fisiología Vegetal), Universidad de Alcalá, Alcalá de Henares, ES, Spain

Marta Callejas Díaz Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain

Jorge El-Azaz Dpto. Biología Molecular y Bioquímica. Facultad de Ciencias Campus de Teatinos S/N, Universidad de Málaga, Málaga, ES, Spain

Daniel Ence School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL, USA

Maria Rosario García-Gil Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Sciences, Umeå, SE, Sweden

Stanislav A. Geras'kin Russian Institute of Radiology and Agroecology, Obninsk, Russian Federation

David S. Gernandt Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

Santiago C. González-Martínez INRAE, Univ. Bordeaux, BIOGECO, Cestas, FR, France

Natalie Graham Scion (New Zealand Forest Research Institute), Whakarewarewa, Rotorua, New Zealand

Delphine Grivet Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain

María Ángeles Guevara Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Laura Hernández-Escribano Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain

Agathe Hurel INRAE, Univ. Bordeaux, BIOGECO, Cestas, FR, France

Ahmed Ismael Scion (New Zealand Forest Research Institute), Whakarewarewa, Rotorua, New Zealand

Jeremy S. Johnson Department of Environmental Studies, Prescott College, Prescott, Arizona, USA

Carmen Jurado-Mañogil Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain

Jaroslav Klápště Scion (New Zealand Forest Research Institute), Whakarewarewa, Rotorua, New Zealand

Jun-Jun Liu Canadian Forest Service, Natural Resources Canada, Victoria, BC, Canada

Carol A. Loopstra Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA

Mengmeng Lu Department of Biological Sciences, University of Calgary, Calgary, Canada

Miriam López-Hinojosa Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Juan Majada CETEMAS, Forest and Wood Technology Research Centre, Asturias, ES, Spain

Lorenzo Federico Manjarrez Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Pablo Martínez-García Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain

Pedro J. Martínez-García Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain

Jorge Mas-Gómez Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain

Isabel Mendoza-Poudereux Biotechnology and Biomedicine (BiotecMed) Institute and Plant Biology Department, University of Valencia, Valencia, ES, Spain

Célia Miguel iBET, Instituto de Biologia Experimental e Tecnológica, Oeiras, Portugal;
BioISI - Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, PT, Portugal

Inês Modesto Department of Plant Biotechnology and Bioinformatics and VIB Center for Plant Systems Biology, Ghent University, Ghent, Belgium;
ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, PT, Portugal;
BioISI - Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisboa, PT, Portugal

Marian Morcillo Biotechnology and Biomedicine (BiotecMed) Institute and Plant Biology Department, University of Valencia, Valencia, ES, Spain

David B. Neale University of California-Davis, Davis, CA, USA

C. Dana Nelson USDA Forest Service, Southern Research Station, Southern Institute of Forest Genetics, Saucier, MS, USA;
Forest Health Research and Education Center, Lexington, KY, USA

María Belén Pascual Facultad de Ciencias, Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Málaga, ES, Spain

Pedro Perdiguero iBET, Instituto de Biologia Experimental e Tecnológica, Oeiras, Portugal;
Centro de Investigación en Sanidad Animal (CISA-INIA), Madrid, ES, Spain

Gary F. Peter School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL, USA

Alberto Pizarro Departamento de Ciencias de la Vida (Fisiología Vegetal), Universidad de Alcalá, Alcalá de Henares, ES, Spain

Christophe Plomion INRAE, Univ. Bordeaux, BIOGECO, Cestas, FR, France

Tania Quesada School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL, USA

Annie Raffin INRAE, UEFP, Cestas, FR, France

Rosa Raposo Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain

Andreia S. Rodrigues ITQB NOVA, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, PT, Portugal;
Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, NL, Netherlands

Dainis E. Rungis Latvian State Forest Research Institute, Salaspils, Latvia

Jerome Salse INRAE, Univ. Clermont Auvergne, GDEC, Clermont-Ferrand, FR, France

Manoj K. Sekhwal School of Forestry, Northern Arizona University, Flagstaff, Arizona, USA

Richard A. Sniezko USDA Forest Service, Dorena Genetic Resource Center, Cottage Grove, Oregon, USA

Lieven Sterck Department of Plant Biotechnology and Bioinformatics and VIB Center for Plant Systems Biology, Ghent University, Ghent, Belgium

Kristian Stevens Department of Plant Pathology, University of California, Davis, CA, USA;
Department of Evolution and Ecology, University of California-Davis, Davis, CA, USA

Mari Suontama SKOGFORSK (The Forestry Research Institute of Sweden), Sävar, SE, Sweden

Leopoldo Sánchez INRAE, ONF, Orléans, BioForA FR, France

Jean-Francois Trontin BioForBois, FCBA Technological Institute, Wood & Construction Industry Dpt, Cestas, FR, France

Polina Y. Volkova Russian Institute of Radiology and Agroecology, Obninsk, Russian Federation

Angelika F. Voronova Latvian State Forest Research Institute, Salaspils, Latvia

Alejandra Vázquez-Lobo Centro de Investigación en Biodiversidad y Conservación, Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico

María Dolores Vélez Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR, CSIC), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid, ES, Spain;
Unidad Mixta de Genómica y Ecofisiología Forestal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)/Universidad Politécnica de Madrid (INIA/UPM), Madrid, ES, Spain

Matthew Weiss Department of Biology, Northern Arizona University, Flagstaff, Arizona, USA

Rafael Zas MBG-CSIC, Pontevedra, ES, Spain

Abbreviations

ABA	Abscisic Acid
AFLP	Amplified Fragment Length Polymorphism
AIF	Apoptosis-inducing factor
AUX	Auxin
BAC	Bacterial Artificial Chromosome
BRs	Brassinosteroids
cDNA	Complementary DNA
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CNV	Copy number variation
COS	Conserved Orthologous Sequences
CP	Candidate or validation Population
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTK	Cytokinin
DBH	Diameter at Breast Height
DEG	Differentially Expressed Gene
EAA	Environmental Association Analysis
EM	Expectation Maximization
eQTL	Expression quantitative trait locus
ESTP	Expressed Sequenced Tags Polymorphism
EST	Expressed Sequenced Tags
ET	Ethylene
ETI	Effector-Triggered Immunity
FA	Fluctuation Asymmetry
FT	Flowering locus T-like protein
GA	Gibberellin
GBLUP	Genomic-based Best Linear Unbiased Prediction
GBS	Genotyping by Sequencing
gDNA	Global DNA
GEA	Genome-wide Environmental Analysis
GEBV	Genomic Breeding Values
GMO	Genetically Modified Organism
GS	Genomic Selection
GWAS	Genome-wide Association Study
HM	Heavy Metal
HMW	High Molecular Weight
HR	Hypersensitive-like Response

HSPs	Heat-shock proteins
ICRP	International Commission on Radiation Protection
Indel	Insertion/Deletion
IRAP	Inter-retrotransposon amplified polymorphism
ISH	In situ Hybridization
IUCN	International Union for Conservation of Nature
iWUE	Intrinsic Water Use Efficiency
JA	Jasmonic Acid
LA	Linkage Analysis
LC	Liquid Chromatography
LCM	Laser Capture Microdissection
LD	Linkage Disequilibrium
LiDAR	Light Detection and Ranging
lncRNA	Long non-coding RNA
LOD	Logarithm of the Odds
LTR	Long Terminal Repeat
MAPK	Mitogen-activated Protein Kinase
MAS	Marker-Assisted Selection
MGR	Major Gene Resistance
miRNA	MicroRNA
MITE	Miniature Inverted Repeat Transposable Element
MLM	Mixed Linear Model
MNPs	Multiple Nucleotide Polymorphisms
MP	Mate pair
MS	Mass Spectrometry
MWAS	Metagenome-wide Association Study
NBS	Nucleotide Binding Site
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NIRs	Near-Infrared Spectroscopy
NMR	Nuclear Magnetic Resonance
NRM	Numerator Relationship Matrix
ONT	Oxford Nanopore Technology
ORF	Open Reading Frame
PAV	Presence/Absence variation
PBLUP	Pedigree-based Best Linear Unbiased Prediction
PCA	Principal Component Analysis
PCD	Programmed Cell Death
PE	Paired end
PEG	Polyethylene Glycol
PPC	Pine Pitch Canker
PR	Pathogenesis-related
PTI	Pattern-Triggered Immunity
PWN	Pine Wood Nematode
QDR	Quantitative Disease Resistance
QTL	Quantitative Trait Loci
QTN	Quantitative Trait Nucleotide

RAPD	Random Amplified Polymorphic DNA
RBIP	Retrotransposon-based Insertional Polymorphism
RDA	Redundancy Analysis
REMAP	Retrotransposon Microsatellite Amplified Polymorphism
RFLP	Restriction Fragment Length Polymorphism
RIVP	Retrotransposon Internal Variation Polymorphism
RNAi	RNA interference
ROS	Reactive Oxygen Species
RRM	Realized Relationship Matrix
SA	Salicylic Acid
SE	Somatic Embryogenesis
SLs	Strigolactones
SNP	Single Nucleotide Polymorphism
sRNA	Small RNA
SSAP	Sequence-Specific Amplification Polymorphism
ssBLUP	Single-Step Best Linear Unbiased Prediction
SSR	Simple Sequence Repeat
SUP	Single-Uredinial Pustule
TE	Transposable Element
TF	Transcription Factor
TIR	Terminal Inverted Repeat
TP	Training Population
TSDs	Target Site Duplications
VOCs	Volatile Organic Compounds
WGD	Whole-genome Duplication
WPBR	White Pine Blister Rust
WUE	Water use efficiency

1 Advances in the Genomic and Transcriptomic Sequencing of North American Pines

Alejandra Vázquez-Lobo,
David S. Gernandt, Pedro J. Martínez-García,
and Amanda R. De La Torre

Abstract

Genetic and evolutionary questions are being addressed in pines using a host of high-throughput sequencing strategies, including whole-genome sequencing, transcriptome sequencing, and target enrichment of nuclear genes. Some of the questions being addressed include the genetic basis of pathogen and drought resistance, differential expression, genetic mapping, phylogeography, and phylogenetics. Pine genomes are enormous, ranging from 20 to 40 Gb. At present, draft genomes are available for only two pine species, *P. taeda* (loblolly pine) and *P. lambertiana* (sugar pine), but most other approximately 80 species of North American pines have been represented in evolutionary studies based on complete plastomes, low-copy nuclear genes, and transcriptomes. A number of online databases have been developed and made publicly available for comparative studies of pines and other conifers.

1.1 Nuclear Genomes

Pine species are naturally distributed in the Northern Hemisphere and are also planted as commercial species in the Southern Hemisphere. Due to their importance for commercial forestry, they are considered the world's most economically important forest species. Efforts to understand their complex biology and evolution were limited by the absence of reference genomes. Pines, as other conifers, are slow-growing, long-lived species and possess enormous genomes (20–40 Gb) with a high number of repeat elements (Wegrzyn et al. 2014). Limitations of short-read sequencing technologies, computational power, and assembly software made the sequencing of pine genomes a daunting task just 10 years ago (De La Torre et al. 2014, 2019). To date, only two North American pine species have been sequenced: *Pinus taeda* and *Pinus lambertiana*.

The first sequenced genome was *Pinus taeda* (loblolly pine) in 2014, the most planted forest

A. Vázquez-Lobo (✉)
Centro de Investigación en Biodiversidad y Conservación, Universidad Autónoma del Estado de Morelos, 62209 Cuernavaca, Morelos, Mexico
e-mail: alejandra.vazquez@uaem.mx

D. S. Gernandt
Departamento de Botánica, Instituto de Biología, Universidad Nacional Autónoma de México, 04510 Ciudad de México, Mexico
e-mail: dgernandt@ib.unam.mx

P. J. Martínez-García
Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain
e-mail: pjmartinez@cebas.csic.es

A. R. De La Torre
Northern Arizona University, 200 E Pine Knoll Dr, Flagstaff, AZ AZ86011, USA
e-mail: Amanda.de-la-torre@nau.edu

A. R. De La Torre (ed.), *The Pine Genomes*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-030-93390-6_1

tree in North America and a key species for commercial forestry in the southwestern United States (McKeand et al. 2021). Whole-genome shotgun sequencing was used to analyze the 22 Gb genome and develop the first two genome assemblies (versions 1.0 and 1.01) of the *P. taeda* genome (Wegrzyn et al. 2014; Zimin et al. 2014). Version 1.0 was based on the MaSuRCA (Zimin et al. 2013) assembly of paired-end reads from a haploid female gametophyte (or megagametophyte) and long insert linking read pairs ("super-reads") from diploid needle tissue (Zimin et al. 2014). This version resulted in a draft genome sequence of 20.15 Gb (spanning 23.2 Gbp) with an N50 scaffold size of 66.9 kbp (Zimin et al. 2014), from which 82% was composed of repetitive elements (Neale et al. 2014). Version 1.01 employed scaffolding from independent genome and transcriptome assemblies (Wegrzyn et al. 2014). Structural annotation identified 50,172 gene models with long intron lengths varying from 2.7 to 100 kbp (Neale et al. 2014). Current version available from the TreeGenes database (see below) is version 2.01, which contains 51,751 protein sequences (file Pita.2_01.pep.fa, last accessed in June 2021).

Another economic and ecologically important pine species in North America is sugar pine (*Pinus lambertiana*). *P. lambertiana* populations are severely affected by the exotic fungal pathogen *Cronartium ribicola* (white pine blister rust, WPBR) throughout their natural range (Weiss et al. 2020). Therefore, interest in identifying the genes coding resistance to the disease was a significant motivation to decode its genome. The sequencing and assembly of *P. lambertiana* followed a similar procedure to that used in *P. taeda* (in fact, both species were sequenced by the same group of researchers at University of California-Davis, US). Paired-end libraries for short-read Illumina sequencing were constructed from haploid megagametophyte tissue, error-corrected and later used to construct "super-reads" with MaSuRCA 2.3.0 (Zimin et al. 2013). Mate pairs from diploid tissue libraries were cleaned and filtered and added to the haploid data for genome assembly with SOAPdenovo2 (Luo et al. 2012). Newly developed Pacific Biosciences (PacBio) and Illumina RNA-seq data were used for additional scaffolding steps (Gonzalez-Ibeas et al. 2016). The total length of assembly version 1.0 including all scaffolds and contigs >200 bp was 27.6 Gbp from a 31 Gbp estimated genome size (Stevens et al. 2016). An important contribution of this assembly was the identification of candidate genes for *Cr1* (major gene for WPBR resistance) that could significantly contribute to marker-assisted breeding efforts (Stevens et al. 2016). In a more recent study, long-reads from 10X Genomics (www.10Xgenomics.com) were used to build and improve the assembly, generating an eightfold improvement over the original NG50 scaffold of 247 kb (Crepeau et al. 2017).

1.2 Plastid Genomes and Target Gene Sequencing

An important aim in the generation of bioinformatic resources for pines has been the search for markers for evolutionary studies. Nuclear genomes of pines are characterized by their high levels of gene duplication and high frequency of repeat regions, making it difficult to identify useful nuclear markers for phylogenetic and population genetic analyses, which usually assume orthology relationships of genetic variants. Plant mitochondrial genomes have low substitution rates and relatively high rates of rearrangement and transfer to and from the nucleus. Until recently most DNA-based phylogenetic and population studies in *Pinus* have been based on plastid markers. To take into account coalescent processes, phylogenetic studies of low-copy nuclear genes were adopted (e.g., Syring et al. 2007; Willyard et al. 2007; DeGiorgio et al. 2014).

The first fully sequenced plastid genome of a gymnosperm was that of the hard pine (subgenus *Pinus*) *Pinus thunbergii* (Wakasugi et al. 1994). The plastome of *P. koraiensis*, a soft pine (subgenus Strobus), was made available in public sequence databases a few years later. Adoption of short-read sequencing accelerated the pace of plastome sequencing until presently complete or nearly complete genomes are available for more than 100 pine species, including all species from North America. Plastome sequencing has allowed inference of phylogenies with well-resolved relationships; however, there is clearly discordance between organellar gene trees and nuclear gene trees of pines (e.g., Wang and Wang 2014; Gernandt et al. 2018), and some relationships among species remain uncertain. Due to the size of the pine genome, alternative approaches to whole-genome sequencing were needed. Therefore, efforts were made to identify suitable low-copy nuclear regions for evolutionary studies.

Target enrichment has been used to acquire sequences for most (Neves et al. 2013) or a fraction (Gernandt et al. 2018) of putative low-copy nuclear genes in pines for studies ranging from genetic mapping to phylogenetics. Biotinylated RNA oligonucleotides are used to enrich genomic libraries for specific regions of interest such as exomes (the entire protein coding fraction of the genome), which can then be characterized with massively parallel sequencing (Gnirke et al. 2009). Target enrichment can be combined with genome skimming to also include in the same sequencing runs the high-copy fraction of genomes, particularly nuclear ribosomal DNA and complete plastomes, a strategy called Hyb-Seq (Weitemier et al. 2014). Hyb-Seq can be performed on degraded or historical samples and has a low per-sample price (Hale et al. 2020). Because the method includes flanking sequences of targeted genes or exons, it provides additional information on the history of markers that may have undergone gene duplication or loss. This method has been used to characterize hundreds of genes for phylogenetic studies of the three most species-rich clades of exclusively North American pines, subsects. *Australes*, *Ponderosae*, and *Cembroides* (Gernandt et al. 2018; Montes et al. 2019; Willyard et al. 2021) and to study population genetics and local adaptation (Peláez et al. 2020).

Can evolutionary questions be addressed better by characterizing hundreds or a few thousand nuclear genes and complete plastomes with Hyb-Seq or by characterizing transcriptomes? It has been argued that genes and data derived from genes, in particular, transcriptomes should not be analyzed with coalescence methods because of the likelihood that they have undergone recombination. This is particularly the case for those genes that are divided into exons dispersed more broadly across the genome and represent unlinked/independent estimates of gene tree relationships (Springer and Gatesy 2016).

1.3 Transcriptomic Resources

The characterization of transcriptomes is a basic tool for the annotation of reference genomes, so for the annotation of the *P. taeda* reference genome, transcriptomes of different tissues at different stages of development for the species were generated. This allowed the identification of more than 80,000 transcripts, of which about 45,000 genes were successfully mapped (Wegrzyn et al. 2014). Similarly, for *P. lambertiana*, by characterizing a reference transcriptome from different tissues and taking advantage of different sequencing platforms, close to 30,000 transcripts have been functionally annotated (Gonzalez-Ibeas et al. 2016). As mentioned above, the sugar pine is threatened by the WPBR, as are its white bark pine relatives with similar distributions. Through a comparative analysis of transcriptomes of the western white pine (*P. monticola*), limber pine (*P. flexilis*), white bark pine (*P. albicaulis*), and sugar pine (*P. lambertiana*), signals of positive selection were found in different genes, including candidates to WPBR resistance (Baker et al. 2018). Transcriptome sequences for loblolly pine and sugar pine are available in the TreeGenes database (see below).

Identification of the genetic basis of processes and phenotypes in pines requires a more detailed analysis, considering biotic or abiotic factors, and comparing tissues and species. For example, characterization of transcriptomes of *P. patula* and *P. tecunumanii* from plantations in South Africa (species from Mexico and Central America) using tissues infected with a fungus (*Fusarium circinatum*) and differential expression analyses has allowed the identification of genes involved in the response to pathogens in these species (Visser et al. 2019, 2018, 2015). Similarly, through differential expression analysis (DE) specific genes have been identified for the response to a fungal infection (*Dothistroma septosporum*) in *P. contorta* (Lu et al. 2021). Through DE transcriptomic analyses, genes involved in wood maturation in *P. radiata* (Li et al. 2011) and in resin tapping in *P. elliotti* (de Oliveira Junkes et al. 2019) have also been identified.

The development of bioinformatic resources for pines of arid environments is of special interest, since in many cases these species are the only forest resource in said environments. For species of arid climates in Mexico, investigations have been carried out on *P. pinceana*, through the characterization of transcriptomes of individuals from different populations in the range of distribution of the species (Figueroa-Corona et al. 2021) and for *P. cembroides*, a differential expression analysis has been carried out to identify the changes in gene expression in juvenile and adult leaves (Webster et al. in prep).

While low-copy genes are informative for evolutionary inferences, identification of functional genes requires characterization of the transcriptome and detection of differentially expressed genes. Recent advances have been made in this regard, by obtaining and characterizing new transcriptomes for 107 pine species from megametophytes or young needles for an evolutionary study of the genus *Pinus* (Jin et al. 2021). This study included 66 species distributed in America, of which 31 are mainly distributed in the United States and Canada and 35 are from Mexico, the Caribbean, and Central America.

1.4 Databases for Genomic and Transcriptomic Resources

1.4.1 Plaza

The database and online resource PLAZA (http://bioinformatics.psb.ugent.be/plaza/) version 4.0 were created to allow comparative, evolutionary, and functional genomic analyses among plant species through a user-friendly web interface (Van Bel et al. 2018). PLAZA allows users to browse genomes, gene families, and phylogenetic trees; to find functional information through BLAST; and to explore genome organization through different visualization tools (e.g., Ks graphs, Skyline plots, WGDotplot) based on gene collinearity or synteny information (Proost et al. 2015). PLAZA Gymnosperms includes structural and functional annotation of 16 gymnosperm species, including 777,165 genes clustered in 30,041 multi-species gene families (last accessed in June 2021). In the case of gymnosperm species lacking reference genome sequences, PLAZA uses curated transcriptomic data to identify genes and gene families and make them available for comparative genomics analyses. To date, PLAZA contains data on only three Pinus species: *Pinus taeda*, *Pinus sylvestris,* and *Pinus pinaster* (last accessed in June 2021).

1.4.2 TreeGenes

The Dendrome project and the associated TreeGenes database (https://treegenesdb.org) were created in the early 1990s as a repository to store genetic linkage and Expressed Sequence Tags (ESTs) data with a focus on commercial Pinaceae species (Wegrzyn et al. 2008, 2019; Falk et al. 2018). Over the years, TreeGenes expanded to include curated data in addition to data provided by users (Falk et al. 2018). To accommodate the needs of larger datasets as a result of the high-throughput sequencing, TreeGenes incorporated more efficient models for data storage and later moved to the Tripal framework, a more flexible, efficient, and sustainable platform (Falk et al. 2018). The Tripal Gateway framework supports cross-site query, data transfer, access to analytical pipelines (e.g., Galaxy), and different modules such as Tripal Plant PopGen Submit (TPPS), Tripal Sequence Similarity Search (TSeq), and OrthoQuery (Falk et al. 2018). The TSeq module allows sequence similarity search against genes, TreeGenes UniGenes, proteins, and full genome through NCBI BLASTX, BLASTN, or BLASTP. Genetic, phenotypic, and/or environmental data submission from users can be