Population Genomics
Editor-in-Chief: Om P. Rajora

Om P. Rajora *Editor*

Population Genomics: Crop Plants

Springer

Population Genomics

Editor-in-Chief

Om P. Rajora, Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada

This pioneering Population Genomics Series deals with the concepts and approaches of population genomics and their applications in addressing fundamental and applied topics in a wide variety of organisms. Population genomics is a fast emerging discipline, which has created a paradigm shift in many fields of life and medical sciences, including population biology, ecology, evolution, conservation, agriculture, horticulture, forestry, fisheries, human health and medicine.

Population genomics has revolutionized various disciplines of biology including population, evolutionary, ecological and conservation genetics, plant and animal breeding, human health, genetic medicine, and pharmacology by allowing to address novel and long-standing intractable questions with unprecedented power and accuracy. It employs large-scale or genome-wide genetic information across individuals and populations and bioinformatics, and provides a comprehensive genome-wide perspective and new insights that were not possible before.

Population genomics has provided novel conceptual approaches, and is tremendously advancing our understanding the roles of evolutionary processes, such as mutation, genetic drift, gene flow, and natural selection, in shaping up genetic variation at individual loci and across the genome and populations, disentangling the locus-specific effects from the genome-wide effects, detecting and localizing the functional genomic elements, improving the assessment of population genetic parameters or processes such as adaptive evolution, effective population size, gene flow, admixture, inbreeding and outbreeding depression, demography, and biogeography, and resolving evolutionary histories and phylogenetic relationships of extant and extinct species. Population genomics research is also providing key insights into the genomic basis of fitness, local adaptation, ecological and climate acclimation and adaptation, speciation, complex ecologically and economically important traits, and disease and insect resistance in plants, animals and/or humans. In fact, population genomics research has enabled the identification of genes and genetic variants associated with many disease conditions in humans, and it is facilitating genetic medicine and pharmacology. Furthermore, application of population genomics concepts and approaches facilitates plant and animal breeding, forensics, delineation of conservation genetic units, understanding evolutionary and genetic impacts of resource management practices and climate and environmental change, and conservation and sustainable management of plant and animal genetic resources.

The volume editors in this Series have been carefully selected and topics written by leading scholars from around the world.

Om P. Rajora
Editor

Population Genomics: Crop Plants

Springer

Editor
Om P. Rajora
Faculty of Forestry and Environmental
Management
University of New Brunswick
Fredericton, NB, Canada

ISSN 2364-6764 ISSN 2364-6772 (electronic)
Population Genomics
ISBN 978-3-031-63001-9 ISBN 978-3-031-63002-6 (eBook)
https://doi.org/10.1007/978-3-031-63002-6

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

Population Genomics Book Series

This Population Genomics book series is dedicated to my (late) parents, and my wife Malti and children Apoorva, Anu, and Maneesha.

Om P. Rajora

Population Genomics: Crop Plants

This book is respectfully dedicated to my (late) father, Shri Than Singh, who was a prominent farmer, a prominent person, and a leader, well known for his honesty, integrity and generosity.

Om P. Rajora

Preface

Population genomics has revolutionized various disciplines of plant biology, especially population, evolutionary, ecological and conservation genetics, plant breeding, and conservation and sustainable management of plant genetic resources by allowing to address novel and long-standing intractable questions with unprecedented power and accuracy. Population genomics provides comprehensive genome-wide information and new insights that were not possible earlier.

Population genomics concepts and approaches along with bioinformatics tools and models have created a paradigm shift in several disciplines of crop plants biology by making significant, unprecedented advances in both basic and applied research. Crop plants have been domesticated from their wild progenitors over several centuries and have undergone severe genetic bottlenecks and selection sweeps through selection and breeding. This has resulted in a narrow genetic base of crop plants. Sustainability and enhancement of crop production is essential for food security to meet increasing demands of increasing world population. Climate change is also posing challenges to sustainability of crop plants and crop production. Sustainable crop production can be facilitated by conservation and enhancement of plant genetic resources, greater and accelerated genetic improvement of crop plants, genetic diversity enrichment in their pre-breeding and breeding programs, and improved ecological, environmental, and climate adaptation of crop plants. Genetic diversity provides the raw material for evolution and adaptation of organisms, especially under changing environmental and disease conditions. Therefore, a deeper and precise understanding of genetic/genomic diversity and structure of wild and domesticated populations and species, origin, evolution, demographic history, center of diversity, domestication history, genetic/genomic basis of domestication syndrome, genomic footprints of domestication, selection and breeding, interspecific and intraspecific phylogenetic relationships, introgression from wild species, genes and genomic regions underlying traits of interest (including productivity, biotic and abiotic stress tolerance), and ecological and climate adaptation of crop plants is required. Population genomics concepts and approaches are unraveling key, novel, and deeper insights into these above aspects of crop plants with

unprecedented power and accuracy. In addition, population genomics approaches are revealing key insights into speciation, systematics, and de-domestication and allowing to construct pangenomes. Population genomics has also facilitated identifying genotype-phenotype associations and building prediction models for estimating genetic/breeding value of individuals, thus facilitating genomics-assisted early selection and accelerated breeding of crop plants. Moreover, population epigenomics and population transcriptomics studies have begun in crop plants which can contribute to enhancing our understanding of acclimation, adaptation, and disease and insect resistance of populations.

As a part of the pioneering Population Genomics book series, this book discusses the progress and perspectives of population genomics in addressing various fundamental and applied crucial aspects in crop plants. The book provides insights into a range of emerging topics in crop plants including pangenomes, genomic diversity and population structure, demography, evolution, domestication, de-domestication, speciation, taxonomy, phylogenomics, molecular breeding, population epigenomics, population transcriptomics, biotic and abiotic stress tolerance, ecological and climate adaptation, gene banks genomics, and conservation of plant genetic resources. The chapters are written by leading and emerging research scholars in crop plants population genomics.

The book has 22 chapters which are organized into three parts. The first part has six chapters which discuss population genomics aspects of pangenomes, organellar genomes, demographic history, evolution, speciation, adaptation, gene banks, and genetic resource conservation in crop plants. The second part includes four chapters which discuss the application of population genomics concepts and approaches in enhancing crop plants breeding and valorization of genetic diversity in breeding and pre-breeding programs. The third part consists of 12 chapters covering the progress and prospects of population genomics research and application in major crop plants. Each chapter focuses on a crop plant and first provides a review of genomic, transcriptomic, epigenomic, and plant resources available for population genomics studies and then discusses the progress made in population genomics aspects including pangenome, genetic diversity and population structure, origin, evolution, domestication, speciation and admixture, phylogenomics, genome-wide association studies, genomic selection, population epigenomics, population transcriptomics, and conservation and sustainable management of plant genetic resources. Finally, future perspectives are discussed.

The book is envisioned for a wide readership, including undergraduate and graduate students, research scholars, and professionals and experts in the field. It fills a vacuum in the field and is expected to become a primary reference in population genomics of crop plants worldwide.

I thank all the authors who have contributed to this volume.

Fredericton, NB, Canada Om P. Rajora

Contents

Contributors

Akira Abe Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Patrick Adebola International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

Ryan J. Andres USDA-ARS Plant Science Research Unit, Raleigh, NC, USA
Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, USA

Adriana Maria Antunes Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brazil

Stéphanie Arnoux INRAE, GAFL, Avignon, France

Asrat Asfaw International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

Robert Asiedu International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria

Arthur Bernardeli Universidade Federal de Viçosa (UFV), Viçosa, MG, Brazil
University of Nebraska-Lincoln (UNL), Lincoln, NE, USA

Cécile Berthouly-Salazar Laboratoire mixte international Adaptation des Plantes et microorganismes associés aux Stress Environnementaux (LAPSE), Dakar, Senegal
DIADE, Université Montpellier, Institut de Recherche pour le Développement, Montpellier, France

Ramesh S. Bhat Department of Biotechnology, University of Agricultural Sciences, Dharwad, India

Pranita Bhatele M. H. College of Home Science & Science for Women, Autonomous, Jabalpur, Madhya Pradesh, India

Dan G. Bock Department of Biology, Washington University in St. Louis, St. Louis, MO, USA

Aluízio Borém Universidade Federal de Viçosa (UFV), Viçosa, MG, Brazil

Sophie Bouchet INRAE, Université Clermont Auvergne, Clermont-Ferrand, France

Concetta Burgarella Université de Montpellier, Institut pour la Recherche et le Développement, Unité Diversité, Adaptation, Développement des plantes, Montpellier, France
Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

Mathilde Causse INRAE, GAFL, Avignon, France

Subhash Chandra ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Peter Civan GDEC, INRAE, Université Clermont Auvergne, Clermont-Ferrand, France

Rafael Augusto da Costa Parrella Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Philippe Cubry Université de Montpellier, Institut pour la Recherche et le Développement, Unité Diversité, Adaptation, Développement des plantes, Montpellier, France

Cynthia Maria Borges Damasceno Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Alice Danguy-Des-Deserts GDEC, Université Clermont Auvergne, Clermont-Ferrand, France
GenPhySE (Génétique Physiologie et Systèmes d'Elevage), Université de Toulouse, INRAE, ENVT, Castanet-Tolosan, France

Jeffrey C. Dunne Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, USA

Sudip K. Dutta Gus R. Douglass Institute, Department of Biology, West Virginia State University, Institute, WV, USA

Jean-Michel Elsen GenPhySE (Génétique Physiologie et Systèmes d'Elevage), Université de Toulouse, INRAE, ENVT, Castanet-Tolosan, France

Yonghai Fan College of Agronomy and Biotechnology, Southwest University, Chongqing, China

Ailton José Crispim Filho Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brazil

Olivier François Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble, France

Paul Gepts Department of Plant Sciences/MS1, Section of Crop and Ecosystem Sciences, University of California, Davis, CA, USA

Kaushik Ghose Department of Plant and Soil Science, Texas Tech University, Lubbock, TX, USA

M. V. Channabyre Gowda Department of Genetics and Plant Breeding, University of Agricultural Sciences, Dharwad, India

Sanjay Gupta ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Hideki Hirakawa Facility for Genome Informatics, Kazusa DNA Research Institute, Chiba, Japan

Ashlesha Holkar Biotechnology Department, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India
Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

James B. Holland USDA-ARS Plant Science Research Unit, Raleigh, NC, USA
Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, USA

Sachiko N. Isobe Department of Frontier Research and Development, Kazusa DNA Research Institute, Chiba, Japan

Ndjido Ardo Kane Institut Sénégalais de Recherches Agricoles (ISRA), Centre d'Étude Régional pour l'amélioration de l'Adaptation à la Sécheresse, Thiès, Senegal
Laboratoire mixte international Adaptation des Plantes et microorganismes associés aux Stress Environnementaux (LAPSE), Dakar, Senegal

Michael B. Kantar Department of Tropical Plant and Soil Sciences, University of Hawaii, Honolulu, HI, USA

Rucha Kavishwar ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Aoi Kudoh Laboratory of Crop Evolution, Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Giriraj Kumawat ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Chikara Kuwata Chiba Prefectural Agriculture and Forestry Research Center, Chiba, Japan

Amnon Levi USDA, ARS, U.S. Vegetable Laboratory, Charleston, SC, USA

Jiana Li College of Agronomy and Biotechnology, Southwest University, Chongqing, China
Academy of Agricultural Sciences, Southwest University, Chongqing, China
Engineering Research Center of South Upland Agriculture, Ministry of Education, Chongqing, China

Lin-Feng Li Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, and Coastal Ecosystems Research Station of the Yangtze River Estuary, School of Life Sciences, Fudan University, Shanghai, China

Shengting Li College of Agronomy and Biotechnology, Southwest University, Chongqing, China

Xiaodong Li College of Agronomy and Biotechnology, Southwest University, Chongqing, China

Carlos Lopez-Ortiz Gus R. Douglass Institute, Department of Biology, West Virginia State University, Institute, WV, USA

Kun Lu College of Agronomy and Biotechnology, Southwest University, Chongqing, China
Academy of Agricultural Sciences, Southwest University, Chongqing, China
Engineering Research Center of South Upland Agriculture, Ministry of Education, Chongqing, China

Jurandir Vieira de Magalhães Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Shivakumar Maranna ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Tony Maritim Biotechnology Department, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India
Tea Breeding and Genetic Improvement Division, KALRO-Tea Research Institute, Kericho, Kenya
Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

Lavanya Mendu Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT, USA

Venugopal Mendu Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT, USA

Xiaoxi Meng Department of Horticultural Science, University of Minnesota, St. Paul, MN, USA

Cécile Monat DIADE, University of Montpellier, IRD, Montpellier, France
South Green Bioinformatics Platform, IRD, Bioversity, CIRAD, INRAE, Montpellier, France
Université Clermont Auvergne, INRAE, GDEC, Clermont-Ferrand, France

Vennampally Nataraj ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Purushothaman Natarajan Gus R. Douglass Institute, Department of Biology, West Virginia State University, Institute, WV, USA

Satoshi Natsume Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Padma Nimmakayala Gus R. Douglass Institute, Department of Biology, West Virginia State University, Institute, WV, USA

Yue Niu College of Agronomy and Biotechnology, Southwest University, Chongqing, China

Muluneh Tamiru Oli Iwate Biotechnology Research Center, Kitakami, Iwate, Japan
Department of Animal, Plant, and Soil Sciences, School of Life Sciences, La Trobe University, Melbourne, VIC, Australia

Amanda Avelar de Oliveira Escola Superior de Agricultura Luiz de Queiroz/ Universidade de São Paulo (ESALQ/USP), Piracicaba, Brazil

Bruna Mendes de Oliveira Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brazil

Renato Gonçalves de Oliveira Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brazil

Janaína de Oliveira Melo Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Diamantina, MG, Brazil

Kenneth M. Olsen Department of Biology, Washington University in St. Louis, St. Louis, MO, USA

Manish K. Pandey Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India

Travis A. Parker Department of Plant Sciences/MS1, Section of Crop and Ecosystem Sciences, University of California, Davis, CA, USA

Maria Marta Pastina Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Shri Hari Prasad Centre for Plant Biotechnology and Molecular Biology, Kerala Agricultural University, Thrissur, Kerala, India

Cunmin Qu College of Agronomy and Biotechnology, Southwest University, Chongqing, China
Academy of Agricultural Sciences, Southwest University, Chongqing, China
Engineering Research Center of South Upland Agriculture, Ministry of Education, Chongqing, China

Rishiraj Raghuvanshi ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Om P. Rajora Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, NB, Canada

Milind B. Ratnaparkhe ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Supriya M. Ratnaparkhe Indore Biotech Inputs and Research (P) Ltd., Indore, Madhya Pradesh, India

Umesh K. Reddy Gus R. Douglass Institute, Department of Biology, West Virginia State University, Institute, WV, USA

Marcela Pedroso Mendes Resende Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brazil

Chris Richards USDA National Laboratory for Genetic Resources Preservation, Fort Collins, CO, USA

Loren H. Rieseberg Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

Renaud Rincent GQE-Le Moulon, INRAE, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, Clermont-Ferrand, France

François Sabot DIADE, University of Montpellier, IRD, Montpellier, France
South Green Bioinformatics Platform, IRD, Bioversity, CIRAD, INRAE, Montpellier, France

Muhammet Şakiroğlu Department of Bioengineering, Adana Alparslan Türkeş Science and Technology University, Adana, Turkey

Luis Fernando Samayoa Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, USA

Gyanesh K. Satpute ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Christopher Sauvage INRAE, GAFL, Avignon, France

Nora Scarcelli DIADE, Univ Montpellier, IRD, Montpellier, France

Robert Eugene Schaffert Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Romit Seth Biotechnology Department, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India
Plants for Human Health Institute – North Carolina State University, Kannapolis, NC, USA

Laura M. Shannon Department of Horticultural Science, University of Minnesota, St. Paul, MN, USA

Ram Kumar Sharma Biotechnology Department, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India
Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

Vinay Sharma Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India

Motoki Shimizu Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Kenta Shirasawa Department of Frontier Research and Development, Kazusa DNA Research Institute, Chiba, Japan

Maria Lúcia Ferreira Simeone Embrapa Milho e Sorgo, Sete Lagoas, MG, Brazil

Ajay K. Singh ICAR-National Institute of Abiotic Stress Management, Khurd, Baramati, Maharashtra, India

Kunwar Harendra Singh ICAR-Indian Institute of Soybean Research, Indore, Madhya Pradesh, India

Vander Fillipe de Souza Universidade Federal de Lavras (UFLA), Lavras, MG, Brazil

Subulakshmi Subramanian Amrita School of Biotechnology, Amrita University, Kollam, Kerala, India

Yu Sugihara Laboratory of Crop Evolution, Graduate School of Agriculture, Kyoto University, Kyoto, Japan

Prashant Suravajhala Amrita School of Biotechnology, Amrita University, Kollam, Kerala, India

Hiroki Takagi Iwate Biotechnology Research Center, Kitakami, Iwate, Japan
Ishikawa Prefectural University, Nonoichi, Ishikawa, Japan

Ryohei Terauchi Laboratory of Crop Evolution, Graduate School of Agriculture, Kyoto University, Kyoto, Japan
Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Heather Tuttle Department of Horticultural Science, University of Minnesota, St. Paul, MN, USA

Rajeev K. Varshney Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India

Yves Vigouroux Université de Montpellier, Institut pour la Recherche et le Développement, Unité Diversité, Adaptation, Développement des plantes, Montpellier, France

Dalia Vishnudasan Amrita School of Biotechnology, Amrita University, Kollam, Kerala, India

Part I
Pangenomes, Demography, Evolution, Speciation, and Conservation

Pangenomics in Crop Plants

Cécile Monat and François Sabot

Abstract Identifying diversity at many scales is the keystone to understand how genomes evolved and plants are so adaptable, even if they are under extreme constraint. Understanding the processes that create diversity and how the genome manages this variability is of high importance to be able to face today's challenges for breeding and genetic resource conservation of crop plants. For about past 15 years, we ushered into the era of pangenomics and started to learn what the true intra-species genomic diversity is and how it shapes the structure and the expression of the genome, especially in crop plants. The pangenome is the complete repertoire of sequences for a given population (often a species). It can be divided in compartments, the first one is the core genome which contains the sequences shared by the whole population. The second compartment is the dispensable genome which regroups the sequences shared by some individuals of the population but not all of them. As a sub-part of the dispensable genome, an individual-specific-genome regroups the sequences that are uniquely found in one individual of the population. By comparing individual gene and sequence contents within a population, pangenomics is a great tool to investigate how sequences evolve within a species and how genetic diversity is shaped within it. More and more studies are including relative wild species to have a larger overview of how genetic and phenotypic diversity was shaped through domestication and selection processes. In this chapter, we will present this concept of pangenomics and the associated methods and progress made, particularly in crop plants and future perspective.

C. Monat
DIADE, University of Montpellier, IRD, Montpellier, France

South Green Bioinformatics Platform, IRD, Bioversity, CIRAD, INRAE, Montpellier, France

Université Clermont Auvergne, INRAE, GDEC, Clermont-Ferrand, France

F. Sabot (✉)
DIADE, University of Montpellier, IRD, Montpellier, France

South Green Bioinformatics Platform, IRD, Bioversity, CIRAD, INRAE, Montpellier, France
e-mail: Francois.sabot@ird.fr

Om P. Rajora (ed.), *Population Genomics: Crop Plants*,
Population Genomics [Om P. Rajora (Editor-in-Chief)],
https://doi.org/10.1007/13836_2020_84,

Keywords Copy number variations · Crops · Genomic diversity · Pangenome · Pangenomics · Plants · Presence/absence variations · Structural variations

1 Introduction

Understanding genetic diversity within a species is required for various population, evolutionary and conservation genetic studies, and this knowledge is essential to understand species' phenotypic variation, as well as mechanisms allowing environmental adaptation (Hirsch et al. 2014; Carlos Guimaraes et al. 2015). However, estimation of genetic diversity is based on the data, methods, and technologies available at the time of analysis. Thus, while Carl Von Linné or Georges-Louis de Buffon (Lawrence Farber 2000) classified animals and plants based on visual phenotypic observations, nowadays biologists deal with large molecular dataset, and diversity estimations are based on genetics and genomics approaches and dataset.

Recent genetic diversity studies are mostly performed using SNP (single nucleotide polymorphism) markers (Springer et al. 2009; Hirsch et al. 2014) detected through sequencing approaches, and the allelic diversity in coding and non-coding sequences is generally thought to be the main cause of phenotypic variations. Up to recently it was widely assumed that individuals from the same species have a similar genetic constitution (Swanson-Wagner et al. 2010) and differ only by these (few) SNPs. However, since the rise of NGS (next generation sequencing) and of the subsequent massive sequencing efforts, it appeared that structural variation (SV), in the form of not only insertion/deletions (InDel) ranging from few bp and up to tens of kb or even megabase-sized, but also chromosomal rearrangements (translocations, inversions, duplications), is much more important than expected (Beyter et al. 2019) (Fig. 1). These variations might be due to CNV (copy number variation) of the same sequence, or even missing sequences in some individuals (PAV, presence/absence variations) (Da Silva et al. 2013; Montenegro et al. 2017). The different levels of diversity (SNP, SV, CNV, and PAV) add a challenge in the accurate representation of the genome. It is now clear that we need more assembled individual genomes (and more variation information) to capture the whole- genetic/genomic diversity of a given species. Thus, we need to have a wider overview of the genomic diversity by sequencing more than one genome per species; for that, approaches

Fig. 1 Schematic representation of no structural variation (no SV), copy number variation (CNV), presence/absence variation (PAV), and example of Inversion (Inv.) as a structural variation between two sequences. Each colored bloc represents a genomic segment

using pangenomes have been deployed (Gan et al. 2011; Li et al. 2014; Hirsch et al. 2014; Schatz et al. 2014a).

The purpose of this chapter is to present an overview of what is pangenomics as applied to crop plants and why it is important for the future. We will first present the pangenome concept, how it was initially developed in the microorganisms world and then deployed in other biological domains. We will focus afterward on pangenomics in plants and more precisely on crop plants. Finally, we will briefly present the methodologies to create and analyze pangenomes.

2 Pangenome Concept

The access to DNA sequences, *i.e.* to the whole genetic and genomic information, has not evolved at the same pace for different phyla, mainly because of the disparity of genome size and complexity between them. For instance, basal eukaryotes (ex: *Saccharomyces cerevisiae*, 13 Mb) and bacteria (ex: *Escherichia coli*, 4.6 Mb) mostly have small simple genomes, easy to sequence and assemble. On the other hand, plants have large (Gb sized or more) and complex genomes; thus, only the true first pangenomic studies were performed on small bacterial genomes (Tettelin et al. 2005, 2008) before plants.

The concept was introduced first by Tettelin et al. (2005), as follows: the pangenome consists of a "core genome containing genes present in all strains and a dispensable genome composed of genes absent from one or more strains and genes that are unique to each strain" (Fig. 2) (Tettelin et al. 2005). This definition has been since then re-used many times (Tettelin et al. 2008; Liang et al. 2012; Lukjancenko et al. 2012; Mann et al. 2013; Lukjancenko 2013; Kahlke 2013; Soares et al. 2013) and extended to transcripts, proteins, pseudogenes as well as to TEs (Transposable Elements) (Lukjancenko 2013; Nguyen 2014; Hirsch et al. 2014; Tranchant-Dubreuil et al. 2019). A pangenomic analysis allows to characterize the folder of genes/sequences available within a given group, as well as to estimate the genomic missing part (*i.e.,* the number of new individuals to be sequenced to reach this almost complete folder) (Tettelin et al. 2008).

Pangenomic studies are performed on a set of organisms/individuals supposed to be representative of a family, a genus, or more generally a species. These analyses provide a large array of phylogenomic information at different levels, from general implications for the group to individual-specific genetic information (Kahlke 2013; Soares et al. 2013). As a corollary, two individuals may not be similar to each other because of the genes they share, but also for the ones they missed (Snipen and Ussery 2010).

Pangenomics is also a way to characterize how conserved are the genomic sequences between individuals within a population. We may, by such means, better understand the genetic variation underlying ecological adaptations and have a more complete view on how genetic variation is important for adaptation and survival of a given group of organisms (Hansen et al. 2012). Indeed, as shown by Caputo et al.

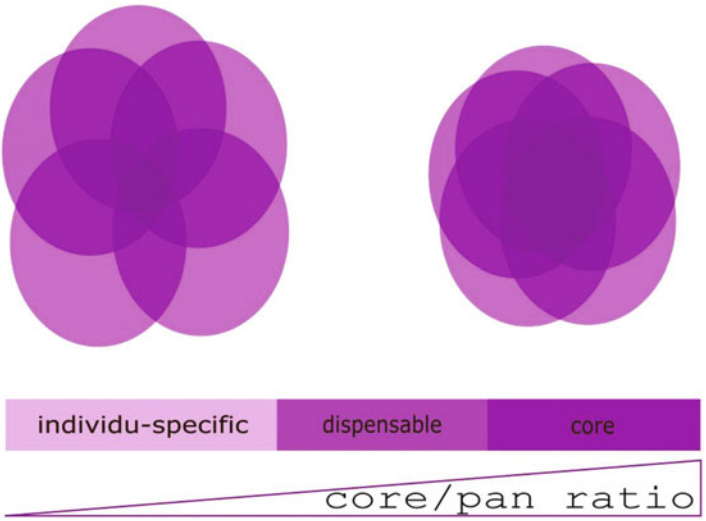

Fig. 2 Pangenome with high or low core/pan ratio. Each individual is represented by a different circle. Intersection between all circles represents the core genome, parts without intersection represent the individual-specific genome, and intersections with not all circle the dispensable genome

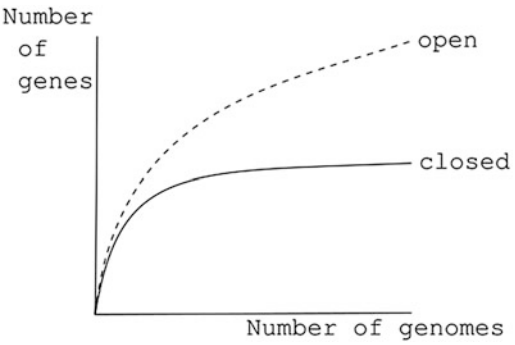

Fig. 3 Open or closed pangenome

(2015), a core/pan ratio (*i.e.,* size of core *vs* size of pangenome, in bp) greater than 90% indicates a high rate of genome conservation between sampled individuals. On the other hand, if this ratio is lower than 85%, then the species may present a certain potential of adaptability because of a higher genomic diversity. Indeed, in this case, if we consider the population, individuals are different enough from each other to be able to respond to rapid and/or abrupt changes and increase the chances of survival for the species as a whole (Fig. 3).

Some studies found that analyzing many genomes of the same species allows to discover more intra-specific genomic diversity than expected (Tettelin et al. 2008). Working on many individuals from the same species leads to identify genomic diversity which might not be discovered if we look only at the inter-species variability (Lukjancenko et al. 2012).

2.1 Core Genome

The core genome is made of genes/sequences present in all individuals of the studied group (Lukjancenko et al. 2012; Lukjancenko 2013; Soares et al. 2013). In bacteria, those genes are conserved and seem to be important genes for basic biological processes (Tettelin et al. 2005; Soares et al. 2013), and thus have been qualified as essential genes (Kahlke 2013). However, as more individuals are added to the study, the less core genes are found (Tettelin et al. 2005; Lukjancenko et al. 2012), and the core genome is generally over-estimated. In the same way, Collins and Higgs (2012) have shown that genes considered as essential (because their function seems to be essential for the survival of the individual) might be absent from the core genome. So, the level of definition of essential is critical in the way of interpretation of a core genome. In this regard, the core genome would compose not only of putative essential (functionally essential) genes but also of genes evolving slowly enough to be present in each genome ("pangenomically" defined as essential), not required for their host survival. Finally, complementation of function can occur through different genes in different individuals, and thus, if considering the sequence, the core genome may be really low, while considering function, the same core genome may be larger.

2.2 Dispensable Genome

Dispensable genes are found in at least two individuals among those studied, but not in all (Tettelin et al. 2005; Liang et al. 2012). These genes mostly belong to secondary biological processes, and are usually associated with adaptation, resistance to biotic and abiotic stresses or, for microorganisms, new host colonization (Tettelin et al. 2008; Mann et al. 2013; Hirsch et al. 2014). For instance, in plants, and more particularly in rice *Oryza sativa*, Schatz et al (2014a) showed that some of them are related to disease resistance genes.

The dispensable genes are supposed to be the main contributors of the variability among individuals within species (Kahlke 2013). Their origin is supposed to be distinct between different phyla: in bacteria, they probably originate from horizontal transfers (Medini et al. 2005), while in eukaryotes they may be related to genes lost, segmental or whole-genome duplications, neogenesis, or TEs activity (Hirsch et al. 2014).

2.3 Individual-Specific Genome

Genes belonging to a single individual are grouped under the term individual-specific genome which is a sub-part of the dispensable genome. (Tettelin et al. 2005; Lukjancenko 2013; Soares et al. 2013). In microorganisms, including bacteria, they are mostly related to phages, horizontal transfers or TEs (Kahlke 2013). For plants, as in rice (*Oryza sativa*), Schatz et al. (2014a) have shown that most of these sequences are repetitive sequences, some being related to genes. They also found that the distribution of these unique regions is not precisely located in a specific region, but they are well distributed across the genome. Different studies have shown that new genes, or genes recently evolved, have generally smaller CDS (coding DNA sequence) compared to older/strictly conserved genes (Cai and Petrov 2010; Capra et al. 2010; Lipman et al. 2002), such as those of the individual-specific-genome, which suggests they might be new or recently evolved (Schatz et al. 2014a). However, the classification as a genome-specific sequence is quite debatable (see (Tranchant-Dubreuil et al. 2019)), as its attribution may be related only to sampling effect or sequencing limits.

2.4 Pangenomics Goes from Bacteria to Plants

Since the beginning of pangenomics (Tettelin et al. 2005), many projects have been performed on different organisms (bacteria, *archaea*, animals, plants, virus, and recently fungi) with a high variation in the range of studied genomes (from 2 individuals to 14,129; Lu et al. 2015) and the number of species involved (between 1 and 3; Darling et al. 2010; Gordienko et al. 2013; Boussaha et al. 2015; Ghatak et al. 2016). Many different methods have been used to be adapted for each data set (see Vernikos et al. 2015 and later in this chapter for more details).

2.5 Pangenomics in Plants

In plants, there have been many pangenomics studies listed in Table 1, 2, and 3. Plants pangenomes seem to be very dynamic and variable depending on species, but their large genome size poses the difficulties for representative sequencing as well as scalable analyses. To override these problems, early studies were limited to chromosomal regions and not the whole genome (Wang and Dooner 2006). Morgante et al. (2007) have first translated the pangenome definition from bacteria to plants, in order to better understand genomic variation. According to this study, performed with two maize (*Zea mays*) lines on four genomic regions, they have shown that only 50% of the sequences are shared, which might constitute the core genome. Similar observations were obtained in barley (*Hordeum vulgare*) and rice (Morgante et al.

Table 1 Pangenomics studies in major crop plants

Species	Number of genotypes	Strategy	References
Zea mays	2 inbred lines	Comparison of BAC assemblies for a few loci	Brunner et al. (2005)
Zea mays	2 inbred lines	Comparison of BAC assemblies on whole genome	Morgante et al. (2005)
Zea mays	27 inbred lines	Reference-based mapping	Gore et al. (2009)
Zea mays	Inbred lines	Comparative genomic hybridization	Springer et al. (2009)
Zea mays	6 inbred lines	Reference-based mapping	Lai et al. (2010)
Zea mays	Inbred lines	Comparative genomic hybridization	Belo et al. (2010)
Zea mays and *Zea mays ssp. parviglumis*	Inbred lines	Comparative genomic hybridization	Swanson-Wagner et al. (2010)
Zea mays	103 inbred lines	Reference-based mapping	Chia et al. (2012)
Zea mays	21 inbred lines	Reference-based mapping on RNA samples	Hansey et al. (2012)
Zea mays	503 inbred lines	*De novo* transcriptome assembly	Hirsch et al. (2014)
Zea mays	14,129 inbred lines	Genetic mapping followed by machine learning approaches	Lu et al. (2015)
Zea mays	1 inbred line	Reference-quality *de novo* genome assembly	Hirsch et al. (2016)
Triticum aestivum	2 accessions (chr3B)	Reference-based mapping and draft assembly of unmapped reads	Liu et al. (2016)
Triticum aestivum	19 accessions	Reference-based mapping and draft assembly of unmapped reads	Montenegro et al. (2017)
Triticum aestivum	2 accessions (700 Mb of chr2D)	Reference-based mapping	Thind et al. (2018)
Oryza sativa	2 accessions	Comparative genomic hybridization	Yu et al. (2011)
Oryza sativa	50 accessions	Reference-based mapping and draft assembly of unmapped reads	Xu et al. (2011)
Oryza sativa	3 inbred lines	High-quality *de novo* genome assemblies	Schatz et al. (2014b)
Oryza sativa	1,483 accessions	Metagenome-like *de novo* assembly	Yao et al. (2015)

(continued)

Table 1 (continued)

Species	Number of genotypes	Strategy	References
Oryza sativa	2 accessions	Comparison of BAC assemblies on whole genome	Zhang et al. (2016)
Oryza glaberrima	3 accessions	*De novo* draft genome assemblies	Monat et al. (2016)
Oryza sativa	3,010 accessions	*De novo* draft genome assemblies and reference-based mapping	Sun et al. (2017)
Oryza sativa	66 accessions	*De novo* draft genome assemblies	Zhao et al. (2018)
Oryza sativa	3,010 accessions	Reference-guided *de novo* assembly	Wang et al. (2018)
Oryza sativa	12 accessions	High-quality *de novo* genome assemblies	Zhou et al. (2020)
Brassica napus and *B. rapa*	BAC segments from 17 accessions	*De novo* draft assembly	Cheung et al. (2009)
Brassica rapa	1 doubled haploid line, 1 inbred line and the reference genome	*De novo* draft genome assembly	Lin et al. (2014)
Brassica oleracea and *B. macrocarpa*	10 lines	Iterative mapping and assembly	Golicz et al. (2016a, b)
Brassica napus	2 accessions	*De novo* draft genome assembly then reference-based guided	Bayer et al. (2017)
		Pseudomolecule-level assembly	
Brassica napus	53 accessions	Iterative mapping and assembly	Hurgobin et al. (2018)
Cicer arietinum	35 accessions	Draft reference-based mapping	Thudi et al. (2016a, b)
Cicer arietinum	129 accessions	Draft reference-based mapping	Thudi et al. (2016b)
Glycine soja and *Glycine max*	31 accessions	Reference-based mapping	Lam et al. (2010)
Gycine max	4 accessions	Comparative genomic hybridization and reference-based mapping on exome data	Haun et al. (2011)
Glycine max	4 accessions	Comparative genomic hybridization and reference-based mapping on exome data	McHale et al. (2012)
Glycine soja	7 accessions	*De novo* draft genome assembly	Li et al. (2014)

(continued)

Table 1 (continued)

Species	Number of genotypes	Strategy	References
Glycine max	41 accessions	Comparative genomic hybridization and reference-based mapping	Anderson et al. (2014)
Sorghum bicolor	3 inbred lines	Reference-based mapping	Zheng et al. (2011)
Hordeum vulgare ssp. vulgare and *H. vulgare ssp. spontaneum*	14 accessions	Comparative genomic hybridization	Muñoz-Amatriaín et al. (2013)
Ipomoea trifida	2 accessions	*De novo* draft genome assemblies and cross mapping	Hirakawa et al. (2015)
Solanum tuberosum	12 simple and double monoploid clones	Reference-based mapping	Hardigan et al. (2016)
Capsicum annuum, *C. baccatum*, *C. chinense* and *C. frutescens*	383 accessions	*De novo* draft genome assemblies and reference-based mapping	Lijun et al. (2018)
Solanum lycopersicum, *S. pimpinellifolium*, *S. cheesmaniae* and *S. galapagense*	725 accessions	Map-to-pan	Gao et al. (2019)
Sesamum indicum	5 accessions	Reference-guided and *de novo* assembly	Yu et al. (2019)
Cajanus cajan	90 accessions	Reference-based mapping and draft assembly of unmapped reads	Zhao et al. (2020)

2007). In another study, this time with 8 lines of maize on the *bz1* locus, Wang and Dooner (2006) estimated that core genome is not as large as it was expected earlier from the analysis of two individuals (from 25 to 84%), and that the dispensable genome, for its part, increases when more genomes are included, as seen in bacteria.

Plants genomes have some specificity making their analysis more complex than for bacteria or microorganisms, explaining why pangenomics studies came later to plants. Among those, we can cite the high level of polyploidy, *i.e.* partial or even whole-genome duplication, particularly frequent in angiosperm species (Cheung et al. 2009). Some of these recent duplications may impair the pangenome identification: thus, in rapeseed (*Brassica napus*), various subgenome loci are distinguishable only through a few SNP and InDels (Cheung et al. 2009). However, Lin et al. (2014) performed a pangenomics study with three genomes of *Brassica rapa*, the diploid progenitor of *Brassica napus* and shown that on average 1,200 genes are individual-specific (and thus subgenome-specific), which provided a supplementary evidence of increasing interest in pangenomics studies.

Table 2 Key results from pangenomics studies in crop plants

Species	Reference	Number of pan-genes	% of core genes	Pangenome size	% core size	Example of SVs
Oryza sativa	Schatz et al. (2014b)	40,362	92.16%	341 Mb	88.82%	PAV of Phosphorus uptake1 (*Pup1*), major rice QTL associated with tolerance to phosphorus deficiency in soils. CNV and InDel of *S5*, major locus for hybrid sterility in rice that affects embryo sac fertility. CNV of Submergence 1 (*Sub1*), major QTL on chromosome 9 determining submergence tolerance in rice
Brassica ssp.	Golicz et al. (2016a, b)	61,379	81.3%	587 Mb	NA	PAV of auxin-related genes (*AUX, IAA, GH3, PIN, SAUR, TIR, TPL, and YUCCA* PAV of flowering related genes (*MAF5, SEP2, ARP4, GID1B, FPF1-like, FHY1, GA2, GA3, and CO* CNV of Flowering locus C (*FLC*), an important regulator of vernalization and flowering time. PAV of glucosinolate-related genes *CYP79A2, SUR1, and SOT18*
Capsicum ssp.	Lijun et al. (2018)	51,757	55.7%	4.31 Gb	NA	SV of capsanthin/capsorubin synthase (*Ccs*), locus involved in carotenoid contents. SV of phytoene synthase gene (*Psy*), locus involved in carotenoid contents. Sv of Pungent gene 1 (*Pun1*), locus involved in capsaicinoid biosynthetic pathway
Sesamum indicum	Yu et al. (2019)	26,472	58.21%	554 Mb	46.70%	PAV between modern cultivars and landraces for genes with functions related to energy metabolism, growth and development, as well as biomass accumulation PAV between landraces and modern cultivars for genes with functions related to environmental adaptation, signal transduction, protein folding, sorting, degradation, transport, and catabolism

Table 3 Pangenomics studies in Arabidopsis and other plants

Species	Number of genotypes	Strategy	References
Arabidopsis thaliana	3 accessions	Reference-based mapping	Ossowski et al. (2008)
Arabidopsis thaliana	4 accessions	Reference-based mapping	Santuari et al. (2010)
Arabidopsis thaliana	80 accessions	Reference-based mapping and draft assembly of unmapped reads	Cao et al. (2011)
Arabidopsis thaliana	18 accessions	Iterative reads mapping combined with *de novo* assembly and reference-based mapping	Gan et al. (2011)
Arabidopsis thaliana	2 accessions	Reference-based mapping	Lu et al. (2012)
Arabidopsis thaliana	3 accessions	High-quality *de novo* genome assembly	Pucker et al. (2016)
Arabidopsis thaliana	2 accessions	High-quality *de novo* genome assembly	Pucker et al. (2019)
Populus nigra, *Populus deltoides* and *Populus trichocarpa*	22 accessions	Reference-based mapping on both DNA and RNA samples	Pinosio et al. (2016)
Populus alba, *P. davidiana*, *P. euphratica*, *P. lasiocarpa*, *P. nigra*, *P. deltoides*, *P. cathayana*, *P. simonii*, *P. ussuriensis* and *P. maximowiczii*	10 accessions	Reference-based mapping	Zhang et al. (2019)
Utricularia gibba	13 specimens	Reference- and other species-based mapping and draft assembly of unmapped reads	Alcaraz et al. (2016)
Medicago truncatula	16 accessions	*De novo* assemblies	Zhou et al. (2017)
Brachypodium distachyon	54 inbred lines	*De novo* draft genome assembly	Gordon et al. (2017)