Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Jeffrey Bennetzen · Sherry Flint-Garcia
Candice Hirsch · Roberto Tuberosa *Editors*

The Maize Genome

Springer

Compendium of Plant Genomes

Series editor

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

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant Arabidopsis thaliana in 2000, whole genomes of about 70 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

Interested in editing a volume on a crop or model plant? Please contact Dr. Kole, Series Editor, at ckole2012@gmail.com

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

Jeffrey Bennetzen · Sherry Flint-Garcia
Candice Hirsch · Roberto Tuberosa
Editors

The Maize Genome

Springer

Editors
Jeffrey Bennetzen
Department of Genetics
University of Georgia
Athens, GA, USA

Sherry Flint-Garcia
USDA-ARS
Columbia, MO, USA

Candice Hirsch
Department of Agronomy
and Plant Genetics
University of Minnesota
St. Paul, MN, USA

Roberto Tuberosa
Department of Agricultural
and Food Sciences
University of Bologna
Bologna, Italy

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-319-97426-2 ISBN 978-3-319-97427-9 (eBook)
https://doi.org/10.1007/978-3-319-97427-9

Library of Congress Control Number: 2018950816

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, the 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 mapping of polygenic clusters or QTLs. Molecular markers were also used in 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 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 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, 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. 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, development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this preface for the first volume of the new series 'Compendium of Plant Genomes,' a net search tells me that complete or nearly complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and 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 is growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated Web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful both to students and to 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, potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

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

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

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

Kalyani, India Chittaranjan Kole

Preface to "The Maize Genome" Volume

It is now three decades since the mapping of QTLs for agronomic traits, including yield, was first reported in maize. Following this pioneering and groundbreaking work, the pace of progress in maize genomics and its breeding applications have been nothing short of spectacular. This progress continued to accelerate, as witnessed by the publication of the first assembly of the maize genome a decade ago. This second milestone paper prompted and paved the way to a wealth of manuscripts and the discovery of several genes/QTLs with a relevant role in maize growth and field performance.

Based upon this premise, this volume builds on such knowledge and provides a glimpse into some of the recent advances in the study and characterization of maize genome structure, evolution and function, and how this information can be harnessed to enhance the effectiveness of genomics-assisted breeding as well as gene/QTL cloning and study. Suitable platforms, genetic materials, and databases now bridge forward and reverse genetics approaches and allow for an unprecedented level of genetic and functional resolution, particularly for quantitative traits. Maize genomics now provides breeders with a formidable toolbox for tailoring hybrids better adapted to face the challenges posed by climate change, while ensuring an environmentally sustainable and profitable production of one of the most important crops for mankind.

Overall, the chapters in this volume emphasize the importance of deeply characterizing the maize genome in order to identify rare haplotypes with beneficial effects that are not yet represented in elite germplasm. Large-scale resequencing coupled with an equally deep analysis of the transcriptome, proteome, and metabolome will accelerate the cloning of agronomically valuable loci, paving the way to a more effective harnessing of biodiversity, more accurate modeling, and, most importantly, the fine-tuning of key sequences via gene editing.

We hope that this volume will provide maize scientists with a better appreciation of the complexity underpinning phenotypic variability while stimulating their curiosity and interest in undertaking new studies to further enhance our understanding of such complexity.

The editors are grateful to the authors of the different chapters for reviewing the published research work in their area of expertise and, in some cases, sharing their unpublished results to update the articles. We also

appreciate their cooperation in meeting the deadlines and in revising their manuscripts, whenever required. This notwithstanding, the editors remain responsible for any errors that inadvertently might have crept in during the editorial work.

Athens, USA — Jeffrey Bennetzen
Columbia, USA — Sherry Flint-Garcia
St. Paul, USA — Candice Hirsch
Bologna, Italy — Roberto Tuberosa

Contents

Part I

Genome Sequencing and Genotyping

1 Draft Assembly of the F2 European Maize Genome Sequence and Its Comparison to the B73 Genome Sequence: A Characterization of Genotype-Specific Regions

Johann Joets, Clémentine Vitte and Alain Charcosset

Abstract

Maize is well known for its exceptional structural diversity, including copy number variants (CNVs) and presence/absence variants (PAVs), and there is growing evidence for the role of structural variation in maize adaptation. F2 is a European maize line resulting from a long-term independent evolution relative to the reference American line B73. It also presents strong heterosis when crossed to American lines related to B73 or PH207, which has been instrumental for the development of hybrid breeding in Northern Europe. De novo genome sequencing of the French F2 maize inbred line revealed 10,044 novel genomic regions larger than 1 kb, making up 88 MB of DNA, that are present in F2 but not in B73 (PAV). This set of maize PAV sequences allowed us to annotate PAV content and to identify 395 new genes. We showed that most of these genes display numerous features that suggest they are either rapidly evolving genes or lineage-specific genes. Using PAV genotyping on a collection of 25 temperate lines, we also analyzed and provided the first insights about PAV frequencies within maize genetic groups and linkage disequilibrium in PAVs and flanking regions. The pattern of linkage disequilibrium within PAVs strikingly differs from that of flanking regions and is in accordance with the intuition that PAVs may recombine less than other genomic regions. As it was shown by several other authors, most PAVs are ancient, while we show that some are found only in European Flint material, thus pinpointing structural features that may be at the origin of adaptive traits involved in the success of this material. We conclude by some words on future directions.

1.1 F2 Is Characteristic from a European Hybridization Event

The story of European maize traces back to its first introduction in 1493 by Columbus after his first trip to America. Being adapted to the tropical climate of the Caribbean, these varieties could be cultivated only in warm regions of the Mediterranean basin and would have been too late flowering to produce seeds in cooler environments. After this seminal trip, explorations lead to the rapid discovery of the northeast American coast, up to cool temperate climates of northern Canada. Most Native American people of the east coast and neighboring inland regions

J. Joets (✉) · C. Vitte · A. Charcosset
Genetique Quantitative et Evolution – Le Moulon, INRA, CNRS, AgroParisTech, Université Paris-Sud, Université Paris-Saclay, Gif-Sur-Yvette, France
e-mail: johann.joets@inra.fr

J. Bennetzen et al. (eds.), *The Maize Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-319-97427-9_1

were relying to a large extent on the cultivation of specific maize varieties, referred to as Northern Flints because of their hard kernel texture. Their short planting to flowering interval was making them adapted to temperate environments. Genetic and historical investigations show that these temperate varieties were rapidly introduced into Europe and cultivated on a significant scale in Northern countries like Germany before 1539 (see Tenaillon and Charcosset 2011 for a review). Genetic analyses highlight that these two main introductions of maize into Europe at some step hybridized, possibly also with introductions of lesser importance, leading to varieties specific to mid-latitude European regions such as the Pyrenean valleys (Brandenbourg et al. 2017). Varieties from these introductions produced staple food in these regions until the late 1960s.

After WW2, traditional European varieties have been used to develop inbred lines, which were tested for their ability to produce hybrid varieties. Among these lines, F2 which stands for France n 2 was developed from the Lacaune population, cultivated on a cool South West France plateau at approximately 800–1000 meters elevation (Cauderon 2002). It proved outstanding in its ability to produce superior hybrids when crossed to inbred lines from North American origin, referred to as Dents because of their soft endosperm texture leading to a depression on the kernel crown. These first European Flints by American Dent hybrids were particularly successful for grain production in Northern Europe. This success can be interpreted as the combination of environmental adaptation features (adaptation to cool spring in particular) contributed by European Flints with yield potential contributed by American dents. Modern hybrid breeding for grain or silage production in North European regions is still based to a large extent on this pattern. As for F2 itself, it remained extremely successful and used in hybrids until the mid-1990s, especially when crossed to the American Dent lines PH207- or B73-related lines. Since that time, it has served as one of the three major progenitors of modern European Flint lines, along with lines F7 and Ep1.

Genotypic evaluations have confirmed a striking divergence between European Flint (i.e., F2) and American Dent lines (i.e., B73) (see Gouesnard et al. 2017). There are also striking phenotypic characteristics that differ between the two lines (Table 1.1). All elements therefore concur to expect large differences between the genomes of B73 and F2, possibly related to heterosis and adaptive traits.

1.2 B73- and F2-Specific Genome Region Discovery and Combination into a Draft B73–F2 Pan-Genome Sequence

Maize SV discovery at the whole-genome scale through comparative genomic hybridization arrays (aCGH)-based analysis of low copy regions led to detection of thousands of PAVs and CNVs between two American maize inbred lines (Springer et al. 2009; Beló et al. 2010). Probing of structural variation through a global analysis of read depth in over 100 maize lines showed that over 90% of the maize genome shows some degree of CNV between lines (Chia et al. 2012). While they allowed cost-effective and genome-wide discovery of PAVs/CNVs in multiple samples, these aCGH- and remapping-based studies did not allow discovering novel regions absent from B73. Discovery of over 2,000 new non-B73 genes was performed using massive mRNA sequencing on over 500 inbred lines, thus providing a cost-effective approach to solve this issue (Hirsch et al. 2014). Nevertheless, discovery of new genes with such mRNAseq-based strategy is dependent on sequencing depth and on the number of tissues and conditions analyzed. It is therefore likely to miss new genes with very low expression or expressed in very specific conditions. Moreover, this type of strategy does not provide sequence breakpoints, thus hampering exploration of underlying mechanisms, and is limited to analysis of the genic portion of the genome. Genome sequencing and de novo assembly can ultimately provide precise breakpoint positions, distinction between CNV and PAV, access to novel sequences, variant size information, and exploration of non-genic space. Targeted assembly of non-B73

Table 1.1 Summary of main phenotypic/adaptive differences between F2 and B73

Trait	F2	B73
Cold adaptation	Mid-tolerant	Sensitive
Leaf number[a]	14.1	20.6
Plant height (cm)[a]	142	210
Flowering time[a]	Early	Late (+20 days)
Kernel number/ear[a]	221	473
Endosperm	Hard (Flint)	Soft (Dent)

[a]Estimations from Bouchet et al. 2017

regions from elite Chinese and American lines led to the discovery of 5.4 MB of new sequence absent in the reference genome assembly (Lai et al. 2010). However, the low sequencing depth used (5X) limited the reconstruction of full-length PAV sequences. Because discovered PAVs were short and incomplete, complete annotation and anchoring to the reference genome were challenging, thus impeding functional prediction and breakpoint detection. Sequence assembly of the PH207 genome provided a matrix for reciprocal comparison of PH207 and B73 gene coverage using remapping of massive sequencing reads. It led to the discovery of over 2,500 genes, which were found specific to one genotype either partly or fully (Hirsch et al. 2016). However, analyses were focused on gene-annotated regions only, so this study did not identify the boundaries of the SVs containing these genes. In a complementary work, we produced a draft sequence assembly of the F2 genome and identified over 10,000 genomic regions present in F2 and absent from B73 (Darracq et al. 2018). New F2 regions make up 90 MB (4% of F2 genome size). Using RNAseq data from 12 tissues and conditions, we identified near 400 genes expressed in F2 PAVs. Expression breadth revealed that PAV genes are expressed in a limited set of conditions and at a lower rate than average B73 genes, consistent with previous results (Hirsch et al. 2016). Hence, while most F2-specific genes are likely present in our assembly (which covers 65% of the F2 genome), we likely did not explore enough conditions to have a RNAseq support for all new genes, and further transcriptome studies may help unravel more F2 specific genes.

Genome comparison studies provide a starting point to unravel the molecular origin and the function of maize structural variants. A consensus assembly that represents many individuals is likely to improve use of sequence-based chromatin and transcription data, as well as SNP detection. Decreasing the amount of spurious alignments would help to better estimating transcript abundance or heterozygosity prediction. How to best combine genomic sequences from several maize inbreds for aligning Illumina reads in a compute-efficient way remains a challenge (Consortium 2016; Hurgobin and Edwards 2017). While using each genome separately is an option, the rapid increase of whole-genome sequences will soon make it too computationally costly. Rather, we propose to build pan-genomic sequences by adding up the non-B73 genomic sequences to the B73 genome sequence. As a proof of concept, we built a first B73–F2 pan-genomic sequence, by adding up the 90 MB of F2-specific sequences to the 2.1 GB B73 genome sequence (Darracq et al. 2018). In the following sections, we will show how our approach can be used for studying (i) characteristics of PAVs and underlying genes, (ii) PAV LD properties, (iii) PAV history among maize inbreds, and (iv) perspectives for improved discovery and use in post-genomic studies.

1.3 F2 Non-B73 Genes Are Expressed in Other Maize Lines, but Are not Well Conserved Outside Maize

The 395 novel predicted genes present in F2 and absent of B73 are all supported by RNAseq experiments. In a comparison of RNAseq-based

abundance of F2 PAV genes versus B73–F2 shared genes, we showed that F2 PAV genes are expressed in less tissues than shared genes. This suggests that RNAseq-based identification of genes in F2 PAVs may have missed some genes due to lack of transcriptomic data in a large enough set of tissues/conditions. When comparing F2 PAV genes with transcriptome datasets from maize and related species, we showed that 90% have a blast best hit with a maize sequence, from another genotype other than B73, and only 8% have a best hit in closely related Poaceae species (20 in *Sorghum*, 3 in *Setaria*, 3 in *Saccharum*, 1 in *Miscanthus*, 1 in *Panicum*, 1 in *Tripsacum*, and 2 in *Oryza*). Interestingly, most orthologous sequences that were found derived from expressed sequence tags (ESTs) that were produced in the early 2000s for large numbers of maize genotypes, tissues, and conditions. By comparison, only 12 novel F2 sequences align to sequences from the pan-transcriptome assembled by Hirsch et al. (2014), which included more than 500 genotypes but from a single tissue. This suggests tissue/condition specificity of PAV gene expression and highlights the need for enlarging RNAseq datasets to improve discovery, annotation, and characterization of genotype-specific genes.

Functional annotation by search of sequence similarity with UniProtKB/Swiss-Prot proteins and InterPro protein domains allowed annotation of 91 F2 PAV genes. Among these, 17 (20%) are putatively involved in stress response and plant defense, 11 (12%) in biosynthetic processes, 10 (12%) in development, 5 (6%) in protein synthesis, and 5 (6%) in chromatin remodeling. For B73, PAV annotation was based on existing RefGen v2 5a annotation and provided a molecular functional prediction for 25 B73 PAV genes. Grouping of these molecular functions highlighted six sequences (25%) putatively involved in metabolism, four (16%) in stress response and plant defense, four (16%) in protein degradation, and two (8%) in cytoskeleton/microtubule. These results suggest that F2 PAV genes and B73 PAV genes are enriched in functions involved in stress response. Similarly, an enrichment of function related to stress response was observed in a set of maize PAV genes identified from a comparison between PH207 and B73 (Hirsch et al. 2016). Hence, transcriptome profiling in abiotic and biotic stress conditions is likely to greatly increase prediction and annotation of genotype-specific genes. Interestingly, in a recent study analyzing the diversity of 67 maize genomes from landrace representatives from the major maize genetic groups, including European lines, we uncovered that genes involved in abiotic stress tolerance have played a role in maize adaptation to European conditions (Brandenburg et al. 2017). This opens interesting perspectives in deciphering the role of PAVs in maize adaptation.

While this study allowed for prediction of PAV functions, protein prediction was successful for only 23% of the F2 novel genes sequences. This suggests that F2 PAV genes may be less conserved than other genes. To test this, we compared PAV and non-PAV genes in maize in terms of both number of genes with protein similarity, and levels of similarity to the protein sequence in an increasingly distant species set, from *Sorghum bicolor* to *Arabidopsis thaliana*. As predicted, the proportion of proteins with no significant similarity with other plant proteome is higher for F2 PAV genes (Fig. 1.1a), and when a protein is found, average identity is markedly (12 to 25%) lower for F2 PAV gene proteins than for B73 FGS proteins (Fig. 1.2b). This lower conservation suggests that PAV genes identified in F2 compared to B73 could have evolved more rapidly than non-PAV genes or emerged recently as novel genes.

With shorter size, shorter expression breadth, enrichment in stress-related functions, and lower conservation at the protein level than average genes, F2 PAV genes have many characteristics of orphan genes (Arendsee et al. 2014). Orphan genes either emerge de novo from non-genic sequence or derive from ancient gene duplications followed by divergent accumulation of mutations beyond recognition. Nevertheless, functional characterization of these genes is still challenging. Because discovery and annotation of PAV genes are a major goal in maize and plant biology, many laboratories are generating

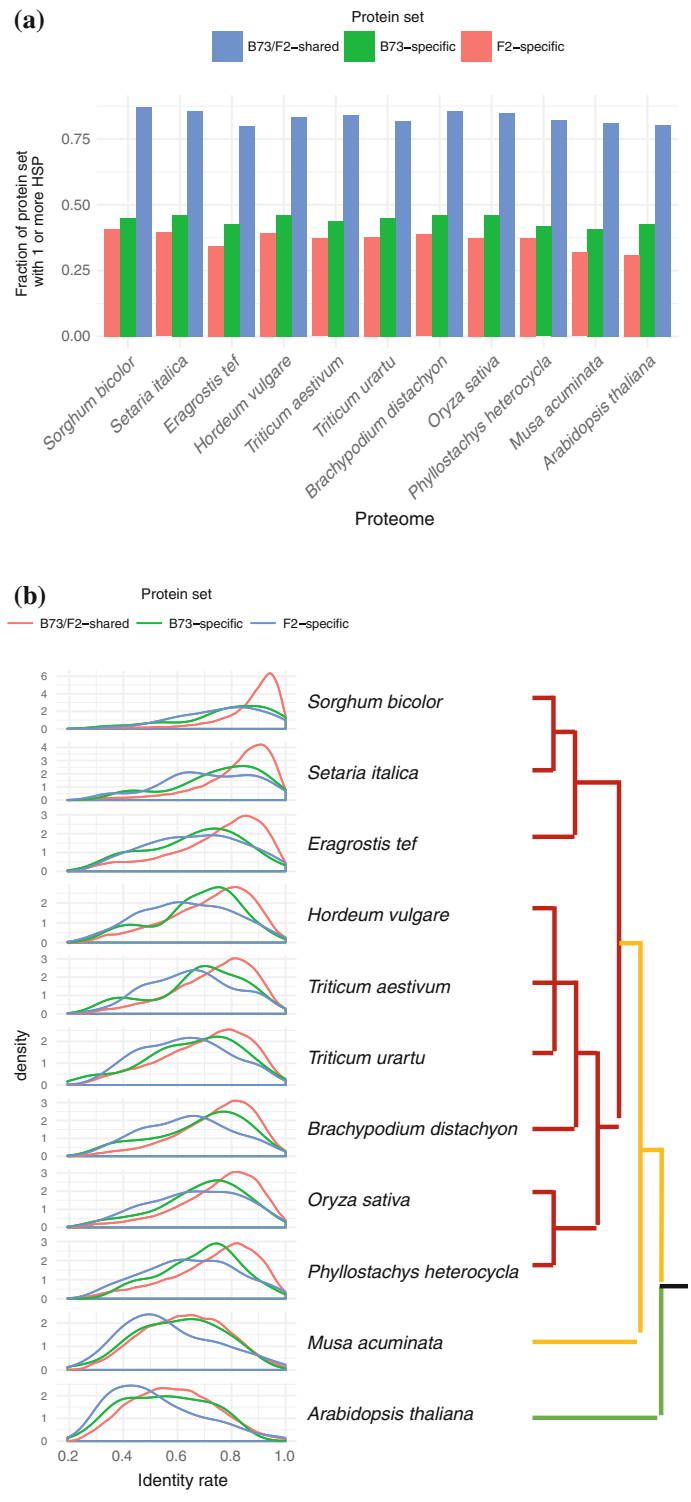

Fig. 1.1 Conservation of B73 and F2 presence/absence variation (PAV) proteins compared to B73–F2 shared proteins. **a** Fraction of protein sets (B73–F2 shared proteins, B73-present/F2-absent proteins, F2-present/B73-absent proteins) with at least one blastp hit (tilled HSP) (E value > = 10^{-3}) with several whole plant proteomes. **b** Distribution of identity rate of blastp best hit (tilled HSP) for the three protein sets against 11 whole plant proteomes. Plant proteomes are sorted according to the genetic distance with maize from sorghum to Arabidopsis, which is the most distant of maize. Length of branches of the phylogenic tree are arbitrary, red branches are for grasses, orange for monocot, and green for eudicot

RNAseq and proteome datasets to help in this task. We believe that this effort will provide important information for better understanding the origin and role of orphan genes.

On the other hand, it has been argued that most of the dispensable genes are members of duplicated gene or large gene family members (Swanson-Wagner et al. 2010). The absence of the gene could therefore be complemented by another member of the family. Of the 395 novels genes discovered in F2, only 116 exhibit greater than 50% identity over at least 80% of their length with a protein of B73, and therefore, 70% of these proteins have no or distant similarity with protein in B73. While this is certainly an underestimation of the number of unique PAV genes as the B73 and F2 genomes are not complete, it is possible that a significant fraction of PAV genes, and possibly biological functions, are absent in some genotypes.

1.4 The Dispensable Genome: A Genomic Faction that Recombines Less Than the Rest of the Genome

Linkage disequilibrium (LD) is the non-random association between alleles at different loci. LD contains information about recombination, demographic history, and gene conversion. LD between copy number variation and flanking SNPs has been found to be higher than between SNPs in genomic regions neighboring CNVs (Schrider and Hahn 2010). This was attributed to the fact that many CNVs have changed genomic location through recurrent duplications and deletions compared to other loci (Schrider and Hahn 2010). In the case of PAV, LD pattern between SNPs within the PAV or between a PAV and flanking marker should not follow these of CNVs. As presented above, PAV genes have a particular mutation pattern and this may impact local LD. But most PAVs do not harbor genes, and whether the whole PAV region evolves at a different rate than other loci remains to be elucidated. To get a first insight on LD pattern between PAVs and their flanking regions, we estimated LD extent for each PAV coded as 0 (absence)/1 (presence) or using the SNP located within the PAV and with shortest distance to the breakpoint. LD was then estimated between this reference polymorphism and SNPs of the flanking region, with increasing distance. While the first approach involves all individuals, for the second, LD can be estimated only when SNPs can be evidenced within the PAV, hence only in the subset of individuals that carry the present allele. For this, we developed a statistical approach to genotype PAV presence and absence alleles using low depth (3x–5x) resequencing data aligned on our B73–F2 pan-genome sequence and applied it on a dataset from a panel of 25 maize lines representing American and European maize genetic groups (Darracq et al. 2018). We compared these LD patterns with those estimated for reference genomic regions and their flanking regions. We showed that LD pattern between PAVs and their flanking regions resembles the same pattern observed between random genes and their flanking regions (Fig. 1.2a). While this might be due to our detection approach to discover PAVs, this first analysis shows that for these PAVs in our panel, LD decreases rapidly. This suggests that PAVs are likely not to be captured by genotyping SNPs, unless these are located within less than 1 kb of the PAV breakpoint.

To investigate whether PAVs recombine less than other genomic regions, we compared LD patterns within PAVs to LD patterns in their flanking regions. While LD depends on demographic history of the lines tested, this effect should be the same for two adjacent genomic regions such as a PAV and its flanking regions, thus giving a relative difference of local recombination rates. For this analysis, within-PAV LD was estimated by comparing pairwise SNPs located inside the variant sequence to pairwise SNPs in the PAV upstream or downstream flanking regions. On average, LD is stronger within a PAV as compared to the flanking regions (Fig. 1.2b). Hence, PAVs seem to recombine less than their flanking regions.

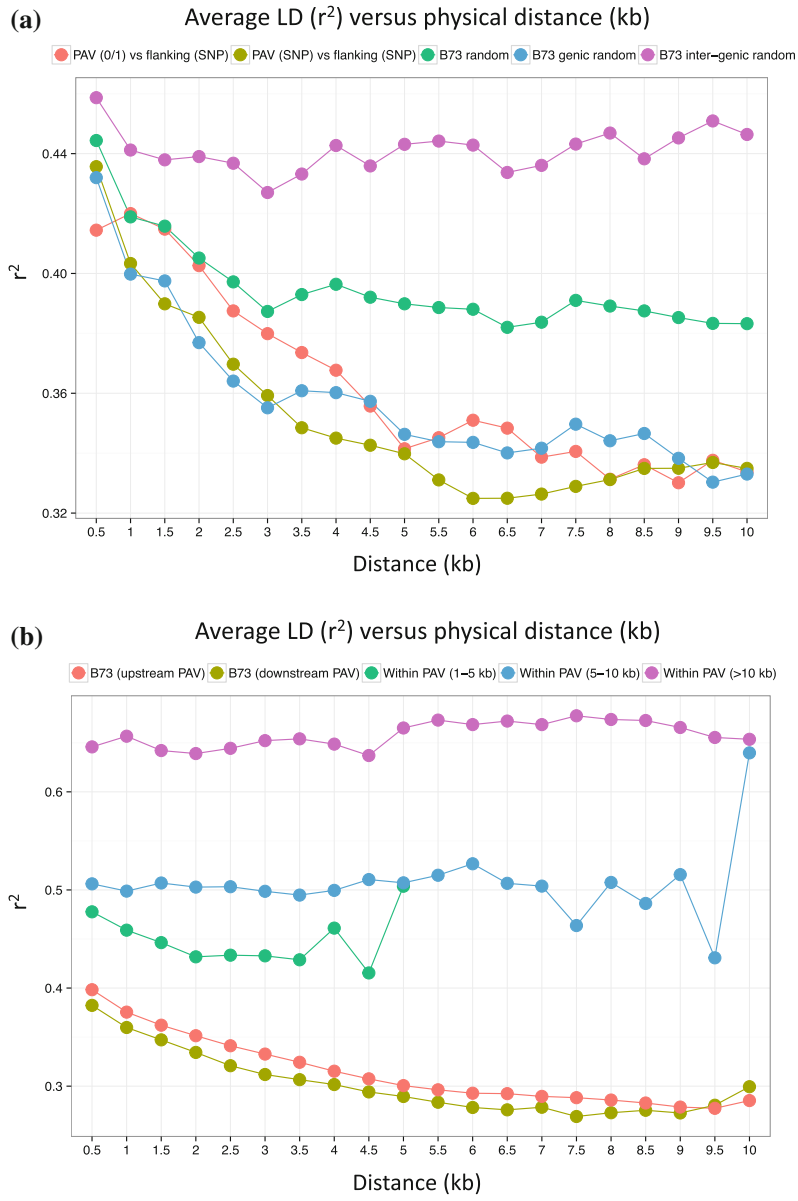

Fig. 1.2 Linkage disequilibrium (LD) decay pattern in presence/absence variation (PAV) regions. **a** LD decay between PAV and flanking region compared to LD decay between gene or TE and flanking regions, respectively. **b** Within-PAV LD compared to LD in flanking regions. PAVs were grouped into three classes according to their size; 1–5 kb PAVs green, 5–10 kb PAVs blue, >10 kb PAVs pink

This result may be due to the fact that PAV sequences can undergo recombination only when present in both gametes, a situation that is less frequent than for shared flanking regions. Of course, this situation depends on the PAV allele frequency, which also depends on the age of the PAVs, so we expect a large range of recombination rates among PAVs. Indeed, when considering PAVs individually, contrasting LD patterns can be observed. For instance, cases of very strong LD are found (Fig. 1.3 left), while in some cases LD patterns reveal subsets of recombining regions within the PAV sequence (Fig. 1.3 right). This difference is likely due to the differences in the date of appearance of the PAV in the population, its frequency in the population upon creation, as well as the temporal dynamics of this frequency.

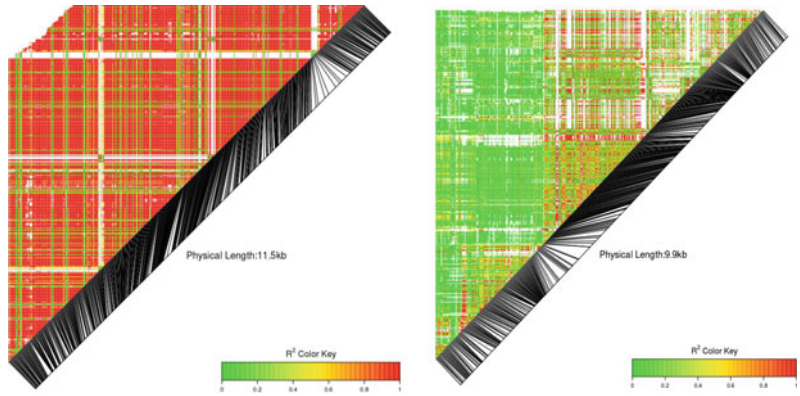

Fig. 1.3 Two examples of contrasted within presence/absence variation (PAV) linkage disequilibrium (LD) patterns. Left most of SNPs are in very high LD all along the PAV. Right two regions of high LD are separated by a breakpoint

1.5 Analysis of PAV Alleles at the Population Level

To investigate to what extent B73–F2 PAVs are conserved among maize genetic groups, we used the genotyping of PAV sequences in our temperate maize panel to estimate frequencies in the different genetic groups. As expected, F2 novel regions were more often present in other European Flints than in any other set of inbreds. Only a small number was detected in the Stiff Stalk group, to which B73 belongs and where the "absent" allele was found. Inbred lines from France or close proximity (Pyrenean) shared more variants with F2 than lines from any other origin, independent of their classification into European Flint and Northern Flint groups, thus reflecting the history of the European germplasm. PCA-based analyses from PAV or SNPs showed very similar classification, showing that SNPs and PAVs have segregated similarly.

A large proportion of PAVs are shared between F2 and at least one other European Flint or one Corn Belt Dent line, which were the most represented groups in our panel. PAVs that were found present in all the four genetic groups were also generally found at high frequency in all groups, suggesting an ancient and shared origin. Consistently, a comparative genomic hybridization experiments on 19 maize lines and 14 teosinte, the wild ancestor of maize, found that 86% of the SVs (CNV and PAV) that were identified were also present in teosinte (Swanson-Wagner et al. 2010). However, 347 PAVs were present only in maize but not in teosinte, and among them, 257 were present in only two to three maize lines suggesting these variants could be specific to maize. We also observed that when PAVs are present in only one genetic group their frequency is low in this group, suggesting the occurrence of recently emerged PAVs. Interestingly, among the 4,218 PAVs that we scored, 396 were found only in European Flints and 134 only in F2 (Darracq et al. 2018). Genotyping of these putative European-specific PAVs in larger maize panels will allow precise allele frequencies and group specificity to be determined.

1.6 Tomorrow's Challenges in Maize Structural Variation

Over the past decade, there has been a growing attention for structural variation in plant evolution. In maize, several genomic studies, including ours, have described some of the characteristics of CNVs and PAVs. But such studies are still in their infancy, and many questions remain to be solved. First, because the maize genome is highly repetitive, discovering structural variants in the repetitive fraction is still a challenge, and most structural variants that have been discovered are from low copy regions. Some studies have made

the choice to focus on genes, which is a cost-effective way of finding SVs with possible phenotypic impact (Hirsch et al. 2014). Using a non-targeted, without a priori approach, we could discover full-length PAVs containing both genic and non-genic regions and characterize their breakpoints. This gave us access to their full sequence content and made LD analyses possible. However, only a subset of our F2 PAVs could be anchored, either because their breakpoints could not be unambiguously anchored or because the assembly was not complete enough to extend PAVs to their biological breakpoints. In both, cases, these issues are linked to the highly repetitive nature of the maize genome, which impairs both unambiguous alignments of short reads in remapping experiments or in whole-genome assembly. This issue might soon be solved, as several maize whole-genome assemblies are under progress. High-quality metrics obtained from new assembly methodologies will open the way to whole-genome sequence comparison, thus eliminating the problem of aligning short reads. Such assemblies are now available for American lines (B73, PH207, W22, CML247) and European lines (EP1, F7). We will soon double this number by adding seven new genome sequences from lines of interest for the European community, and with contrasted genome sizes as well as the complete set of NAM founder parents.

A second challenge is to discover genes standing within these structural variants. As we presented, the particular features of these genes make them difficult to annotate, and the generation of large datasets of RNAseq and proteomic data in many tissues and conditions will be necessary to solve this problem. For this reason, for our seven genotypes and for B73, we are generating deep mRNAseq datasets from a set of tissues from standard- and abiotic-constrained conditions.

Once discovery and annotation of SV will be resolved, the next step will be to combine the information given by these new datasets to make the best use of it. Several laboratories are working on this question, and discussions are emerging. But this is only the beginning, and the maize community needs to organize.

Clearly, pan-genome sequence will be very useful for better analyzing phenotypic data at the molecular (methylome, transcriptome, proteome) or plant scale to find the underlying genetic components. Using the entire genomic information in GWAS will therefore be a major task in the coming years, and typing both SVs and SNPs will be necessary. We developed a pan-genome strategy that allows efficient alignment of resequencing data, as well as an efficient statistical methodology to classify PAVs as present or absent. This methodology can be used across a combination of a large number of maize lines. However, considering the history of maize, and the relatively limited bottleneck involved in its domestication, reconstructing haplotypes representing the entirety of maize genetic diversity will likely require retrieving information from hundreds of maize lines. This number is likely too high for producing public whole-genome sequence assembly resources for all of them, and defining a cost-effective strategy to do so will be an incoming task. We believe discussions at the community level will help build homogeneous datasets that can profit the whole community.

References

Arendsee ZW, Li L, Wurtele E (2014) Coming of age: orphan genes in plants. 19:698–708. https://doi.org/10.1016/j.tplants.2014.07.003

Beló A, Beatty MK, Hondred D et al (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120:355–367. https://doi.org/10.1007/s00122-009-1128-9

Bouchet S, Bertin P, Presterl T, Jamin P, Coubriche D, Gouesnard B, Laborde J, Charcosset A (2017) Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. Heredity 118:249–259

Brandenburg J-TT, Mary-Huard T, Rigaill G et al (2017) Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. PLoS Genet 13:e1006666. https://doi.org/10.1371/journal.pgen.1006666

Chia J-MM, Song C, Bradbury PJ et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44:803–807. https://doi.org/10.1038/ng.2313

Cauderon A (INRA, ed) (2002) L'INRA dans l'amélioration des plantes des «Trente Glorieuses» à la lumière des préoccupations actuelles, in: L'Amélioration des Plantes, continuités et ruptures (http://www.inra.fr/gap/viescientifique/animation/colloqueAP2002/Cauderon.pdf)

Consortium CP-G (2016) Computational pan-genomics: status, promises and challenges. Brief Bioinform. https://doi.org/10.1093/bib/bbw089

Darracq A, Vitte C, Nicolas S, Duarte J, Pichon JP, Mary-Huard T, Chevalier C, Bérard A, Le Paslier MC, Rogowsky P, Charcosset A, Joets J (2018) Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. BMC Genom 19:119

Gouesnard B, Negro S, Laffray A, Glaubitz J, Melchinger A, Revilla P, Moreno-Gonzalez J, Madur D, Combes V, Tollon-Cordet C, Laborde J, Kermarrec D, Bauland C, Moreau L, Charcosset A, Nicolas S (2017) Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank. Theor Appl Genet 130: 2165–2189

Hirsch C, Hirsch CD, Brohammer AB, et al (2016) Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28:tpc.00353.2016. https://doi.org/10.1105/tpc.16.00353

Hirsch CN, Foerster JM, Johnson JM, et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell Online tpc.113.119982. https://doi.org/10.1105/tpc.113.119982

Hurgobin B, Edwards D (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? Biology 6:21. https://doi.org/10.3390/biology6010021

Lai J, Li R, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030. https://doi.org/10.1038/ng.684

Schrider DR, Hahn MW (2010) Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications. Mol Biol Evol 27:103–111. https://doi.org/10.1093/molbev/msp210

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez LA, Barbazuk A, Jeddeloh JA, Nettleton D, Schnable PS, Ecker JR (2009) Maize inbreds exhibit high levels of copy number variation (cnv) and presence/absence variation (pav) in genome content. PLoS Gen 5(11): e1000734

Swanson-Wagner RA, Eichten SR, Kumari S et al (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20:1689–1699. https://doi.org/10.1101/gr.109165.110

Tenaillon MI, Charcosset A (2011) A European perspective on maize history. Comp Rendus Biol 334(3): 221–228

The Maize Pan-Genome

2

Alex B. Brohammer, Thomas J. Y. Kono and Candice N. Hirsch

Abstract

The pan-genome of a species is comprised of genes/sequences that are present in all individuals in the species (core genome) and genes/sequences that are present in only a subset of individuals within the species (dispensable genome). In maize, the study of the pan-genome began in the 1940s through cytogenetic experiments and has seen an increased focus in research over the last decade largely driven by advances in genome sequencing technologies. It is estimated there are at least 1.5x as many genes in the pan-genome (greater than 60,000 genes) as there are in any individual's genome (~40,000 genes), with even more variation outside the gene space being observed. This variation has been associated with phenotypic variation and is hypothesized to be an important contributor to the high levels of heterosis often observed in maize hybrids. Due to the high level of variation and the existing genetic and genomic resources, maize has become a model species for plant pan-genomics studies. This chapter will review the mechanisms that can create genome content variation, tools that are available to study the pan-genome, the history of maize pan-genome research ranging from the early cytogenetic studies to today's genomics-based approaches, and the functional consequences of this variation.

2.1 Introduction

By definition, the pan-genome refers to the non-redundant set of sequences distributed throughout the population of a particular species. A pan-genome consists of two sets of sequences: those present in every individual in the population, the core genome, and those present in only a subset of individuals, the dispensable genome. The dispensable genome can be further partitioned based on a frequency spectrum. Genes present in low frequencies are part of the "cloud" set, while those in intermediate and high frequencies are part of the "shell" and "soft core" sets, respectively (Koonin and Wolf 2008).

The concept of a pan-genome was introduced by the bacterial community to describe the extensive variation in genome content between species (Tettelin et al. 2005; Medini et al. 2005; Hogg et al. 2007; Tettelin et al. 2008). Technological advances and reduced sequencing technology costs have permitted the pan-genome concept to be extended beyond bacterial species to the plant and animal kingdoms (Li et al. 2010; Computational Pan-Genomics Consortium

A. B. Brohammer · T. J. Y. Kono · C. N. Hirsch (✉)
Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul 55108, MN, USA
e-mail: cnhirsch@umn.edu

J. Bennetzen et al. (eds.), *The Maize Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-319-97427-9_2

2016). Within the plant kingdom, pan-genome analyses have been applied to a number of model and crop species such as *Arabidopsis thaliana* (Cao et al. 2011; 1001 Genomes Consortium 2016), *Brachypodium distachyon* (Gordon et al. 2017), *Brassica oleracea* (Golicz et al. 2016), *Glycine soja* (Li et al. 2014), maize (*Zea mays*; Hirsch et al. 2014), *Medicago truncatula* (Zhou et al. 2017), *Oryza sativa* (Yao et al. 2015), soybean (*Glycine max*; Anderson et al. 2014), and wheat (*Triticum aestivum*; Montenegro et al. 2017).

Depending on the number of genomes that need to be surveyed to capture the full suite of dispensable genes in a species, a pan-genome can be considered open or restricted. The former is common of bacterial species, where with each additional genome that is sequenced new genes are added to the species pan-genome (Tettelin et al. 2008). In contrast, restricted genomes like maize are typical of plant and animal species, where the majority of the pan-genome is captured in a relatively limited set of individuals. In maize, through a transcriptome-based analysis it was estimated that approximately 350 lines were needed to capture the suite of dispensable genes transcribed in the seedling (Hirsch et al. 2014).

Genome content variation in pan-genomes is often described in the context of gene copy number variation (CNV) and gene presence/absence variation (PAV). Copy number variation describes the situation in which additional copies of a particular gene exist in one individual compared to another, and PAV is simply the extreme form of CNV, where one individual possesses one or more copies and another has zero copies of the gene. Genome content variants can result from recombination-based mechanisms, replication-based mechanisms, or other molecular mechanisms and can be divided into two broad categories based on whether they lead to a balanced or unbalanced outcome. This chapter will expand on these mechanisms that generate genome content variation in plant pan-genomes, tools to measure genome content variation, historical and contemporary knowledge on the maize pan-genome, and the functional importance of this variation in driving phenotypic variation within the species.

2.2 Mechanisms that Generate Genome Content Variation

2.2.1 Transposable Elements

Transposable elements (TEs) are genomic elements that have the ability to move in the genome either through a copy-and-paste or cut-and-paste mechanism. Transposable elements were first identified by Barbara McClintock through studying disruption of pigments in maize kernels (McClintock 1950) and comprise approximately 85% of the maize genome (Schnable et al. 2009). In addition to having direct effects on protein-coding sequence and transcript regulation (Tenaillon et al. 2010), TEs also provide multiple avenues for generation of genome content variation. Some classes of TEs "capture" and shuffle gene fragments or entire genes during transposition such as Pack-MULEs and *Helitrons*. Additionally, TEs are a form of dispersed homologous sequence throughout the genome, which can lead to ectopic recombination and the generation of novel gene sequences (Bennetzen and Wang 2014). Finally, the presence of TEs can stimulate meiotic recombination, presumably through the generation of transposase-induced double-strand breaks (Yandeau-Nelson et al. 2005). Subsequent error-prone repair of these breaks then provides further opportunity for genome content variation.

2.2.2 Unequal Recombination

Unequal recombination occurs when homologous chromosomes do not pair exactly during meiosis, and recombination results in gametes with differing DNA content. This is particularly prone to occur in regions of the genome that are already duplicated, because paired sequences may be locally homologous, but may not be globally homologous. Recombination between these improperly paired chromosomes then

generates some gametes with more DNA than the progenitor cell, and some gametes with less DNA. Genes arranged in tandem duplicate arrays are common in maize (Messing et al. 2004; Schnable et al. 2009) and provide opportunities for genome content variation via unequal pairing and recombination of duplicated sequences. For example, the *A1-b* locus in maize is a naturally occurring tandem duplication of the anthocyaninless1 (*a1*) gene that has been well characterized for unequal recombination (Yandeau-Nelson et al. 2006). In this case, unequal pairing of the duplicated genes occurred preferentially between homologous chromosomes, but could also occur between sister chromatids. Unequal recombination rates at the duplicated locus were similar to equal recombination rates at non-duplicated *a1* loci, suggesting that unequal recombination is a common phenomenon at this locus.

2.2.3 Non-allelic Homologues

Similarly to unequal recombination, segregation of single-copy homologues in non-allelic positions can also lead to changes in gene copy number in the genome (Emrich et al. 2007). Mating between two individuals carrying single-copy homologues in non-allelic positions will result in progeny that are hemizygous for each of the homologues. Independent assortment, or meiotic recombination if the homologues are physically linked, generates gametes that have variable copy number for the homologues. Inbred progeny produced from these gametes then have zero, one, or two copies of the non-allelic homologues, resulting in apparent *de novo* copy number variation. An example of this phenomenon in maize is two loci involved in elongation of fatty acid precursors for surface lipids, gl8a and gl8b. These two loci are unlinked paralogs with 96% nucleotide sequence identity in B73 that can form *de novo* copy number variation (Dietrich et al. 2005). On a genome-wide scale, several dozen genes were documented to be non-allelic homologues in a single recombinant inbred line population that showed apparent *de novo* copy number variation through segregation of the non-allelic homologues (Liu et al. 2012). This *de novo* copy number variation was hypothesized to contribute to the phenotypic transgressive segregation observed in the population across a number of phenotypic traits.

2.2.4 Horizontal Gene Transfer

Horizontal gene transfer (HGT) refers to the asexual transfer of genes between organisms of divergent evolutionary lineages. Maintenance of a newly transferred gene as a segregating genome content variant depends on several events. First, the horizontally transferred gene must integrate into a cell that gives rise to gametes in order for it to be transmitted into subsequent generations. It must then not be lost due to genetic drift and provide strong enough selective advantage to be maintained in a population. As such, it is hypothesized that horizontally transferred genes that persist as segregating variation within a population have a particularly high likelihood of contributing to phenotyping variation.

Horizontal gene transfer was first observed in bacteria (Freeman 1951) and is now known to be highly prevalent among bacterial species. In bacteria, HGT occurs through random uptake of extracellular DNA, incorporation of viral DNA into the host genome, or direct transfer of plasmids among individuals (Syvanen 2012). While rare in plants, HGT has been observed via viral DNA repeats in *Nicotiana tabacum* (Bejarano et al. 1996). Expressed transfer DNAs from *Agrobacterium rhizogenes* have also been observed in cultivated sweet potato (Kyndt et al. 2015). Plant-to-plant HGT has also been documented in parasitic species. For example, a nuclear gene in *Striga hermonthica*, a hemiparasitic plant that can cause devastating crop loss in species such as *Sorghum bicolor*, has been found to have high similarity to genes from *S. bicolor*, suggesting HGT as an origin for this gene in *S. hermonthica* (Yoshida et al. 2010).

2.2.5 Genome Duplication and Fractionation

When a genome undergoes a whole genome duplication event, it generates four copies of each nuclear gene where there were previously just two. New mutations can then begin to cause the function of the duplicates to diverge. Under classical models, the net direction of molecular evolution will be toward the ancestral state of two functional copies of each gene. Three major paths to this outcome are that one duplicate evolves a new function (Ohno 1970), the copies are retained and each partially loses function (Force et al. 1999), or one of the copies completely loses function (Jacq et al. 1977). Following a whole genome duplication, the most common mechanism to restore the ancestral diploid function is through fractionation (Langham et al. 2004; Tang et al. 2008).

An ancient genome duplication event in the ancestor of maize resulted in two subgenomes in present-day maize. Analysis of the B73 reference genome assembly showed that one subgenome has greater gene retention than the other, and these subgenomes were named "Maize1" and "Maize2," respectively (Schnable et al. 2011). The paralogs lost during fractionation are not completely consistent between individuals within the species and this variation in gene loss during fractionation generates genome content variation within the species (Brohammer et al. 2018). Many genes that show presence/absence variation within maize also show sequence similarity to genes in closely related grass species (Hansey et al. 2012; Hirsch et al. 2014). This suggests that these genes were present before the divergence of the maize lineage from other grass species and were differentially lost among maize individuals.

2.3 Contemporary Tools to Measure Genome Content Variation

2.3.1 Reference-Based Methods

Reference-based methods used to measure genome content variation within species include oligonucleotide arrays and next-generation sequencing (NGS) read mapping. Oligonucleotide arrays were the first reference-based method used for conducting genome-wide surveys of genome content variation within maize (Springer et al. 2009; Beló et al. 2010). A specific technique called array-based comparative genomic hybridization (aCGH) was particularly important to advancing our knowledge of PAV and CNV in maize. In this method, two labeled DNA samples are hybridized to probe sequences designed to target regions throughout the genome, and signal intensity from each labeled sample indicates its relative copy number. A major limitation to aCGH, and arrays in general, is the inability to detect sequences absent from the reference genome since probes are often designed from a single reference individual. Related issues brought about by limitations of probe design from a single reference individual include biased CNV detection toward deletion discovery and a reduced ability to evaluate regions of high sequence diversity.

Unlike aCGH, NGS methods allow for the discovery of the full suite of structural variants within the species including sequences outside the reference genome (Young et al. 2016). There are three common NGS structural variant detection methods: read depth, split read, and read pair. The read-depth method relies on sequence read depth from mapping reads to a reference genome assembly as a proxy for copy number. Both the split-read and read-pair methods take advantage of imperfect mapping to identify genomic rearrangements and allow for the detection of all structural variant classes, including inversions and translocations. Paired-end and mate-pair sequence reads have an expected insert size between the two sets of reads. Deviation from these expected distances between the two reads can be used to identify structural variations. The read-pair method uses reads whose distance or orientation between mapped reads from the same fragment is discordant with the reference genome to detect structural variation. The split-read approach to structural variation detection uses information from paired-end sequence reads where one of the

pairs maps accurately while the other pair maps only partially or fails to map entirely. The split-read approach can also be expanded to splitting an individual read and identifying reads in which only a portion of the read can accurately map to the reference genome as another method to identify structural variation.

Each method of NGS structural variation detection has its own set of biases (Alkan et al. 2011), and each has variable sensitivities. Many of the available structural variation callers were originally developed to work with human cancer data or model mammalian species and may provide unreliable results or require extensive knowledge and tuning of parameters to be properly used with plant genomes. Combining at least two of these structural variation detection methods into a hybrid structural variation caller (i.e., SURVIVOR; Jeffares et al. 2017) that reports consensus structural variations can overcome some of these issues. Additionally, some of these methods rely on imperfect read mapping, which can be prevalent when mapping short NGS reads to highly repetitive plant genomes even in the case of reference genome reads mapping to the reference genome assembly. Increased read coverage and optimization of mate-pair library sizes can mitigate this challenge; however, long-read sequencing technologies offer the most promise for avoiding inconsistent structural variation detection in repetitive regions and for the detection of large structural variants.

2.3.2 Non-Reference-Based Methods

With reference-based variant detection, there is an ascertainment bias that is caused by the reliance on a single reference genome assembly. One method for characterizing gene content variation beyond a single reference genome assembly is through direct comparison of multiple *de novo* genome assemblies. Schatz et al. demonstrated the power of this approach by generating *de novo* genome assemblies of *indica*, *aus*, and temperate *japonica* rice strains, where they identified several megabases of variable sequence between the three strains (Schatz et al. 2014). This approach has also been used in maize where approximately thousands of novel genes were identified in a comparison of *de novo* genome assemblies of elite inbred lines from opposite heterotic groups (Hirsch et al. 2016; Darracq et al. 2018).

Direct comparison of whole genome *de novo* assemblies allows for detailed analysis of variation outside of a single reference genome; however, a major disadvantage is the cost and computational effort required to bring these studies to fruition. This disadvantage is important for pan-genome studies because it often leads to a small number of genotypes being assayed and an underestimate of dispensable genome content within species. An alternative approach is to use the transcriptome as a proxy to evaluate the gene space within a species pan-genome. This approach has the advantage of reducing both the amount of sequencing and computation required in pan-genome studies. In maize, the gene space is only ~97 MB of the genome, and as such, this approach was able to be used to study the maize pan-genome using over 500 accessions (Hirsch et al. 2014).

Recent improvements in assembly algorithms and the continued decline in sequencing costs are making multiple *de novo* genome assemblies within a species more practical (Schatz et al. 2014; Wetterstrand 2018). An example of this shift toward the generation of *de novo* genome assemblies for pan-genome analysis is the assembly and annotation of a panel of 54 *Brachypodium distachyon* accessions by Gordon and colleagues (Gordon et al. 2017). For seven years, only two reference genome assemblies for maize were available: the B73 reference genome, and Palomero Toluqueño, a popcorn landrace (Vielle-Calzada et al. 2009). In the span of just three years, nine additional genome assemblies were made publicly available (W22—GenBank assembly accession GCA_001644905.2; F7 and Ep1—(Unterseer et al. 2017); PH207—(Hirsch et al. 2016); B73—(Jiao et al. 2017); F2—(Darracq et al. 2018); Mo17, B104, and CML247 (Maize Genetics and Genomics Database 2017)).

New and emerging technologies that provide long-range information will help to further