ANNUAL REVIEWS

Annual Review of Plant Biology

Plant Pan-Genomics Comes of Age

Li Lei,[1] Eugene Goltsman,[1] David Goodstein,[1] Guohong Albert Wu,[1] Daniel S. Rokhsar,[1,2] and John P. Vogel[1,3]

[1]DOE Joint Genome Institute, Berkeley, California 94720, USA; email: JPVogel@lbl.gov

[2]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

[3]Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA

Annu. Rev. Plant Biol. 2021. 72:411–35

First published as a Review in Advance on April 13, 2021

The *Annual Review of Plant Biology* is online at plant.annualreviews.org

https://doi.org/10.1146/annurev-arplant-080720-105454

Keywords

pan-genome, population genetics, structural variation, sequence graph, natural variation

Abstract

A pan-genome is the nonredundant collection of genes and/or DNA sequences in a species. Numerous studies have shown that plant pan-genomes are typically much larger than the genome of any individual and that a sizable fraction of the genes in any individual are present in only some genomes. The construction and interpretation of plant pan-genomes are challenging due to the large size and repetitive content of plant genomes. Most pan-genomes are largely focused on nontransposable element protein coding genes because they are more easily analyzed and defined than noncoding and repetitive sequences. Nevertheless, noncoding and repetitive DNA play important roles in determining the phenotype and genome evolution. Fortunately, it is now feasible to make multiple high-quality genomes that can be used to construct high-resolution pan-genomes that capture all the variation. However, assembling, displaying, and interacting with such high-resolution pan-genomes will require the development of new tools.

Contents

INTRODUCTION

Reference genome: a representative genome that serves as the benchmark for an entire species; typically much effort is expended to produce the most complete assembly and annotation possible

Structural variants (SVs): sequence variants >50 bp in size; the most recognized forms include presence/absence variants, copy number variants, inversions, and translocations

Presence/absence variant (PAV): a deletion or insertion relative to a reference genome

Copy number variant (CNV): a duplicated or deleted sequence; PAVs are an extreme form of CNVs

Pan-genome: the nonredundant set of genes or genomic sequences within a species

In the classical view of intraspecific genetic variation, the genome of each individual is described as a set of small variations on a common reference genome. In recent years, however, comparative analysis of genomes or genome segments from multiple individuals of the same species has revealed that a single reference genome is inadequate to capture the genetic diversity of a species. These findings demand a more expansive view of genetic variation, in which genomes within a species may differ in more dramatic ways, including a diversity of structural variants (SVs) that, strikingly, can contain one or more genes. This in turn implies that the functional gene content of a species is far more variable than previously imagined.

Practically, while aligning sequence reads from many individuals to a reference genome effectively identifies small polymorphisms [single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels)], this approach completely misses longer (>50 bp) stretches of novel sequences not found in the reference genome as well as sequences that are highly divergent from the reference genome (24, 25). These larger SVs include presence/absence variants (PAVs) and copy number variants (CNVs) (**Figure 1**). Significantly, SVs can be large enough to contain genes, sometimes many genes. While numerous studies demonstrate the critical role SVs can play on agronomically important traits (e.g., resistance to biotic stress and abiotic stress, flowering time, plant architecture, yield, and grain or fruit quality) (reviewed in 76), we do not have a firm grip on the spectrum, prevalence, and mechanisms for the formation for SVs in plants. This has motivated the construction of plant pan-genomes.

Most simply, the pan-genome can be defined as the nonredundant set of genes within a species (**Figure 2**). However, this gene-based definition misses all the diversity of noncoding sequences and the locations of repetitive elements like transposable elements (TEs), some of which may be important for gene regulation and phenotypic expression in eukaryotes, as well as gene-preserving SVs including inversions and translocations. A more inclusive pan-genome definition would be the nonredundant set of genomic sequences within a species (**Figure 2*c***). Unfortunately, as discussed in detail below, this more broadly defined sequence-based pan-genome is much more difficult to construct, interpret, and display than a gene-based pan-genome with its discrete units of known biological meaning (**Figure 2**). Due to technical challenges, the plant pan-genomes constructed to date are primarily gene-based. Recent improvements in DNA sequencing, however, are greatly decreasing the cost of assembling very high-quality genomes, removing one of the barriers to constructing useful sequence-based pan-genomes for plants.

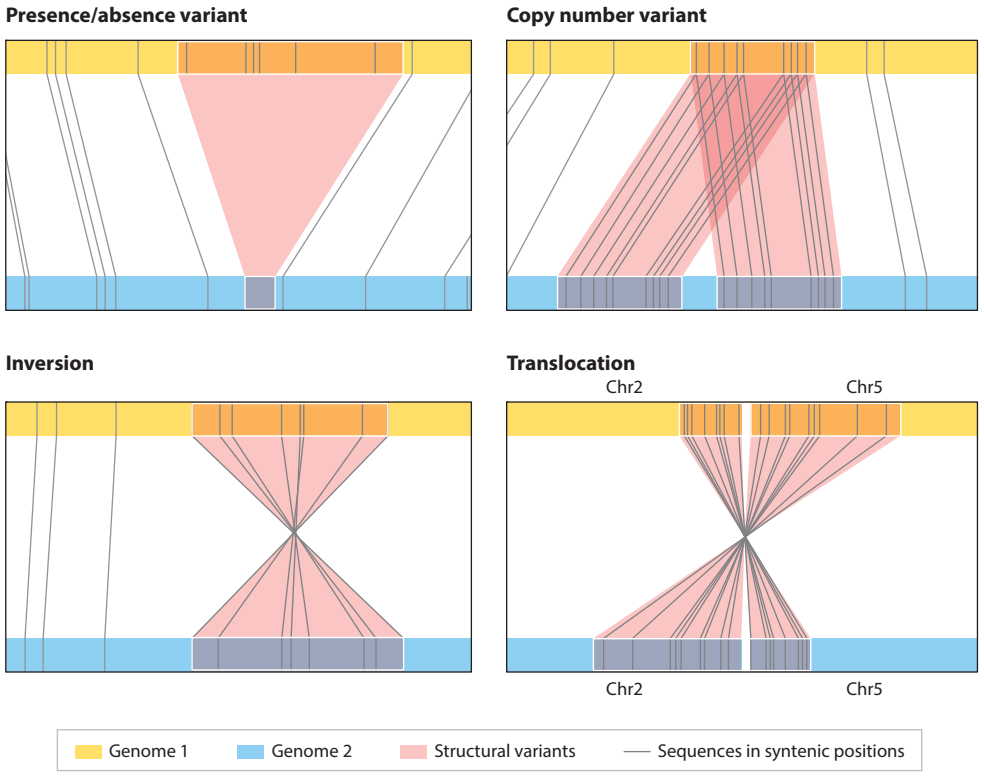


Figure 1

Structural variants (SVs). Genomic variation ranges from single nucleotide polymorphisms (SNPs) to chromosomal translocations. Smaller variants, SNPs, and insertions/deletions less than 50 bp are readily detected by mapping raw reads to reference genomes. In addition, larger deletions and regions of high sequence divergence can be detected as regions where no reads map to the reference genome; however, larger insertions are invisible to read mapping approaches. The detection of larger SVs (here defined as variants greater than 50 bp) is difficult and imprecise using read mapping approaches. Fortunately, SVs can be detected by comparing de novo assemblies, and the accuracy increases with the quality of the assemblies. There are four main types of SVs: presence/absence variants, copy number variants, inversions, and translocations. SVs cover an enormous size range up to many megabases of sequences. While SVs are typically less abundant than smaller polymorphisms, they affect many more nucleotides per genome due to their large size. SV detection can be confounded by sequencing errors, chimeric assemblies, and repetitive DNA, so care must be taken when interpreting these variants. Note that whether a variant is noted as present or absent depends on which genome is considered as the reference.

This review discusses the concept and nomenclature of pan-genomes and provides an overview of the construction, use, and limitations of the plant pan-genomes created to date. In addition, the difficulties of creating, studying, and interacting with plant pan-genomes are explored along with how recent advances may help overcome these difficulties.

DEFINING PAN-GENOMES

The first pan-genome was constructed in 2005 from eight strains of the bacterial species *Streptococcus agalactiae* by comparing de novo assemblies and annotations (78). In this study, the authors divided the pan-genome into two fractions: core and dispensable. The core genome consists of

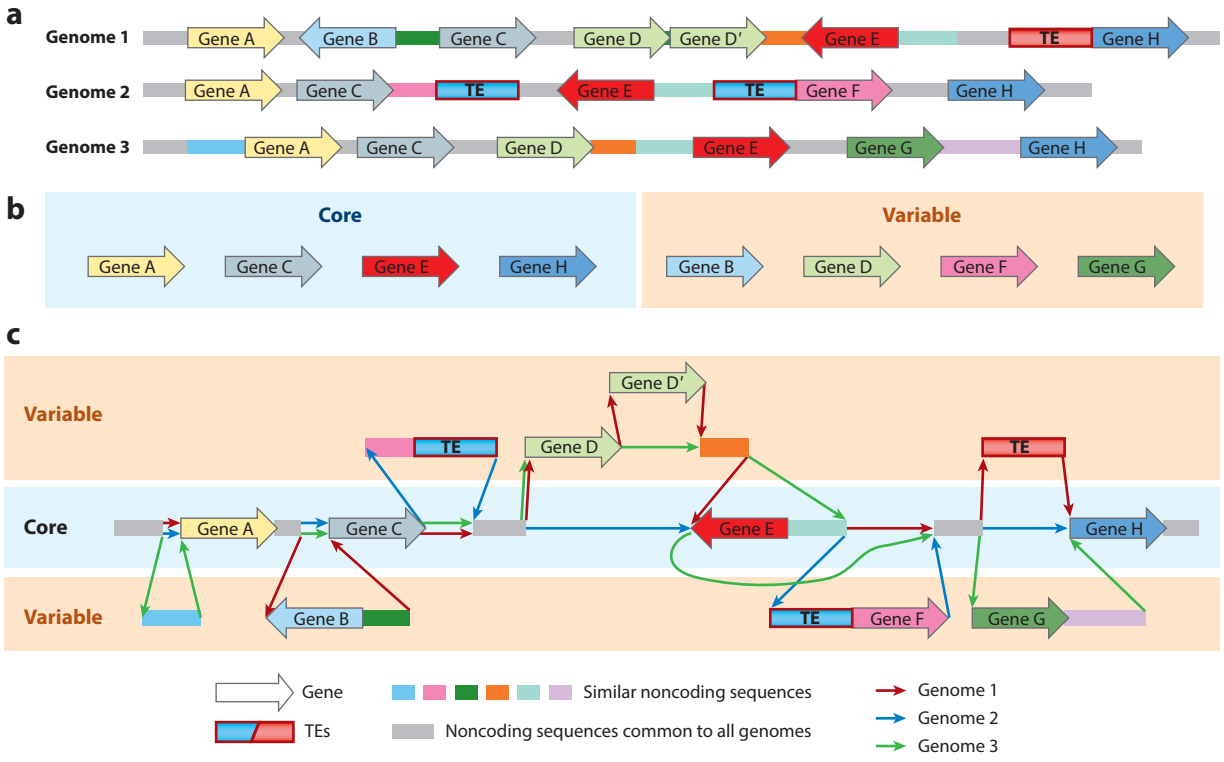


Figure 2

Construction of gene- and sequence-based pan-genomes. (*a*) Syntenic genomic segments from three genomes. Introns and small polymorphisms (SNPs and indels) are omitted for clarity. Note that gene D has a recent tandem duplicate, D′, with nearly identical sequence. Also note that gene E and its 5′ region are inverted in genome 3. (*b*) Gene-based pan-genome constructed by clustering genes (coding sequences) according to similarity. This approach does not capture noncoding sequences or TEs. Notice that gene D′ has been collapsed with gene D because they share very similar sequence. (*c*) Sequence-based pan-genome represented as a sequence graph (see **Figure 3** for sequence graph construction). The nonredundant set of unique sequences from panel *a* is separated into core and variable fractions. The sequence of any individual line can be reconstructed by following the color-coded arrows: red for genome 1, blue for genome 2, and green for genome 3. Note the inversion of gene E is captured by the direction of the green arrows. This approach captures all sequences and their relationships. Theoretically, all polymorphisms from SNPs to translocations can be represented by sequence graphs; however, in practice, such graphs become too computationally intensive for existing algorithms when applied to many genomes. Abbreviations: indels, insertions and deletions; SNPs, single nucleotide polymorphisms; TEs, transposable elements.

Variable gene: a gene absent from one or more lines, also called a distributed, accessory, or dispensable gene; sometimes further divided into categories based on how many lines a gene is found in

the genes present in all strains, while the dispensable genome consists of genes absent from one or more strains (**Figure 2*b***). The dispensable genome was further divided into what the authors called "accessory genes" (found in at least two strains and missing from at least one strain) and "unique genes" (found in only one strain). Significantly, the *S. agalactiae* pan-genome was much larger than any individual genome within this bacterial species. In this review we use the term variable rather than dispensable to denote genes found in only some lines because dispensable implies that the organism could survive without such genes and this is almost certainly not the case (48). Simultaneous loss of multiple members of variable gene families may be lethal, and numerous epistatic interactions are also expected.

A practical limitation of binary terminology (e.g., core versus variable) is its sensitivity to assembly and/or annotation errors, which are inevitable in large genome collections. For example, if a true core gene is missing from one genome because of error, it will be mistakenly counted as a

variable gene. This becomes a significant problem as the number of genomes increases. Taken to the extreme, if enough genomes were sequenced, there could be no core genes identified simply due to technical errors. To correct for this kind of error, some authors divide the pan-genome into additional fractions to represent how many lines a gene is present in. For example, Koonin & Wolf (37) divided bacterial pan-genomes into four fractions: core genes that are found in all genomes, soft-core genes that are found in almost all lines, shell genes that are found in fewer lines than the soft-core but more than just a few lines, and finally cloud genes that are found in only a few lines. This nomenclature has also been adopted for two plant pan-genomes (17, 23) and has the advantage of grouping genes in such a way that robust comparisons can be made. For example, comparing core and shell genes will be a true comparison between genes found in all lines and genes that are truly missing in some lines without the dilution caused by technical errors such as missed core genes or artifactual genome annotations.

Core gene: a gene present in all lines (some publications define a core gene as present in almost all lines)

The widespread use of pan-genomes is hindered by the lack of standardized ways to define, construct, describe, or visualize them. Thus, comparisons between pan-genomes are fraught, and all of the assumptions and thresholds that went into making each pan-genome must be taken into account during any comparisons. Indeed, even when authors use the same nomenclature, the meaning may be different. For example, while most studies define the core genome as the set of genes found in all lines, some studies include genes found in an overwhelming majority of the lines, e.g., 95% (30). Thus, care must be exercised when comparing results from different groups as well as when interpreting the biological meaning of these results.

While a gene-based pan-genome is conceptually straightforward, it raises the question, When are two loci divergent enough to be considered different genes? It is common to group sequences within and between genomes by similarity to define groups of orthologous genes, but this can be especially problematic in plants where one often finds large gene families whose members have evolved over varying timescales from a common ancestral gene. Another consideration is how heavily, if at all, to weigh synteny when deciding which genes in two lines are the same gene. For example, two consecutive similar genes in individuals 1 and 2 could represent two distinct, tandemly duplicated genes, inherited from a common ancestor, or parallel duplications. Recent duplicates may have nearly identical sequences, which leads to cases where there are clearly two genes in one line and one gene in another. Is a gene a discrete physical entity, or is a gene really the known or potential function of the DNA sequence? If one considers genes as discrete transmissible DNA sequences, then recent duplicates should be counted as two genes represented in the pan-genome.

A more conservative approach would be to consider genes with nearly the same sequence, and presumably the same or similar function, as a single pan-gene. Using this approach in the simple example above, a cluster of tandemly duplicated genes would be represented as a single pan-gene. However, all genes in the cluster contribute DNA sequences to the pan-gene cluster capturing population-level variation. This clustered pan-gene approach has the advantage that it is not confused by gene duplication and/or gene movement, and most plant pan-genomes constructed to date use clustering based on similarity to construct a gene-based pan-genome. However, the clustering process inevitably leads us back to the question of when two sequences are divergent enough to be called different genes, and each study uses its own thresholds. As with almost everything biological, divergence is a continuum, and depending on the thresholds used, the pan-genome will be smaller or larger. Thus, as a practical matter, creating a pan-genome requires a number of assumptions and thresholds, and these must be taken into account when interpreting (and especially when comparing) published results, and authors should be up front about the implications of their methods.

While most pan-genomes constructed to date are primarily gene-based because of the relative ease of comparing and categorizing discrete units defined by transcription and translation, the importance of noncoding and repetitive sequences is unquestionable. It would therefore be extremely powerful to define a comprehensive sequence-based pan-genome that includes information about the relative position of all sequences. Unfortunately, interpreting noncoding sequence variation is challenging. Indeed, even for classes of noncoding sequences of known importance, e.g., promoters, our ability to predict and define their boundaries is poor. Thus, there is no straightforward way to functionally partition and compare nongenic sequences as there is for genes. Repetitive sequences like TEs present additional challenges, as they may be found in the genome in various stages of mutational decay. Nevertheless, these sequences play important roles in regulating gene expression, as was elegantly demonstrated by an analysis of TEs and grape color (35). The mobility of these elements makes their inclusion in pan-genomes all the more important because they are responsible for much of the intraspecific sequence diversity (6, 47).

SOURCES OF STRUCTURAL VARIATION IN PAN-GENOMES

Genome size is determined by the net effect of processes that expand and contract DNA sequences. Differences in these processes between individuals create the structural variation observed in pan-genomes. Thus, an understanding of these processes is useful for understanding the challenges of accurately assembling and interpreting pan-genomes. The primary mechanisms contributing to genome expansion are polyploidization (whole-genome duplication), segmental duplication, proliferation of repetitive DNA sequences, and unequal crossing over, often fostered by repetitive DNA (56, 63, 85). The main mechanisms contributing to genome contraction include TE-mediated unequal homologous recombination, illegitimate recombination (nonhomologous recombination), and deletion-biased double-strand break repair (reviewed in 69).

It is well known that all plants are derived from ancient polyploidization events, and many lineages have undergone more recent polyploidization events (67, 84). Polyploidization has two important consequences for pan-genomic SVs: (*a*) the absence of selection pressure to prevent the loss of some redundant genes and sequences and (*b*) recombination-mediated exchanges between nonhomologous chromosomes leading to the doubling or loss of chromosome fragments (85). The relative contribution of these mechanisms may vary between species. For example, in *Arabidopsis*, illegitimate recombination was the main driving force behind its decreased genome size, whereas deletion-biased double-strand break repair contributed to genome shrinkage in the carnivorous plant *Genlisea nigrocaulis* (82). This is relevant to pan-genomic SVs because individuals that share the same polyploidization event may retain, add, or lose different genomic segments.

In addition to polyploidy, several potential mechanisms can cause the de novo formation of SVs, including illegitimate recombination between copies of repeated sequences such as TEs, transfer RNA genes, or segmental duplications; rearrangements associated with DNA repair by nonhomologous end joining; microhomology-mediated break-induced replication, contraction, or expansion of variable number tandem repeats; and mobile element insertions (reviewed in 26, 29). For example, many of the CNVs in rice have been attributed to TE insertions and nonhomologous end joining (4). Some inversions in maize have been demonstrated to be caused by illegitimate recombination and nonhomologous end joining (34, 46, 91, 93, 98). Whatever the cause, the SVs in the variable genome cause extensive genic PAVs. A substantial proportion of genes are affected by SVs directly or indirectly, and those genes can affect many processes, including biotic stress and abiotic stress tolerance, flowering time, plant architecture, yield, and grain or fruit quality (reviewed in 76).

In addition to gene loss, some of the above-listed mechanisms of SVs in the pan-genome may also give rise to new genes. For example, although both whole-genome and smaller duplications are often followed by gene loss, some of the retained genes can evolve new functions due to relaxed selection pressure on homoeologous genes. Importantly, proteins coded by TEs can be co-opted (i.e., domesticated) by the host as an adaptation to evolutionary conflict (32). A classic example of TE domestication is found in *Drosophila*, where the integrity of the chromosome telomeres is maintained not by telomerase but by two non–long terminal repeat retrotransposons (59). Eukaryotes may also acquire genes from prokaryotes via lateral gene transfer, though the occurrence may be rare and restricted (39).

CHALLENGES TO CREATING AND INTERPRETING PLANT PAN-GENOMES

By 2015, numerous studies had demonstrated that large prokaryotic pan-genomes are the rule (81). The degree to which this could be extrapolated to eukaryotes, however, was not clear. The large differences between the biology of prokaryotes and eukaryotes could conceivably influence the size of the pan-genome. In particular, the very concept of a species is not directly equivalent, the mechanism of genetic exchange between individuals of the same species is radically different, and the much higher frequency of horizontal gene flow in bacteria would be expected to dramatically influence the size of the pan-genome. Furthermore, the large size and nontrivial ploidy of plant genomes create major challenges in constructing and interpreting plant pan-genomes, including sequencing cost, assembly quality, annotation variability, and visualization and mining.

The notion of constructing plant pan-genomes was unimaginable before the advent of inexpensive short-read sequencing technology made it feasible to sequence multiple lines from a single species. Nevertheless, due primarily to the repetitive nature of plant genomes, it is impossible to assemble highly contiguous plant genomes from short reads alone. Creating high-quality assemblies is even more challenging for outbred and/or polyploid species due to the need to separate haplotypes of heterozygous individuals. Thus, short-read sequencing technology is most suited to studying gene space and inbred plants with small genomes and is the basis of virtually all plant pan-genomes published to date. The long reads produced by Pacific Biosciences (PacBio) sequencing technology span tens of kilobases and typically contain complete genic haplotypes. As a result, they are extremely powerful for assembling highly contiguous plant genomes because they can span repetitive regions and preserve haplotype phasing across genes. Recent decreases in the cost of PacBio sequencing and the higher accuracy of circular consensus sequencing for PacBio have made it feasible to produce highly accurate and contiguous whole-genome assemblies of multiple individuals from the same species. This will usher in a new wave of high-confidence plant pan-genomes that include noncoding sequences, repetitive sequences, and larger SVs (e.g., inversions) (47).

Given a collection of assembled genomes, the next step in interpretation is annotation, or the identification of functional features such as transcribed sequences and exon boundaries. Genome annotation is still very much an art, and the results of different annotation methods applied to the same genome can differ by thousands of genes. This layer of technical variability further complicates the comparison and interpretation of plant pan-genomes, especially when evaluating the size and significance of the core genome. Thus, when constructing gene-based pan-genomes, it is highly desirable that the same annotation protocol be used for all genomes including, wherever possible, using the same transcriptional supporting data (e.g., transcriptome sequencing) as inputs for each genome. Further complicating matters, we lack the ability to reliably predict and

define regulatory regions and control elements that, for the most part, are completely ignored by annotation pipelines, greatly limiting the utility of the resulting pan-genomes.

In addition to the technical difficulties of creating pan-genomes, the tools for displaying and interacting with pan-genomes present additional challenges. Existing genome browsers and databases, with their focus on linear reference genomes, are ill-suited to display the complexity inherent in pan-genomes created from many individual genomes. This is especially problematic for sequence-based pan-genomes. As described in detail below, new tools and approaches are needed to fully utilize the increasingly accurate pan-genomes being produced.

PLANT PAN-GENOMES

Several plant pan-genomes have been constructed (**Table 1**) using a variety of approaches that resulted in pan-genomes of varying degrees of completeness and accuracy. However, they all indicate that the pan-genome is substantially larger than any individual genome and that a large fraction of genes in any individual are variable genes. The first hint that plant pan-genomes could be much larger than individual genomes came from studies that mapped individual sequence reads to reference genomes. Read mapping easily identifies SNPs and small indels (less than the length of individual read, typically 50 bp) in large numbers of plant genomes (e.g., 25, 50). It can also identify larger deletions or highly divergent regions with respect to the reference genome as areas in which few reads map. This approach has been used by numerous studies to identify substantial amounts of reference sequence that were missing or highly divergent in nonreference genomes. For example, examination of two *Arabidopsis* lines revealed 3.4 Mb of sequence that was deleted or divergent with respect to the reference genome (57), a survey of six divergent lines of the grass *Brachypodium distachyon* identified 10–21 Mb of deleted or highly divergent DNA relative to the reference genome (24), and a survey of 302 soybean lines identified 6,388 deletions comprising 73.6 Mb with respect to two reference genomes (97). Since there is nothing special about the lines used to create the reference genomes, presumably there are similar amounts of nonreference sequence in each nonreference line indicating that the pan-genomes for these representative monocot and dicot species could be considerably larger than any individual genome. In this section, we highlight several plant pan-genome studies as examples of the different approaches ranging from early efforts, which made compromises to minimize sequencing cost, to the most recent studies, which are based on highly contiguous and accurate genome assemblies.

To overcome the size and complexity of the maize genome and estimate the size of the pan-genome, Hirsch et al. (28) used a pan-transcriptomic approach. They sequenced messenger RNA from the seedlings of 503 maize lines and assembled pooled reads that did not map to the reference genome. After removing transcripts that were similar to reference genes (>85% identity over 85% of the length), they identified 8,681 expressed genes that were not contained in the reference genome. While this approach is limited in that it only captures genes expressed in seedlings and is completely blind to noncoding sequences, it did demonstrate that the maize pan-genome was substantially larger than the reference genome. Maize possesses a large and notoriously variable genome, so learning whether these results could be generalized had to wait for studies in plants with less dynamic genomes.

A light shotgun-sampling metagenomic approach that combines low-coverage sequence from many lines prior to assembly is a cost-effective approach that takes advantage of identity by descent across multiple genomes. This kind of approach has been used to create pan-genomes for rice and wheat. Yao et al. (89) constructed a rice pan-genome from 1,483 rice lines. They removed reads that aligned to the reference genome and then pooled the remaining reads according to subspecies (one pool for *indica* and one pool for *japonica*) prior to independently assembling the

Table 1 Plant pan-genomes

Species	Year	Number of individual genomes	Reference genome size (Mb)	Sequencing technology	Pan-genome construction approach	Number of nonreference genes identified[a]	Number of reference genes or average number of genes[b]	Number of genes or gene clusters in the pan-genome[c]	Pan-genome[d]/reference or average genome	Database or data access point	Reference
Zea mays	2014	503	2,106	Illumina	Pan-transcriptome	8,681[e]	22,354[e]	31,398[e]	1.4	https://datadryad.org/stash/dataset/doi:10.5061/dryad.r73c5	28
Glycine soja	2014	7	1,115	Illumina	De novo assembly	NA	55,570	59,080	1.06	NA	43
Brassica rapa	2014	3	284	Illumina	De novo assembly	3,672	39,650	43,882	1.07	http://brassicagenome.net/databases.php	45
Oryza sativa	2014	3	384	Illumina	De novo assembly	1,300	39,062	40,362	1.14	http://schatzlab.cshl.edu/data/rice/	68
Oryza sativa	2015	1,483	384	Illumina	Metagenome-like	1,913 high confidence (8,991 total) for *indica*	35,596	37,509	1.05	NA	89
Brassica oleracea	2016	10	488	Illumina	Iterative mapping and assembly	6,922	54,457	61,379	1.13	http://brassicagenome.net/databases.php	19
Brachypodium distachyon	2017	54	272	Illumina	De novo assembly	7,167 high confidence (33,329 total)	30,751	37,886 high confidence (64,084 total)	1.23 high confidence (2.08 total)	https://brachypan.jgi.doe.gov/	23
Medicago truncatula	2017	15	360	Illumina	De novo assembly	38,000	37,000	75,000	2.03	http://www.medicagohapmap.org/downloads/assemblies	95
Triticum aestivum	2017	18	17,000	Illumina	Metagenome-like	59,430[f]	128,656[f]	140,500[f]	1.09	http://wheatgenome.info/wheat_genome_databases.php	54
Oryza sativa	2018	3,010	373	Illumina	Map-to-pan	3,734	20,142	23,876	1.19	http://iric.irri.org/resources/3000-genomes-project	83

(*Continued*)

Table 1 (*Continued*)

Species	Year	Number of individual genomes	Reference genome size (Mb)	Sequencing technology	Pan-genome construction approach	Number of nonreference genes identified[a]	Number of reference genes or average number of genes[b]	Number of genes or gene clusters in the pan-genome[c]	Pan-genome[d]/reference or average genome	Database or data access point	Reference
Capsicum annuum, *Capsicum baccatum*, *Capsicum chinense*, and *Capsicum frutescens*	2018	383	3,480	Illumina	Map-to-pan	15,862 high confidence (52,286 total)	36,895[g]	51,757 high confidence (89,181 total)	1.4 high confidence (2.42 total)	**http://www.pepperpan.org:8012/index.html**	58
Oryza sativa and *Oryza rufipogon*	2018	67	336	Illumina	De novo assembly	10,872	35,596	42,580	1.2	**http://db.ncgr.ac.cn/RicePan-Genome/index.php**	94
Brassica napus	2018	53	850	Illumina	Iterative mapping and assembly	13,649	80,382	94,013	1.17	**http://brassicagenome.net/databases.php**	31
Sesamum indicum	2019	5	357	Illumina	De novo assembly	6,173	36,189	42,362	1.17	**http://www.sesame-bioinfo.org/pan-genome/**	92
Solanum lycopersicum L.	2019	725	1,179	Illumina	Map-to-pan	4,873	35,496	40,369	1.14	**https://datadryad.org/stash/dataset/doi:10.5061/dryad.m463f7k**	17
Helianthus annuus L.	2019	287	3,600	Illumina	Targeted de novo assembly	13,357	47,848[f]	61,205	1.28	**https://www.sunflowergenome.org/pangenome-data/**	30
Glycine max	2020	27	1,115	PacBio and Illumina	Graph-based de novo assembly	3,317	54,175	57,492	1.06	**https://figshare.com/s/689ae685ad2c368f2568**	47
Brachypodium hybridum D subgenome and *Brachypodium distachyon*	2020	54 (*B. distachyon*), 4 (*B. hybridum*)	500	Illumina	De novo assembly	44,004	30,096	74,100	2.46	**https://phytozome-next.jgi.doe.gov/info/Bhybridum_v1_1**	22

Abbreviation: NA, not applicable.

[a]The number of unique genes or gene clusters is reported depending on how the pan-genome was constructed.

[b]If applicable, the number reported is the number of pan-gene clusters, not the number of annotated genes.

[c]If applicable, the number of high- and low-confidence pan-genes is reported. Low-confidence genes are usually defined as genes only appearing in one or a few genomes.

[d]If applicable, the high-confidence pan-genome is used for this ratio.

[e]Only transcripts are reported.

[f]Estimated by the authors.

[g]NCBI assembly GCA_011745865.1.

reads from each subspecies. While this approach has the advantage of inexpensively sampling much more germplasm than is possible with de novo assembly of each individual line, the assemblies are, by definition, chimeric and it cannot be determined which genes came from which line. This method also misses low-abundance genes because the sequencing depth is too low, especially when polymorphisms are considered. Despite these limitations, 1,913 high-confidence nonreference protein-coding genes (out of 8,991 total nonreference protein-coding genes) were annotated for the *indica* lines, indicating that rice also has a substantial pan-genome. Significantly, the authors conducted a genetic association study that showed that 23.5% of metabolic traits were more significantly associated with nonreference genome markers than with reference genome markers and that 41.6% of SNPs associated with agronomic traits were in variable sequences. Thus, this study demonstrated the importance of variable genes and the value of pan-genomics for crop improvement. The very large hexaploid wheat genome presents challenges for pan-genome construction. Montenegro et al. (54) addressed this challenge using a metagenome approach. Based on 18 cultivars, they estimated that the full wheat pan-genome contains 140,500 genes, of which approximately 81,070 are core.

A pan-genome was constructed from 10 *Brassica oleracea* lines using an iterative mapping and assembly approach that assembled and inserted novel sequences from each line into the pan-genome before adding sequences from the next line (19). For each line, read pairs that both mapped to the current pan-genome (starting with the reference genome) were identified and eliminated from further consideration. Then the remaining read pairs, some with one read that mapped to the current pan-genome, were assembled. The resulting contigs were then inserted into the reference genome using read pairs that linked the current pan-genome to the newly assembled contigs. In this manner they constructed a pan-genome that was 20% larger and contained 3% (2,154) more genes than the reference genome. Interestingly, they noted PAVs in flowering time genes that may be responsible for the early flowering of the rapid cycling line TO1000. The same approach was used to create a pan-genome for 53 natural and synthetic accessions of the allotetraploid *Brassica napus* (31). Examination of the pan-genome revealed that homoeologous exchanges were the mechanism for some of the extensive PAVs observed. While these pan-genomes clearly identified novel pan-genome sequences beyond the reference and demonstrated the functional importance of the pan-genome, their utility is limited because the assembly and incorporation of sequence create chimeras and truncations. In addition, smaller variants (from SNPs to small indels) in each line are not captured.

A so-called map-to-pan approach was used to create pan-genomes for rice, pepper, tomato, and sunflower (17, 30, 58, 83). To construct a sequence-based pan-genome from highly fragmented de novo assemblies, first the contigs were clustered and then the largest contig from each contig cluster was added to the pan-genome, which was then annotated. PAVs in all the lines were then called by mapping raw reads to the pan-genome. This approach allows for the creation of pan-genomes from large numbers of lines with relatively low sequence coverage. However, since the initial assemblies are incomplete and unannotated, the number and type of variants that can be unambiguously assigned to individual genomes are limited. Using this approach, Wang et al. (83) created a rice pan-genome based on 3,010 accessions. They identified more than 10,000 nonreference genes, which produced a gene-based pan-genome about 25% larger than any individual rice genome. Using the same approach, Ou et al. (58) constructed a pan-genome from 383 pepper lines corresponding to four *Capsicum* species. Again, the pan-genome was considerably larger than any single genome; however, by including four species, the pan-genome constructed was also undoubtedly larger than the pan-genome for a single species. The same approach was used to construct a tomato pan-genome from 725 accessions, which identified 4,873 nonreference genes (17). A similar strategy using targeted assembly was used to construct a pan-genome from 287 cultivated

sunflower lines (30). The authors first mapped reads to the reference genome and then assembled reads that did not individually map for each accession. The resulting contigs were clustered and the pan-genome constructed. Since the sequencing was low coverage, the resulting pan-genome is not comprehensive; nevertheless, the authors identified 2,700 nonreference genes that passed all their filtering steps.

Sequence graph: a graphical representation of the alignment of multiple sequences where shared sequences are represented by nodes and relationships between sequences are indicated as edges

Confidently estimating the core and variable genome size and identifying all variants (from SNPs to large indels) in individual genomes require the de novo assembly and annotation of the genomes from a large number of lines. The first pan-genomes that used this approach were created for two model plants, *B. distachyon* and *Medicago truncatula* (23, 95). This was feasible because both species have small genomes that allow nearly complete, if still very fragmented, genomes to be assembled from short read sequences. Genomes from 15 lines were used to create the *M. truncatula* pan-genome, and genomes from 54 lines were used for *B. distachyon*. Both studies created the pan-genome by clustering genes based on sequence similarity. In *M. truncatula*, 67% of the genes in the pan-genome were variable and the pan-genome was 47% larger than the reference genome. The *B. distachyon* pan-genome showed a similarly large percentage of variable genes, with about half the genes in individual genomes being variable. Since the *B. distachyon* pan-genome used a much higher number of genomes, the authors divided the pan-genome into four fractions: core genes that were found in every line, soft-core genes that were found in almost every line, shell genes that were found in three or more lines but not included in the soft core, and cloud genes that were found in only one to two lines. They excluded the cloud genes from most analyses and focused on the remaining high-confidence pan-genome. Even with this conservative approach, the pan-genome contained 23% more gene clusters than the reference genome, and about half of the genes in any genome were variable genes. In addition to creating pan-genomes, both studies also identified all the smaller variants that can be recovered from the individual genome assemblies. Both studies also made rudimentary explorations of the role of TEs on gene expression and pan-genome evolution and found, not surprisingly, that TE insertions close to genes tended to decrease expression. While the fragmented assemblies are not ideal for detecting larger SVs, the authors of the *M. truncatula* pan-genome did catalog such variants. Encouragingly, 88–94% of the SVs in the lines examined were validated by PacBio long reads. These two studies agree with results from earlier studies on *Glycine soja* (43), rice (68), and *Brassica rapa* (45), which all used a similar approach, as well as a recent study on sesame (92), but these studies used many fewer lines and lower-quality assemblies that precluded firm conclusions about pan-genome size.

The next level of pan-genome construction is based on highly accurate and contiguous assemblies that can now be constructed at reasonable cost using PacBio sequencing. The very long (often megabase-scale) PacBio contigs can be placed into chromosome-scale scaffolds using Hi-C, optical mapping, and/or linkage maps. Theoretically, all variants, from SNPs to chromosomal translocations, can be identified by comparing these assemblies. A pan-genome for 27 soybean lines was constructed from high-quality assemblies (26 PacBio assemblies and the reference genome) (47). Similar to other pan-genomes from de novo assemblies, a gene-based pan-genome was constructed by clustering genes based on similarity. In addition, the high contiguity of the assemblies allowed them to call 776,399 SVs with respect to the reference genome. They next consolidated the SVs into 124,222 nonredundant SVs, used those to create a sequence graph (described in detail below) as a representation of the sequence-based pan-genome, and found that about 78.5% of all PAVs were located in repetitive regions. They then made another sequence graph, which excluded PAVs that contained >90% repetitive sequence. This less repetitive graph was used to call 55,402 SVs in 2,898 accessions by mapping sequence reads to the graph. This approach removes the bias inherent in calling variants with respect to a single reference genome and gives a more comprehensive picture of PAVs at the population level. It should be noted that

while sequence graphs were used as part of the analysis to show that TE insertions were correlated with changes in gene expression and variation in agronomic traits, the authors neither used the graphs to construct the pan-genome (see the section titled Sequence-Based Pan-Genomes and Graph-Based Analysis) nor released the graphs in an easily accessible format, which limits their utility to the community.

COMPARISON OF PLANT AND ANIMAL PAN-GENOMES

Although the main focus of this review is the pan-genomes of plant species, the concept of a pan-genome as the collection of all genetic sequences found in a species (or other taxon) is more generally applicable (65). The most widely studied pan-genome among animal species is, of course, that of humans (33, 42, 71, 77), although over the past few years pan-genome projects have been undertaken for fruit flies (12, 40), mice (44), pigs (79), and salmon (7). These studies, performed with various methodologies, are difficult to compare directly to each other and to those of plants. It does appear, however, that plants have quantitatively larger pan-genomes. The earliest comparison of multiple human genomes identified 5 Mb of additional sequences in Asian and African genomes relative to the human reference genome based predominantly on samples of European origin (42). These sequences are mostly short, with only one-sixth of the novel sequence contigs longer than 1 kb. At the level of genes, while the authors found that nearly 100 human transcripts defined by transcriptome data and absent from the human reference could be aligned to the two nonreference genomes, one-third of these novel genes are members of highly variable families associated with immunity, with the remainder being predominantly hypothetical genes of unknown function. Comparable amounts of novel sequences were found in other human populations (33, 77). Using a simple population genetics model of neutral insertion/deletion of new sequences and generously assuming a well-mixed large population size of 6.5 billion, Li et al. (42) suggested that the species-level pan-genome could amount to as much as 40 Mb. A more recent comprehensive study of 910 humans of African descent (71), however, found 296 Mb of nonreference sequences across their sample and found higher levels of novel sequences shared with high-quality Asian genomes, suggesting that the original estimates of Li et al. for novel sequences per genome could be underestimated by a factor of 10. At the gene level, approximately 300 named genes were affected by structural variation. This human pan-genome study, the most complete analysis currently available, suggests that the human pan-genome accounts for 10% more genes than the reference genome. Other animal pan-genomes include pig (estimated at 72.5 Mb additional sequence beyond the reference, based on 11 genomes) (79) and mouse (14–75 Mb of nonreference sequences in each of 16 mouse strains). These are within a factor of three of the human pan-genome size.

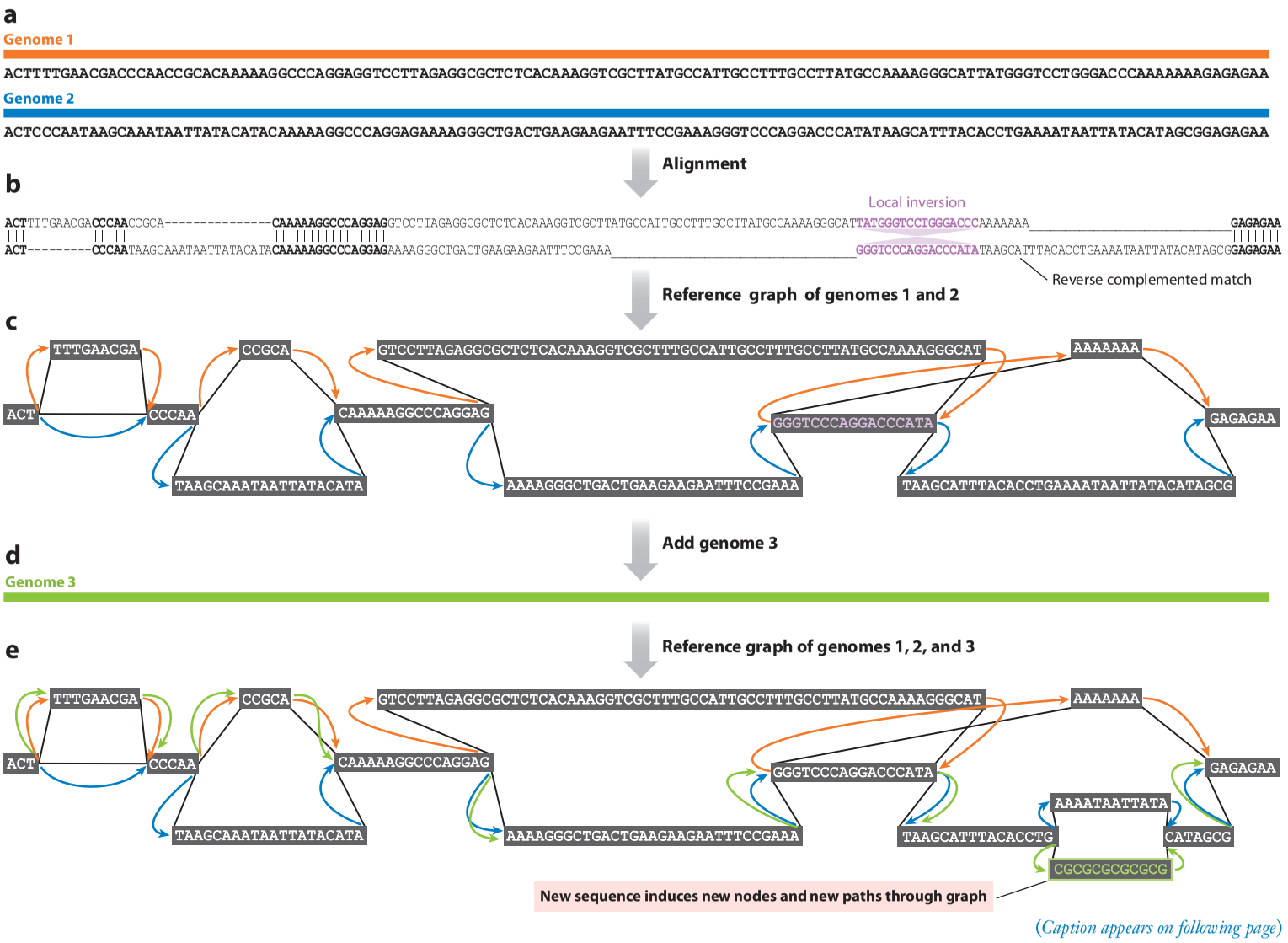
Why are plant pan-genomes so much larger than mammalian pan-genomes? The answer to this question can be found by considering the mutational and population-genetic processes that produce species-level diversity. As with single nucleotide variation, structural variation arises initially as a mutation; neutral variation is subject to genetic drift, while other variation is either swept to fixation (positive selection) or lost (negative selection). Key parameters for analyzing pan-genomes are therefore the mutation rate for structural variation and the effective population size, which governs genetic drift, as well as the relative fraction of SVs that are neutral. Relative to animals, plants have several numerical advantages that lead to expectations of larger pan-genomes. At the mutational level, the large number of recently expanded transposable repetitive elements provides ample substrate for nonallelic homologous recombination that can produce duplications and deletions. Active TEs can also mobilize neighboring sequences and produce structural change. Similarly, hybridization with other taxa can also add novel variation; from the perspective of natural genetic variation, the effect of this kind of gene migration is similar to mutation. The relative

rate of structural mutation between plants and mammals is therefore a key parameter to be determined. Since flowering plant species have a history of ancient duplications, residual effective redundancy may allow more of these SVs (especially deletions) to be neutral. A second crucial parameter is effective population size (36) since a larger population produces more mutations and can accommodate more standing variation due to drift. The larger effective population sizes of plants may explain the more than tenfold enhancement of standing single nucleotide variation compared with mammals. Standing structural variation is expected to scale linearly with effective population size. These mutational and population-genetic factors governing pan-genome size are appealing topics for future study.

SEQUENCE-BASED PAN-GENOMES AND GRAPH-BASED ANALYSIS

Plant pan-genome construction has undergone dramatic technical improvement from initial studies based on short reads and incomplete assemblies to the use of highly accurate and contiguous PacBio assemblies. This development opens the door to creating much more comprehensive sequence-based pan-genomes that include accurately placed noncoding and repetitive sequences. The plant pan-genomes constructed to date are primarily gene-based pan-genomes. Even if these studies use some sort of sequence-based pan-genome assembly as an intermediary step, their most accessible output includes matrices of PAVs and sometimes the corresponding genes from each individual line. Missing from these studies is an accessible representation or catalog of all sequence variants and their relationships to nearby (i.e., linked) sequences. This limits the utility of existing pan-genomes because they cannot easily be mined for nongenic sequences that control phenotypes (e.g., regulatory elements and insertions of TEs that influence gene expression) (6). In general, noncoding sequences are more variable than genes or TEs, and we know less about their biological significance. Some noncoding sequences (e.g., promoters and regulatory elements) play obvious roles, but the significance of the vast majority of noncoding sequences is unknown; however, conservation of some noncoding sequences over evolutionary time suggests an important role (80). Capturing the diversity of a species' nongenic sequences and repetitive elements requires the construction of an accurate, accessible, sequence-based pan-genome.

Sequence graphs have the potential to be a powerful tool to represent, display, and/or construct sequence-based pan-genomes because they can compactly represent intersequence relationships. Distinct sequences are represented as nodes of the graph, and their spatial relationships to other sequences (usually, immediate adjacency or proximity) are represented as edges (**Figure 3**). Because identical or highly similar sequences between genomes collapse into common nodes, even large collections of many assemblies can be encoded compactly. The basis of a sequence graph is typically a set of sequence alignments that reveal common and divergent elements in the data set (**Figure 3**). The more divergent the sequences, the more nodes and edges are induced, and the more complex the graph. Although attractive by their generic nature, alignment-based sequence graphs face several challenges stemming from difficulties in aligning divergent sequences. For example, if related, but highly divergent, sequences fail to align, they will not be recognized as potential allelic or duplicated variants and will not be correctly represented in the graph. Large repeats within genomes also may create ambiguities and errors in the alignment, leading to tangles in the graph that are difficult to interpret. Approaches being developed to work around these problems include preventing the collapse of multiple nodes in the graph and restricting alignment to uniquely anchored chains, which avoids complex graph structures from the start (41). Other approaches are more thorough and inclusive in the early stages of alignment, with a subsequent attempt to unroll the complex structures of the cyclic graph into isolated direct acyclic graphs that lack path collisions and other ambiguities (16, 62).



(Caption appears on following page)

Figure 3 (*Figure appears on preceding page*)

Iterative construction of a bidirected sequence graph of a pan-genome region. (*a*) Two assembled genomes are randomly chosen from the population to act as the initial reference and query sequences. (*b*) A sequence alignment is performed and blocks of collinear alignments, along with their breakpoints, are identified (short sequences are shown for simplicity). (*c*) A sequence graph is constructed from the alignments, with the adjacency of sequences represented as edges (*solid black lines*) connecting the nodes. By design, any unaligned segment will induce a new node in the graph and a new edge to any adjacent aligned segments. The bidirected nature of the graph allows for the expression of the relative orientation of the connected segments, as shown in the case of the node with the purple sequence. This directed connectivity allows sample-coherent paths to be drawn through the graph (*colored arrows*) and puts the graph topology into comparative context. The graph can now serve as a reference graph. (*d*) A third genome is aligned to the graph. The genome has strong but sporadic homology to genomes 1 and 2 (already present in the graph). (*e*) Adding this genome to the graph creates a new path (*green arrows*) for genome 3 that reflects its mosaic similarity to genomes 1 and 2. Additionally, the novel sequence induces a new topology in the form of a bubble with two divergent segments adjacent to common sequences on both sides. The reference graph now represents all three genomes. The iterative process can continue adding genomes until the graph becomes too fragmented to allow confident alignment.

When reliable prior knowledge exists about the variant composition in a pan-genome (typically obtained via read-to-reference mapping), there are computational tools that can transform a linear reference sequence and a set of variant calls into graphs (18). This approach bypasses the computationally expensive all-versus-all alignment step along with the uncertainties of subsequent graph construction, but the trade-off is increased reference bias and a potentially incomplete variant picture, especially among SVs, due to known limitations of current linear reference-based variant callers. It is still debated whether reference-guided graph construction or an all-versus-all alignment approach produces more practically relevant and computationally tractable results.

In a pan-genome sequence graph, every node is sample-coherent, meaning that the complete sequence associated with a node exists in at least one of the input assemblies. To reconstruct a particular genomic region, one simply needs to "walk" along adjacent nodes that originated from the sample(s) of interest (**Figures 2*c*** and **3**). One can expect that homologous regions in a pan-genome will share many nodes in common and, therefore, could be viewed as a collection of similar paths diverging and converging precisely around variant positions. Since the graph contains, ideally, all the assembled sequence in the pan-genome, rare variants could potentially be revealed this way. It must be noted that this model requires either that the input assemblies should be from haploid/inbred organisms or that the contigs are fully haplotype-phased.

Representing and categorizing variants in the form of a pan-genome graph that can be useful to a user present new challenges. First of all, even with a handful of genomes, the topology of a graph can get very complex. Smaller variants nested inside large structural rearrangements add new layers of complexity to the graph, making it difficult to classify a region of variation and define its boundaries (see the section titled Graph-Based Variant Calling). Establishing a hierarchy of variation topology becomes an important task—one that is tightly dependent on the graph construction model chosen. One type of alignment graph called a cactus graph has features that allow it to be transformed into a hierarchical tree of related but independent subgraphs (60). In this transformation, a set of alternative paths that diverge and then converge again becomes a kind of supernode from which the subgraph buds out. As these abstracted subgraphs can themselves be quite complex, each can be further simplified into child nodes and buds, giving the whole structure a cactus-like appearance. This hierarchical decomposition not only reveals the substructure of the pan-genome but also breaks down the problem of variant categorization into a more manageable set of problems (60). Other formalizations of nested structures in direct acyclic graphs have since been proposed with the similar goal of providing a framework for establishing generic models of variation inference in pan-genomes (18).

Establishing a stable genome coordinate system for a pan-genome graph is another significant challenge. Such a coordinate system would allow users to refer to a feature in a way that parallels the use of a reference genome coordinate system. Graphs constructed from all-to-all pairwise sequence alignments lose the linear coordinate system associated with linear references, which poses an obvious challenge to defining variant loci. An important development in this area is the advent of Minigraph, a data model where reference coordinates are established on the graph topology and remain stable as new samples are iteratively added to the graph (41). In this model, the linear coordinate of a node is encoded precisely for the single sample that had induced it during the iterative graph construction, and since every node is sample-coherent, this coordinate will not change as more samples are added. Another suggested proxy for a coordinate system is to use node identifiers directly to infer a pan-genomic locus, avoiding altogether explicit linear coordinates (49). In this scenario, one question that deserves further exploration is whether it is feasible, for the purposes of visualization and analysis, for a variant locus to be defined solely in relation to its graph neighborhood or whether it would still have to be mapped back to the individual linear genomes on an as-needed basis by a particular application. Furthermore, such a system would be unstable as the addition of a new genome to the reference graph could disrupt the node content.

Finally, pan-genomes with a large number of rare variants can result in very large and highly branched graphs, creating substantial computational challenges for both graph construction and the access/query operations a user may want to perform on the graph. Ultimately, there is a limit to how much variation an alignment-based graph can represent, determined in part by the computational efficiency of the graph model, but also by the limitations of the underlying alignment methods. After all, a sequence graph is only as accurate and sensitive as the alignments it was built from, and an ever-increasing fragmentation in the graph will ultimately limit its alignability. Specialized sequence alignment methods and efficient graph representation models are needed to address these challenges. Efforts in this area are spearheaded by the vgteam, the creators of a far-reaching graph-based framework for working with pan-genome variation (18).

GRAPH-BASED VARIANT CALLING

Reference graph data structures can represent both SVs and point mutations using the same semantics (27). However, since lengths of the polymorphisms involved vary dramatically, their discovery in a graph context is far from trivial. While there is no commonly accepted approach, a few methods have been proposed that we see as paving the way toward fully automated pan-genome variation profiling in the future. An intuitive way to discover pan-genomic variation embedded in a graph is by exploring its topology directly. Visually, a simple variant locus can appear as a bubble on a path through the graph. More complex variants can show up as nested bubbles and other patterns that can be more challenging to visualize and formally define (**Figure 4**). Computational algorithms, including the above-mentioned hierarchical graph decomposition, have been developed to automate this process of untangling the complex topology caused by nested variation (61). Once graph-embedded variant sites have been identified, one can envision mapping the reads back to the participating nodes to genotype each individual genome path through each bubble. For small variants, a direct read-to-graph pileup approach offered by VG tools can both call variants embedded in the base graph and detect novel variants, augmenting the graph in the process (55). Despite showing encouraging results in limited studies, much work lies ahead to make these methods more robust, sensitive, and applicable to different types of variation observed in pan-genomes.

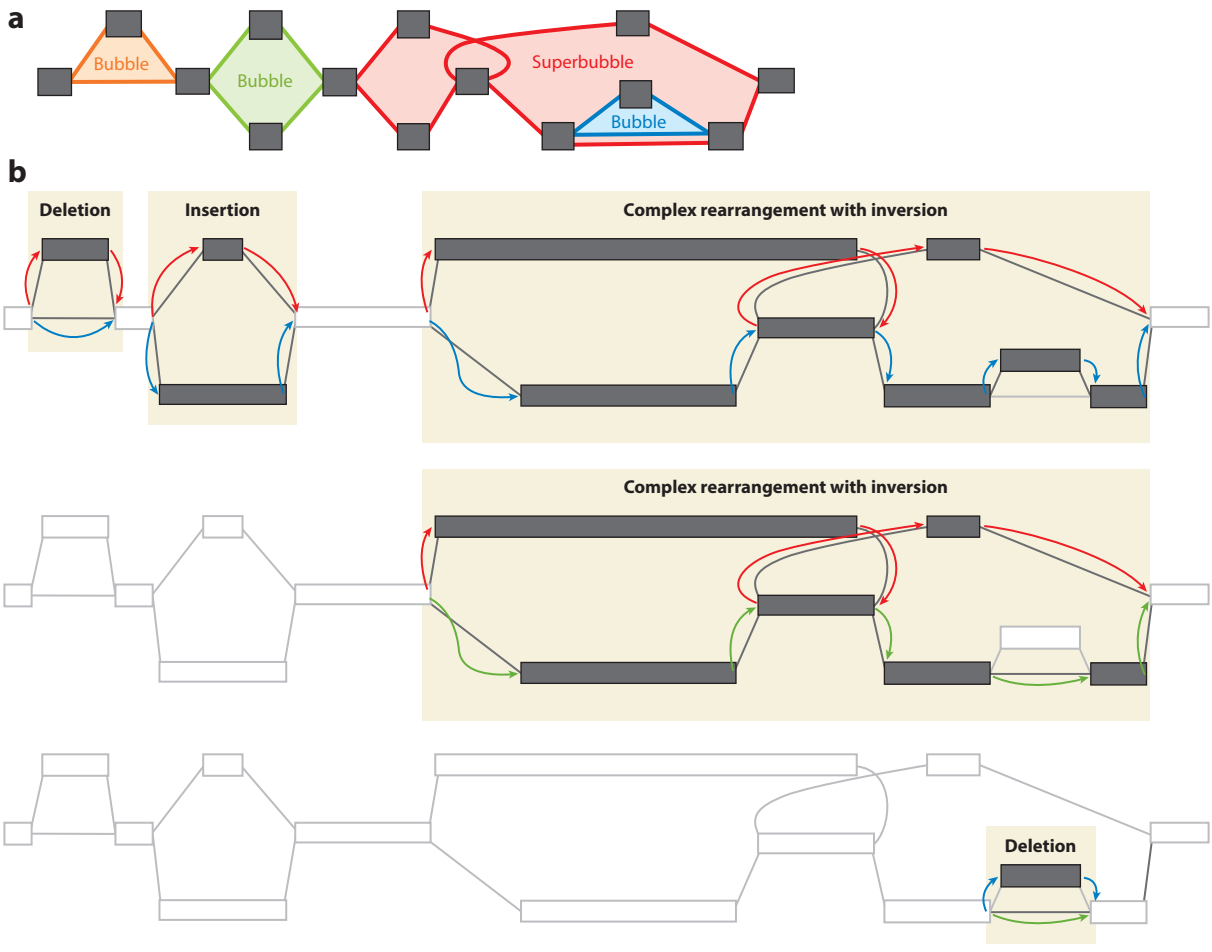


Figure 4

Exploring variation on a pan-genome sequence graph. (*a*) Bubble topologies are detected in the graph, defined as subgraphs with single source and sink vertices. A bubble can represent two or more alternative sequences at a variation site. In this example, a hypothetical algorithm detects simple bubbles. The blue bubble is nested inside a complex superbubble (*red*). The challenge for an algorithm is to detect and classify both the outer and the inner/nested bubble types regardless of node size. (*b*) With the bubbles identified, genome-specific paths (*colored arrows*) can be traced through each bubble and assigned coordinates relative to the chosen reference path (here, in each bubble, the upper sample path represents the chosen reference). Notice that, depending on the reference path chosen, a nested variant can be reported either as part of a larger variant (*top* and *middle diagrams*) or as an isolated case (*bottom*). The variants can then be output as vcf files with sequence and positional information included.

PAN-GENOME VISUALIZATION AND INTERROGATION

At the present time, most publicly hosted plant pan-genomes deploy traditional linear genome browsers [e.g., JBrowse (9), GBrowse (74), and IGV (66, 74)] for viewing base-level variation and gene annotations with respect to a designated reference genome, with PAV clusters displayed in simple tabular views, or use repurposed gene family viewers [e.g., Phytozome's gene family report (21) or Legume Federation's Genome Context Viewer (13)]. RPAN (75), hosting a rice pan-genome with over 3,000 accessions, provides a single JBrowse genome browser built on the linear reference pan-genome, with PAV tracks and a mapped read coverage track for each accession.

Genes in the browser are linked to a tabular view of PAV attributes (such as core versus non-core classification and gene presence frequency across major and minor subgroups of accessions). Searching by location, gene identifier, sequence similarity, or presence frequency within a user-specified subset of accessions is supported, but search results cannot be downloaded, only entire data sets. WheatPan (54) hosts a GBrowse genome browser built around an 18-cultivar hexaploid bread wheat pan-genome. PAV is indicated directly on the browser via pie charts for each gene, and details are available concerning which lines are present or absent. Genes are annotated via UniProt homology. The wheat pan-genome can be searched by region, gene identifier, and sequence similarity. BnPIR (73), the *B. napus* pan-genome information resource, includes a JBrowse instance built around the linear pan-genome constructed from eight high-quality assemblies, with PAV heatmap tracks, a GBrowse-based synteny view of all eight genomes simultaneously, and a separate variant browser. Detailed gene pages with functional annotation assignments are accessible via the Gene Search, and PAV cluster reports are available from the Gene Classification search page. In Phytozome's BrachyPan database (23), JBrowse genome browsers are available for the linear reference pan-genome, as well as for each of the 54 assembled and annotated *B. distachyon* accessions that were used to construct it. All genes from every accession are mapped into each accession's genome browser. Genes are color-coded and grouped into tracks based on whether they are part of core, soft-core, shell, or cloud PAV clusters, and each gene is linked to the BrachyPan PAV cluster for that gene. All gene and PAV cluster information can be downloaded from BrachyPan's BioMart (72) data warehouse, which also supports PAV fingerprint queries (e.g., "find all clusters with these accessions present and all others absent").

While single reference linear genome browsers provide a reasonable entry point to exploring certain types of pan-genome variation (gene PAVs, SNPs, and smaller SVs), they fall short on providing users with a visual overview of larger-scale structural variation across the pan-genome, such as translocations, large insertions, and nested variation (e.g., a subset of accessions all having an insertion relative to the reference and that insertion sequence itself varying across the subset). Ideally, one would like to move seamlessly from broad surveys of the full range of variation captured in the pan-genome graph, to selecting a subgraph based on variant content or being anchored by genes of interest, and finally to drilling down to a view of those variants mapped onto the (or a) reference genome in our traditional linear genome browser view. As noted in recent reviews (14, 16), multiscale plant pan-genome graph visualization remains a work in progress, impeded by the large size, variant density, and complexity of typical plant genomes. Many of the tools for visualizing assembly graphs [e.g., GfaViz (20), Bandage (86), and AGB (53)] struggle with regions larger than 0.1 Gbp (the model grass *B. distachyon* has a reference genome nearly three times as large) and are difficult to decorate with annotation features, making them hard to query when a user or application needs to extract subgraphs and regions of interest. Some of the tools providing base-scale detail have similar maximum region size limits [e.g., sequenceTubeMap (8), VG (27)], though ODGI, for example, can potentially handle genomes on the giga–base pair scale. One tool combining large-scale and local-scale visualization is MoMI-G (90), which uses a Circos (38) plot to display pan-genomic variation on the whole-genome level, synchronized with a table of (filterable) variants and a sequenceTubeMap view of stacked linearized paths representing the pan-genome's constituent haplotypes, with base-level resolution. Selecting variants in Circos or the tabular display loads the corresponding region into the sequenceTubeMap view. As MoMI-G has limitations on the number of haplotypes and region size over which it can display a linearized view, efforts are underway to develop the Pantograph Genome Browser (70), which will enable stacked linearized path views encompassing an entire plant chromosome (giga–base pair scale) and for thousands of haplotypes. Combining a multiscale pan-genome browser like Pantograph with the traditional linear genome browsers' search, display, and data mining tools provided by large

public plant-genome databases would provide a valuable, integrated platform for studying both the structure and biological function of plant pan-genomic diversity.

APPLICATION OF PAN-GENOMES

While plant pan-genomics is still an emerging field that has not yet exerted a large impact on plant research and breeding, a few examples highlight its application to population genetics, the evaluation of wild relatives, and polyploid genome evolution. By definition, pan-genomes inform us about population genomics and the prevalence of genes and polymorphisms across populations, and most pan-genome studies include substantial population analysis. However, these analyses primarily focus on SNPs relative to a reference genome and as such are expansions of previous techniques and studies (28, 30). With the advent of high-quality pan-genomes based on increasingly complete assemblies, researchers are gaining the power to comprehensively evaluate SVs at the population level. Recently, SVs have been used in genetic analysis, including genome-wide association studies (1) based on different types of SVs including PAVs (47, 73) and TE insertion polymorphisms (11). Of particular interest for clonal crops, SVs have been shown to be a major driving force for plant evolution and domestication (2, 96). However, most of these studies are based on SVs relative to a single reference genome with the discovery of insertion polymorphisms limited by read length. A full characterization of the different types of genetic variation can only be achieved with plant pan-genomes, where variations can be genotyped relative to a pan-genome instead of a single reference genome.

Hybridization and genomic introgression have been shown to be major sources of genetic variation in both animals (15, 51, 64) and plants (5, 30, 87). Genomic regions of introgression have been documented as hot spots for structural variations that contain disease resistance genes (3). With the possibility of fully characterizing different types of genetic variation, plant pan-genomes will enable genome-wide high-resolution admixture mapping across multiple scales and help pinpoint causal genetic mutations underlying specific traits.

Population stratification was traditionally studied with limited markers, and later with genome-wide SNPs, to understand evolutionary processes. With the advent of plant pan-genomics, PAVs have been used to characterize population differentiations (17, 23). A complete characterization of the SV spectrum (type, size, frequency) enabled by high-quality (graph-based) pan-genomes will offer a powerful new window into the evolutionary processes behind different types of SVs. Population-specific SVs in particular may be associated with traits specific to that population. Adaptive SVs may be identified using simulations to disentangle demographic and selection processes (52, 88).

Pan-genomes can also be created from multiple species or higher taxonomic groups to explore genome evolution. Not surprisingly, this approach has been applied to bacterial species. For example, core/pan-genome ratios were used to define new bacterial species (10). The rationales for multispecies pan-genome studies in plants include identifying genes and other sequences that could be transferred between species to improve crops, identifying historic introgressions between species, and understanding polyploid genome evolution. Pan-genomes from closely related, interfertile species are similar to a single species pan-genome in that, at least in theory, the genes have a possibility of exchange between individuals. Plant pan-genomes that fall into this category include rice and pepper (58, 94). When pan-genomes include wild relatives of domesticated crops, many genes of possible agricultural use can be identified (e.g., disease resistance genes and genes characteristic of abiotic stress tolerance). For example, Zhao et al. (94) compared wild and cultivated rice genomes and identified potentially useful genes that could be transferred to cultivated rice by constructing a pan-genome from 54 accessions of *Oryza sativa* and 13 *Oryza rufipogon* accessions.

They found 10,872 genes that were at least partially absent in the Nipponbare reference genome. As in other studies, the variable genes were enriched for genes involved in biotic and abiotic stress responses, exactly the type of genes that could be agronomically valuable. Another application of a multispecies pan-genome was used to study genome evolution after the polyploidization events that formed the allotetraploid grass *Brachypodium hybridum*. By constructing a pan-genome containing 52 genomes from one of the diploid progenitors and the corresponding D subgenomes from four polyploid genomes, the authors were able to distinguish genome evolution that happened before and after polyploidization (22). They determined that the D subgenome was evolving much more slowly than a simple comparison between single reference genomes would suggest. The substantial sequence divergence and changes in chromosome number between the diploid parents, *B. distachyon* and *Brachypodium stacei*, serve as substantial barriers to homoeologous recombination, which may contribute to the stability of the subgenomes. By contrast, a pan-genomic analysis of the recent allotetraploid *B. napus* revealed that frequent homoeologous exchanges between the very similar subgenomes accounted for many of the PAVs observed (31).

FUTURE DIRECTIONS

Technological improvements in long-read sequencing (e.g., PacBio circular consensus sequencing) have made it cost-effective to create multiple nearly complete genome assemblies for small- to moderate-sized plant genomes (up to 1 Gbp). These are the ideal raw material for creating highly accurate sequence-based pan-genomes that contain all the genes and noncoding sequences, as well as most repetitive elements in the proper order, orientation, and accurate long-range contiguity. A powerful scenario for using this technology would be to select 10–30 lines that capture the maximum amount of genetic diversity within a species and create reference-quality assemblies and annotations for this set. Significantly, several plant pan-genome studies have estimated the number of genomes required to capture most of the genomic diversity in a species and concluded that fewer than 30 genomes, selected to maximize diversity, are sufficient (19, 23, 28, 83, 94). A graph approach would then be used to create a high-confidence, sequence-based pan-genome from these reference-quality genomes. The resulting pan-genome graph would be used to generate displays that can be viewed at various zoom levels to allow users to visualize and search for all polymorphisms of interest, including large SVs, individual genes, and SNPs. The display would also allow the download and export of selected sequences and alignments for detailed analysis. This would complement a gene-based pan-genome that could easily be constructed from the annotated genomes. To capture the full diversity within a species, hundreds to thousands (such as in entire germplasm collections) of additional lines would be sequenced and assembled using short-read technology. The sequences and annotations from these lines would be aligned and added to the pan-genome graph and additional variants called and cataloged. The final result would be an extremely useful tool for studying natural diversity, identifying markers for breeding, selecting lines with desirable combinations of alleles, assigning function to noncoding sequences, and studying gene function.

DISCLOSURE STATEMENT

D.S.R. is a member of the Scientific Advisory Board of Dovetail Genomics. The other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the US DOE under contract DE-AC02-05CH11231. We thank Dr. Peter Morrell and his lab members at the University of Minnesota for valuable comments.

LITERATURE CITED

1. Akakpo R, Carpentier MC, Hsing YI, Panaud O. 2020. The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.* 226:44–49
2. Allaby R. 2019. Clonal crops show structural variation role in domestication. *Nat. Plants* 5:915–16
3. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, et al. 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182:145–61.e23
4. Bai Z, Chen J, Liao Y, Wang M, Liu R, et al. 2016. The impact and origin of copy number variations in the *Oryza* species. *BMC Genom.* 17:261
5. Baurens FC, Martin G, Hervouet C, Salmon F, Yohomé D, et al. 2019. Recombination and large structural variations shape interspecific edible bananas genomes. *Mol. Biol. Evol.* 36:97–111
6. Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 65:505–30
7. Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nat. Commun.* 11:5176
8. Beyer W, Novak AM, Hickey G, Chan J, Tan V, et al. 2019. Sequence tube maps: making graph genomes intuitive to commuters. *Bioinformatics* 35:5318–20
9. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* 17:66
10. Caputo A, Fournier PE, Raoult D. 2019. Genome and pan-genome analysis to classify emerging bacteria. *Biol. Direct* 14:5
11. Carpentier MC, Manfroi E, Wei FJ, Wu HP, Lasserre E, et al. 2019. Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat. Commun.* 10:24
12. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* 10:4872
13. Cleary A, Farmer A. 2018. Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny. *Bioinformatics* 34:1562–64
14. Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D. 2020. Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* 54:18–25
15. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, et al. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science* 366:594–99
16. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, et al. 2020. Pangenome graphs. *Annu. Rev. Genom. Hum. Genet.* 21:139–62
17. Gao L, Gonda I, Sun H, Ma Q, Bao K, et al. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51:1044–51
18. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36:875–81
19. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, et al. 2016. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7:13390
20. Gonnella G, Niehus N, Kurtz S. 2019. GfaViz: flexible and interactive visualization of GFA sequence graphs. *Bioinformatics* 35:2853–55
21. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40:D1178–86
22. Gordon SP, Contreras-Moreira B, Levy JJ, Djamei A, Czedik-Eysenberg A, et al. 2020. Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nat. Commun.* 11:3670

23. Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, et al. 2017. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 8:2184
24. Gordon SP, Priest H, Des Marais DL, Schackwitz W, Figueroa M, et al. 2014. Genome diversity in *Brachypodium distachyon*: deep sequencing of highly diverse inbred lines. *Plant J.* 79:361–74
25. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, et al. 2009. A first-generation haplotype map of maize. *Science* 326:1115–17
26. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10:551–64
27. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, et al. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 21:35
28. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, et al. 2014. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 26:121–35
29. Huang K, Rieseberg LH. 2020. Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* 11:296
30. Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, et al. 2019. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* 5:54–62
31. Hurgobin B, Golicz AA, Bayer PE, Chan CKK, Tirnaz S, et al. 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnol. J.* 16:1265–74
32. Jangam D, Feschotte C, Betrán E. 2017. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* 33:817–31
33. Kehr B, Helgadottir A, Melsted P, Jonsson H, Helgason H, et al. 2017. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* 49:588–93
34. Knoll A, Fauser F, Puchta H. 2014. DNA recombination in somatic plant cells: mechanisms and evolutionary consequences. *Chromosome Res.* 22:191–201
35. Kobayashi S, Goto-Yamamoto N, Hirochika H. 2004. Retrotransposon-induced mutations in grape skin color. *Science* 304:982
36. Koonin EV. 2009. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* 41:298–306
37. Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36:6688–719
38. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res.* 19:1639–45
39. Ku C, Martin WF. 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. *BMC Biol.* 14:89
40. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. 2016. A thousand fly genomes: an expanded *Drosophila* genome nexus. *Mol. Biol. Evol.* 33:3308–13
41. Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs. arXiv:2003.06079 [q-bio.GN]
42. Li R, Li Y, Zheng H, Luo R, Zhu H, et al. 2010. Building the sequence map of the human pan-genome. *Nat. Biotechnol.* 28:57–62
43. Li YH, Zhou G, Ma J, Jiang W, Jin LG, et al. 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32:1045–52
44. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, et al. 2018. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nat. Genet.* 50:1574–83
45. Lin K, Zhang N, Severing EI, Nijveen H, Cheng F, et al. 2014. Beyond genomic variation—comparison and functional annotation of three *Brassica rapa* genomes: a turnip, a rapid cycling and a Chinese cabbage. *BMC Genom.* 15:250
46. Lister C, Jackson D, Martin C. 1993. Transposon-induced inversion in Antirrhinum modifies *nivea* gene expression to give a novel flower color pattern under the control of *cycloidea*^{radialis}. *Plant Cell* 5:1541–53
47. Liu Y, Du H, Li P, Shen Y, Peng H, et al. 2020. Pan-genome of wild and cultivated soybeans. *Cell* 182:162–76.e13
48. Marroni F, Pinosio S, Morgante M. 2014. Structural variation and genome complexity: Is dispensable really dispensable? *Curr. Opin. Plant Biol.* 18:31–36

49. Marschall T, Marz M, Abeel T, Dijkstra L, Dutilh BE, et al. (Comput. Pan-Genom. Consort.). 2018. Computational pan-genomics: status, promises and challenges. *Brief. Bioinformat.* 19:118–35
50. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, et al. 2009. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *PNAS* 106:12273–78
51. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* 8:14363
52. Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* 35:561–72
53. Mikheenko A, Kolmogorov M, Hancock J. 2019. Assembly Graph Browser: interactive visualization of assembly graphs. *Bioinformatics* 35:3476–78
54. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, et al. 2017. The pangenome of hexaploid bread wheat. *Plant J.* 90:1007–13
55. Novak A, Hickey G, Garrison E, Blum S, Connelly A, et al. 2017. Genome graphs. bioRxiv 101378. **https://doi.org/10.1101/101378**
56. Osada N, Innan H. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLOS Genetics* 4:e1000305
57. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* 18:2024–33
58. Ou L, Li D, Lv J, Chen W, Zhang Z, et al. 2018. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence–absence variation analyses. *New Phytol.* 220:360–63
59. Pardue ML, DeBaryshe PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu. Rev. Genet.* 37:485–511
60. Paten B, Diekhans M, Earl D, St. John J, Ma J, et al. 2011. Cactus graphs for genome comparisons. *J. Comput. Biol.* 18:469–81
61. Paten B, Eizenga JM, Rosen YM, Novak AM, Garrison E, Hickey G. 2018. Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.* 25:649–63
62. Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. *Genome Res.* 27:665–76
63. Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome size diversity and its impact on the evolution of land plants. *Genes* 9:88
64. Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, et al. 2020. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science* 368:731–36
65. Richard GF. 2020. Eukaryotic pangenomes. In *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, ed. H Tettelin, D Meidini, pp. 253–91. Cham, Switz.: Springer
66. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. 2011. Integrative genomics viewer. *Nat. Biotechnol.* 29:24–26
67. Ruprecht C, Lohaus R, Vanneste K, Mutwil M, Nikoloski Z, et al. 2017. Revisiting ancestral polyploidy in plants. *Sci. Adv.* 3:e1603195
68. Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, et al. 2014. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. *Genome Biol.* 15:506
69. Schubert I, Vu GTH. 2016. Genome stability and evolution: attempting a holistic view. *Trends Plant Sci.* 21:749–57
70. Seaman J. 2020. PantoGraph. *The Pantograph Project.* **https://graph-genome.github.io/pantograph.html**
71. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51:30–35
72. Smedley D, Haider S, Durinck S, Pandini L, Provero P, et al. 2015. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* 43:W589–98
73. Song JM, Guan Z, Hu J, Guo C, Yang Z, et al. 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* 6:34–45
74. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Res.* 12:1599–610

75. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, et al. 2017. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res*. 45:597–605
76. Tao Y, Zhao X, Mace E, Henry R, Jordan D. 2019. Exploring and exploiting pan-genomics for crop improvement. *Mol. Plant* 12:156–69
77. Telenti A, Pierce LCT, Biggs WH, Di Iulio J, Wong EHM, et al. 2016. Deep sequencing of 10,000 human genomes. *PNAS* 113:11901–6
78. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *PNAS* 102:13950–55
79. Tian X, Li R, Fu W, Li Y, Wang X, et al. 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci. China Life Sci.* 63:750–63
80. Turco G, Schnable JC, Pedersen B, Freeling M. 2013. Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. *Front. Plant Sci.* 4:170
81. Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* 23:148–54
82. Vu GTH, Schmutzer T, Bull F, Cao HX, Fuchs J, et al. 2015. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome* 8:eplantgenome2015.04.0021
83. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557:43–49
84. Wendel JF. 2015. The wondrous cycles of polyploidy in plants. *Am. J. Bot.* 102:1753–56
85. Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr. Opin. Genet. Dev.* 49:1–7
86. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31:3350–52
87. Wu GA, Terol J, Ibanez V, López-García A, Pérez-Román E, et al. 2018. Genomics of the origin and evolution of *Citrus*. *Nature* 554:311–16
88. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. 2018. Differential effect of selection against LINE retrotransposons among vertebrates inferred from whole-genome data and demographic modeling. *Genome Biol. Evol.* 10:1265–81
89. Yao W, Li G, Zhao H, Wang G, Lian X, Xie W. 2015. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol*. 16:187
90. Yokoyama TT, Sakamoto Y, Seki M, Suzuki Y, Kasahara M. 2019. MoMI-G: modular multi-scale integrated genome graph browser. *BMC Bioinformat*. 20:548
91. Yu C, Zhang J, Peterson T. 2011. Genome rearrangements in maize induced by alternative transposition of reversed *Ac/Ds* termini. *Genetics* 188:59–67
92. Yu J, Golicz AA, Lu K, Dossa K, Zhang Y, et al. 2019. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* 17:881–92
93. Zhang J, Peterson T. 2004. Transposition of reversed *Ac* element ends generates chromosome rearrangements in maize. *Genetics* 167:1929–37
94. Zhao Q, Feng Q, Lu H, Li Y, Wang A, et al. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50:278–84
95. Zhou P, Silverstein KAT, Ramaraj T, Guhlin J, Denny R, et al. 2017. Exploring structural variation and gene family architecture with de novo assemblies of 15 *Medicago* genomes. *BMC Genom*. 18:261
96. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, et al. 2019. The population genetics of structural variants in grapevine domestication. *Nat. Plants* 5:965–79
97. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33:408–14
98. Ziolkowski PA, Blanc G, Sadowski J. 2003. Structural divergence of chromosomal segments that arose from successive duplication events in the *Arabidopsis* genome. *Nucleic Acids Res.* 31:1339–50