Trends in Plant Science

CellPress OPEN ACCESS

Review

New whole-genome alignment tools are needed for tapping into plant diversity

Baoxing Song [1,2,*] Edward S. Buckler,[3,4,5] and Michelle C. Stitzer[3,6,*]

Genome alignment is one of the most foundational methods for genome sequence studies. With rapid advances in sequencing and assembly technologies, these newly assembled genomes present challenges for alignment tools to meet the increased complexity and scale. Plant genome alignment is technologically challenging because of frequent whole-genome duplications (WGDs) as well as chromosome rearrangements and fractionation, high nucleotide diversity, widespread structural variation, and high transposable element (TE) activity causing large proportions of repeat elements. We summarize classical pairwise and multiple genome alignment (MGA) methods, and highlight techniques that are widely used or are being developed by the plant research community. We also outline the remaining challenges for precise genome alignment and the interpretation of alignment results in plants.

Highlights

High-quality telomere-to-telomere plant genome assemblies are now readily conducted at the population scale for model and non-model species. These assemblies are the result of rapid improvements in genome sequencing over recent decades.

The challenges of structural and sequence diversity make well-developed alignment methods in animal genomics unsuitable for many plant genomes. This lag in comparative technologies for plant genomes has delayed investigations of genome evolution and genetics, and has focused attention only on protein-coding genes.

Pairwise and multiple genome alignment approaches optimized for plant genomes are necessary to make significant advances in understanding plant structural variation, genomic rearrangements, and sequence evolution.

Plant genome alignment is fundamental and essential

Whole-genome alignment (WGA) is the process of identifying homologous regions within a collection of assembled genomes and then performing base-pair resolution sequence alignments. WGA generates a positional relationship between sequences within genomes, enabling investigation of genomic function and how evolutionarily related sequences have diverged from their common ancestor (Box 1). WGA – both intraspecific and interspecific – is rapidly being incorporated into studies as the cost and technology barriers to genome assembly are lowered. Intraspecific WGA can be used in population and quantitative genetic studies to more accurately identify causal variants, and future advances in **graph genome** (see Glossary) methods can further facilitate these surveys. WGA for interspecific comparisons is established for a wide range of research aims, including genome evolution, understanding phylogenetic relationships, and functional sequence identification. These questions can be asked in an ever-growing range of model and nonmodel systems as genome assembly becomes routine.

The majority of WGA algorithms and tools were initially developed by the human genome research community to align the genomes of human, mouse, rat, and chimpanzee [1], for example BLASTZ [2], MUMmer [3], and LAST [4]. Benchmarking of WGA performance has largely been conducted using simulated data that mimic the landscape of sequence divergence that is common in mammal species [5]. However, the structure and chromosome evolution of plant and mammal genomes differ from each other [6] (Figure 1). Plant genomes vary in size more than any other taxa [6], and the largest known plant genome is 2400-fold larger than the smallest [7,8]. Polyploidy is a major contributor to genome size differences in plants, where at least 35% of species are polyploid [9]. Another major contributor to genome size differences is **transposable element (TE)** content, and genome size is a linear function of TE content [10]. The activity of TEs often results in the duplication of identical sequences and represents an important substrate for the emergence and expansion of repetitive sequences in the host genome. The activity (TE **presence or absence variation, PAV**) and decay of TEs also cause long **indel** variants

[1]National Key Laboratory of Wheat Improvement, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agriculture Sciences in Weifang, Weifang, Shandong 261325, China
[2]Key Laboratory of Maize Biology and Genetic Breeding in Arid Area of Northwest Region of the Ministry of Agriculture, College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712100, China
[3]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA
[4]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853, USA
[5]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY 14853, USA

[6]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

*Correspondence:
baoxing.song@pku-iaas.edu.cn (B. Song) and
mcs368@cornell.edu (M.C. Stitzer).

Box 1. The impact of sequencing technology development on estimated sequence diversity

The ability to compare the genomic sequences of different individuals and species has grown at breakneck speed over the past few decades and can now provide extremely valuable insights into the association between genomic diversity and phenotypic variance as well as into the evolution of population genomic sequences. From early biochemical studies using isozymes, we are now able to directly investigate genome sequence diversity among different lines/accessions of the same species or between closely related species. The whole-genome sequencing (WGS) approach is performed using massively parallel sequencing technologies (e.g., Illumina), and sequencing reads are mapped to the reference genome via **short-read alignment** methods [108]. Mutations of each sequenced individual are identified, usually single-nucleotide polymorphisms (SNPs) and indels.

A major limitation of the short-read method is that it has limited ability to genotype long indels and highly polymorphic regions. An artificial distinction exists between short indels (<50 bp), termed indels, and indels >50 bp which have been classified as structural variation (SV) [109–114]. There is no biological mechanistic reason for a specific threshold (e.g., 50 bp), except for the technical ability of the insert size or read length to cover both edges of a variant. Read-mapping approaches show a limited ability to investigate nested variants (Figure 3). In addition, to call a variant, short reads must map to the reference genome. For highly discordant regions or species distant from the reference genome, this reduces even SNP calls.

The development of long-read sequencing technology (e.g., PacBio and Oxford Nanopore) has made read length less of a limitation. Reads from these technologies can be mapped back to a reference genome and indels called via read alignments [115,116]. These generate even more power when aligning *de novo* genome assemblies generated from these reads. Several telomere-to-telomere genome assemblies without gaps have been established for humans [117,118] as well as for a wide range of plant or crop species such as rice [119,120], maize [121], watermelon [122], arabidopsis [123], tomato, and potato. With large-scale and high-quality genomes, researchers can comprehensively investigate any type of genomic variants using both inter- and intra-species genome alignments. For example, the genome assembly and comparison approach accurately identified the 140 bp MITE insertion at the *VGT1* locus on Mo17 compared to B73 [30], whereas this variant is absent from Panzea project short-read variant calling. The 16 bp indel at *FRIGIDA* (*FRI*) was not identified by the 1001 Genomes Project but could be identified via genome alignment (Figure I).

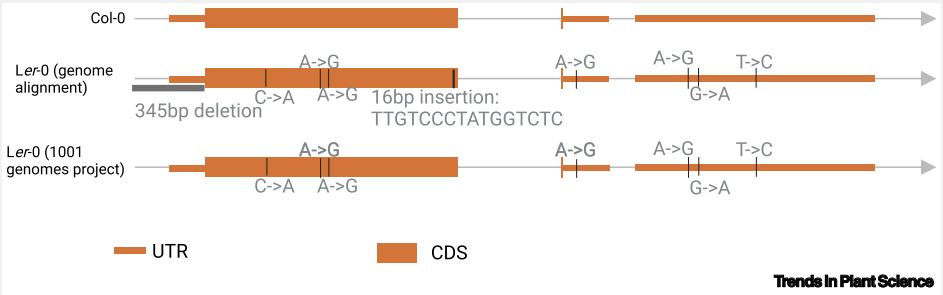

Figure I. Genome alignment can more comprehensively identify variants than short-read mapping. The 16 bp insertion at the *FRI* locus (AT4G00650) on L*er*-0 compared to Col-0 could not be identified via short reads by the 1001 Genomes Project. The genome assembly and comparison approach accurately identified this insertion. The region upstream of *FRI* in L*er*-0 contains hyper-diverse sequences, and the boundaries of this deletion have been reported differently because of differences in alignment parameters [124].

between genomes that can alter gene function and generate phenotypic impacts [11,12]. By contrast, most mammalian genomes have very few active TEs, resulting in slower accumulation of these large-scale indel variants mediated by TE movement. Finally, plants have higher sequence diversity even the level of individual nucleotides [13], likely reflecting larger effective population sizes. Higher sequence diversity means more uncertainty in alignment. These chromosome-scale sequence features raise profound technological challenges for plant WGA.

Recent improvements in genome sequencing and assembly technologies have allowed the assembly of large plant genomes with high continuity, a low error rate, and few gaps [14]. Accurate alignment of plant genomes can help to answer fundamental questions about plant evolution and environmental adaptation. Alignments can identify chromosome rearrangements and WGD

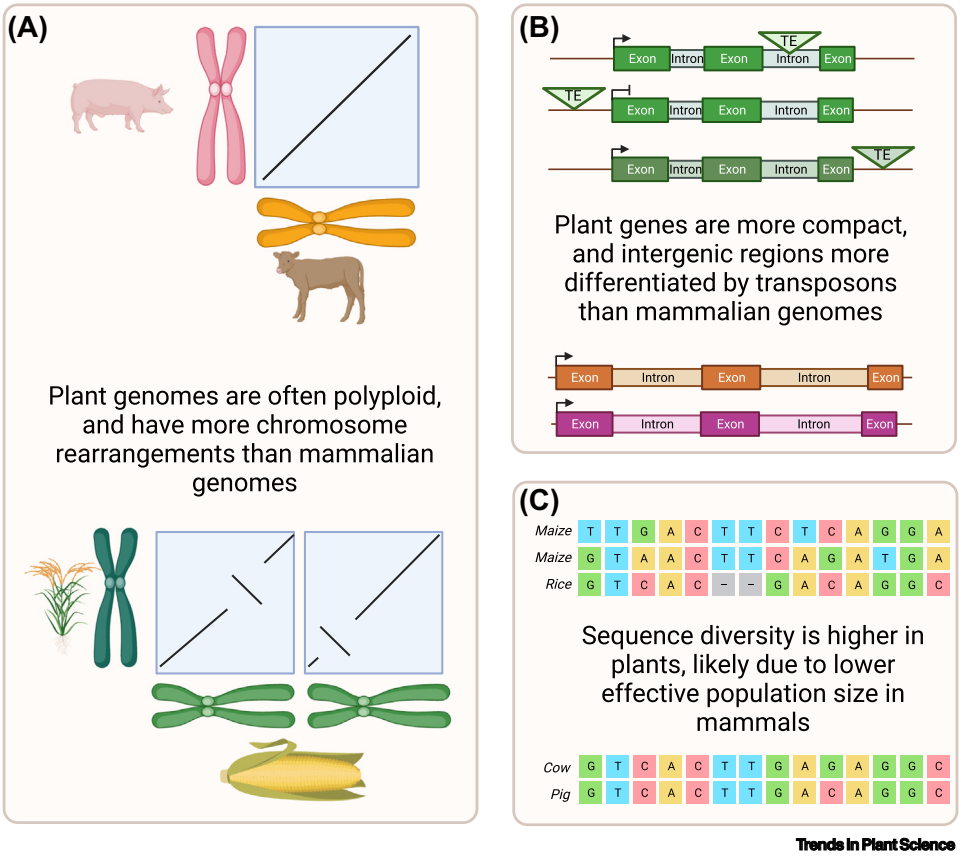

Figure 1. Alignment across different genomic scales. (A) Collinear chromosomal regions are maintained between two species until a rearrangement occurs. In mammals (top) this often takes the form of inversions and chromosome fissions/fusions. In plants (bottom) these processes occur, but additional whole-genome duplications mean that taxa maintain duplicated blocks of chromosomes that can be found on more than one chromosome. This difference means that an aligner that only allows one-to-one alignment will miss biologically relevant sequences. (B) Genes and transposable element (TE) positions are more constant in mammalian genomes than in plant genomes. (C) Nucleotide divergence is higher in plants.

events which can result in barriers to crossing, leading to speciation. Because polyploidization is often followed by rediploidization, alignment can identify how duplicated genes have diverged genetically and functionally, and whether one or both duplicate copies have been lost from a genome. At the base-pair scale, genome alignment can examine single-nucleotide polymorphisms (SNPs) and short or long indel variation which can be used to identify functional variants that cause adaptive phenotypes of interest. Alignment can identify shared distal regulatory sequences even when they have been pushed far away from their target gene by TE insertions. Thus, the study of gene expression evolution benefits from accurate and specific identification of long indels and SNPs in regulatory sequences. Because evolution occurs within the context of a genome, genome alignment is an essential first step to identifying the underlying molecular differences between any two sequence assemblies.

In this review we first introduce nucleotide alignment algorithms, and then highlight why plant genome alignment requires optimization strategies that differ from those developed for mammalian genomes. The goal is not to provide an exhaustive list of all genome alignment tools but instead to cover the biases and limitations of current strategies. We provide an outlook on what is necessary to further optimize plant genome alignment technology.

Glossary

Fractionation: following a whole-genome duplication (WGD) event, fractionation is a process that returns the genome to a diploid state. In this process most duplicated genomic features are lost from one or the other homolog.
Global alignment: end-to-end alignment of two sequences. The classical method is the Needleman–Wunsch algorithm proposed by Saul B. Needleman and Christian D. Wunsch in 1970.
Graph genome: a data structure to represent or store variants between genomes. It displays species variation information by representing sequence and structural variation information in the form of nodes and paths.
Indel: a DNA sequence that differs by insertion or deletion between two sequences.
***k*-mer:** a subsequence of length *k*, where *k* is a parameter with an integral value. In practice, all subsequences of length *k* of a sequence are usually generated.
Local alignment: assumes that two sequences are not similar over the entire length, and finds the regions with the highest level of similarity between the two sequences. The classical method is the Smith–Waterman algorithm proposed by Temple F. Smith and Michael S. Waterman in 1981.
Presence or absence variation (PAV): describes sequences that are present in one genome but are entirely missing in another genome.
Progressive alignment: a method that decomposes the multiple sequence alignment (MSA) problem into a group of alignments for two sequences or two groups of aligned sequences. This approach works by successively constructing pairwise alignments. At each iteration, two multiple alignments are aligned by a procedure that treats it as a sequence, resulting in a combined multiple alignment that can be passed to the next iteration. This procedure terminates with the complete MSA.
Short-read alignment: also known a short-read mapping, the process of aligning reads to a reference genome in the presence of errors and genetic variations.
Subgenome: following a WGD, the set of chromosomes from each parent is called a subgenome. Over evolutionary time, many duplicate genes (but not all) are lost from one homologous region or

Base-pair resolution pairwise genome alignment algorithms

The alignment of two homologous sequences often assumes that two aligned bases are derived from a shared position in a common ancestor. We often have few ways to test the biological validity of this assumption, but computational developments in alignment have begun to incorporate models of how sequences evolve. Sequence alignment is complicated by indel variation because there is no longer a one-to-one relationship between each base pair. When an indel has occurred, aligners must choose when to introduce and extend a gap. To produce the optimized alignment quickly, dynamic programming algorithms are usually used. The dynamic programming sequence alignment algorithms were initially developed by Needleman and Wunsch for **global alignment** [15] and by Smith and Waterman for **local alignment** [16]. The central processing unit (CPU) time and memory cost of these classical algorithms is proportional to the product of the lengths of the two sequences being aligned, which limits their application because costs grow exponentially. For example, a 100 kb comparison requires ~40 GB of memory, more than the capacity of many personal computers (Table S1 in the supplemental information online). A banded sequence alignment approach [17] can reduce both time and memory cost – over 100-fold for a 100 kb sequence – by not computing all cells in the scoring matrix, leading to a limitation that it may not generate the optimal alignment. To reduce memory usage, Hirschberg's algorithm reduces linear space, and this idea has been extended to perform both local and global alignment with affine gap penalty (Table S2) [18,19]. Advances in modern computational hardware designs now allow dynamic programming algorithms in many of the commonly used sequence alignment tools to be implemented using single instruction multiple data (SIMD) technology, and this can accelerate computation by >tenfold by processing multiple data concurrently with a single instruction [20–22]. The computational cost of the recently developed wavefront global alignment (WFA) approach [23] is not directly related to the input sequence length and is instead related to the dissimilarity between the two input sequences (Table S1), resulting in an ability to align over longer distances.

A simple WGA approach would be to perform global alignment for each pair of homologous chromosomes from end to end. However, this is not possible in practice owing to the high computational cost of dynamic programming approaches. Even if computationally possible, dynamic programming algorithms produce alignments that have a fixed order and orientation which cannot capture genomic rearrangements such as inversions or translocations. Moreover, when a WGD has occurred, the correspondence between chromosomes is not one-to-one. Alternative methods are therefore necessary.

The seed-and-extend approach performs large-scale sequence alignments efficiently. This strategy produces an approximately optimal alignment using shared ***k*-mers** as seeds to trigger alignment and then extends the alignment from these shared sequences (Figure 2A). This idea has been widely adopted for various sequence alignment problems, including WGA. Modern WGA tools generally process matched seeds further by non-gap extension, chaining, anchoring, and so forth to speed up or reduce undesired alignments (R.S. Harris, PhD thesis, Pennsylvania State University, 2007). For example, the MUMmer software groups exact matches into clusters and performs alignment for regions between consecutive exact matches as well as flanking regions using the Smith–Waterman algorithm [3]. However, these approaches can fail to generate an alignment when genomic regions have sufficiently high sequence diversity that a seed does not exist, and can also generate high rates of false positive alignments when genomes have abundant repeat elements. However, sensitivity and specificity can be traded off against each other [24–27]. To increase sensitivity, LASTZ and LAST use flexible seeds that allow mismatches to the target sequence [4] (Table 1). To increase specificity, repeat elements are generally annotated and soft-masked [28]. Because these masked sequences are not used as seeds, they do not initiate an alignment.

its partner region asymmetrically. These subgenomes may differ in both gene density and the level of gene expression. One subgenome is prone to retain more genes and is referred to as the 'dominant subgenome', whereas the other loses more genes and is referred to as the 'recessive subgenome'.

Transposable element (TE): a DNA sequence that can change its position within a genome via a 'copy-and-paste' mechanism (retrotransposons) or a 'cut-and-paste' mechanism (transposons). They typically range in length from 50 to 10 000 bp but are sometimes far larger. The activity of TE generates long indel variants between genomes, and retrotransposons introduce repeat sequences within genomes.

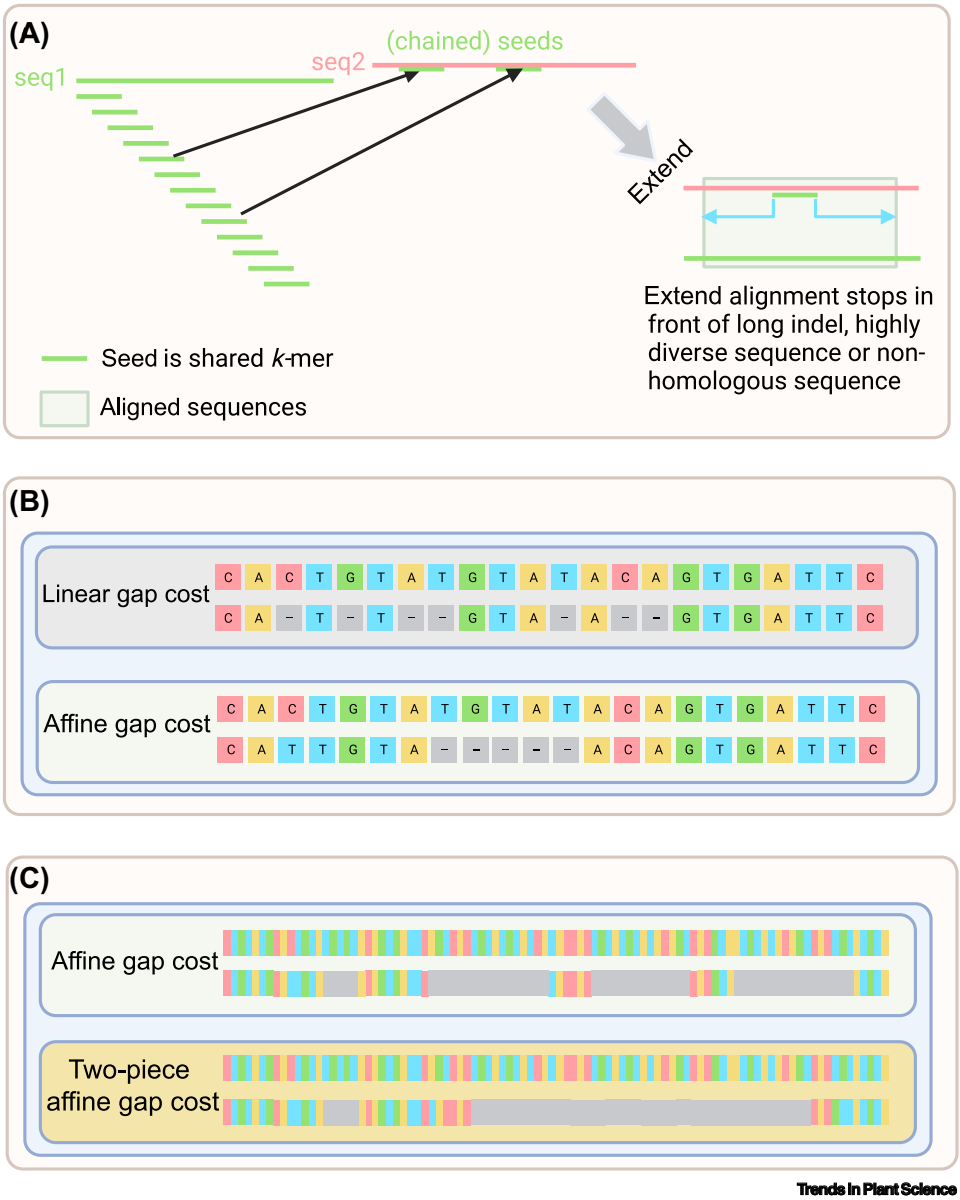

Figure 2. Comparison of sequence alignment approaches and gap cost penalties. (A) The basic principle of the seed-and-extend/seed-chain-extend approach. This approach can identify locally similar blocks in two sequences. (B) The global alignment approach produces alignment for every base pair of the input sequence, even for highly diverse regions and long indels. The affine gap cost penalty generates fewer gap variants than the linear gap cost. (C) The two-piece affine gap cost penalty allows fewer but longer gap variants than the (single-piece) affine gap cost approach.

Another WGA strategy utilizes the large-scale collinearity of phylogenetically closely related genomes. This approach comprises two steps. First, a collinear map between the genomes is constructed using a variety of approaches that use *k*-mer matches, conserved segments identified via local alignment approaches, or annotated genes as anchors. Collinear block identifications are usually modeled as graph optimization problems which can be solved using sparse dynamic programming algorithms [29–32]. Second, a base-pair resolution alignment is obtained for each

Table 1. Comparison of different classical genome alignment strategies used by the plant genome community

	Classical software	Seed strategy	Indel scoring	Does it align long indels (at least 2 kb)?	WGD aware	Description
Pairwise						
Seed-chain-extend alignment	MUMmer	Maximal unique matches (MUM)	Affine gap	No	No	MUMmer is fast, with default settings, its sensitivity is lower than LASTZ or LAST.
	LASTZ, LAST	Spaced seed	Affine gap	No	No	LASTZ and LAST provide flexible options for seeds and scoring parameters. LAST is an update for LASTZ that is faster but is not an exact replacement. For example, LASTZ provides options to infer scores from sequences.
Seed-chain global alignment	Minimap2	Minimizer	Two-piece affine gap	Yes	No	Minimap2 identifies primary chaining using minimizers as anchors, and performs global alignment between adjacent anchors in a chain.
	AnchorWave	Coding gene	Two-piece affine gap	Yes	Yes	AnchorWave uses coding genes as anchors to identify collinear blocks, and performs global sequence alignment for each anchor and inter-anchor regions. AnchorWave conducts WGAs between closely related individual chromosomes and handles each translocation as a long deletion and an long insertion. When aligning genomes of the same species, the genoAli function of AnchorWave expects that the homologous chromosomes in the two genomes would have exactly same chromosome name.
Multiple						
	Pairwise alignment approach	Description				
Cactus	LASTZ	A graph-based progressive multiple genome aligner that reconstructs ancestral genomes by combining sub-alignments. It can perform MGAs for 1000 genomes. However, it has not been well benchmarked for plant genomes and natural variation genomes from the same species.				
ROAST	BLASTZ/LAST/LASTZ	After aligning multiple query genomes to the reference genome, ROAST progressively combines them into multiple genome alignments and generates the output in multiple alignment format (MAF). The MSA_pipeline workflow has automated this process.				
Others						
	Description					
Chain-and-net	Combines fragmented alignments into larger alignment regions. Combined with LAST, for example, it has been used to generate genome-wide reciprocal best hits.					
Quota-alignment	Uses coding genes as anchors, it conducts WGD-aware genome-wide syntenic block identification and also reports syntenic homologous genes, but does not produce base-pair resolution sequence alignments. Several subsequent programs perform similar analysis but use different algorithms. This approach has also been used to compare multiple genomes, for example in MCScan and GENESPACE.					
SyRI or Assemblytics	They take the pairwise genome alignment result from minimap2 or MUMmer as the input to identify consecutive alignments along a sample contig, and call indels up to a maximum of 10 kb in size indirectly. SyRI expects that homologous chromosomes in two genomes have the same chromosome name.					

collinear region by a global alignment algorithm using scoring parameters optimized for long indels (Figure 2B). Such approaches show good performance for long indel alignment and can achieve high sensitivity. Thanks to improvements in the computational efficiency of global

sequence alignment algorithms, tools such as minimap2 [31] and AnchorWave [23,30] now implement this collinear strategy.

To speed up alignment and increase specificity, pairwise WGAs between individuals that share a karyotype are performed separately for each homologous chromosome, using chromosomes with identical names [33]. This strategy requires collinearity between chromosomes, and will not produce an alignment of sequences that have translocated to a different chromosome. AnchorWave [30] implements a function to conduct an end-to-end alignment for each pair of homologous chromosomes and achieves great sensitivity. This approach treats an intra- or inter-chromosome translocation as a deletion at the donor and an insertion at the recipient, and copy-number variations or tandem repeats are treated as indels.

Multiple genome alignment

WGA typically only compares two taxa, but many evolutionary inferences are improved by sampling multiple taxa. Multiple sequence alignment (MSA) is a necessary starting point for MGA. The outputs of MSA are routinely applied to numerous phylogenetic and evolutionary analyses by aligning a group of preselected homologous sequences, often genes. Such MSA approaches generally produce an alignment along the whole length of each input sequence, analogous to global alignment [34–38]. MSA requires additional steps beyond simply combining a set of pairwise sequence alignments. This allows MSA to unify multi-isoform indels [39], especially when applied to the investigation of sequence conservation or variant calling across multiple individuals.

For almost all MSA problems it is not practical to utilize classical dynamic programming approaches because of the exorbitantly high CPU time and memory costs – 1 petabyte (PB) of memory storage is necessary to align a 100 kb region in only three samples (Table S1). To make MSA possible, a **progressive alignment** algorithm [38] is widely implemented to generate an approximate global optimal alignment, for example by MAFFT [34], MUSCLE [35], CLUSTAL [37], and T-coffee [36]. These progressive MSA approaches have been extended to the genomic scale with additional functions to resolve genomic rearrangements. For example, the ROAST [40] pipeline progressively creates multi-species alignments from the output of pairwise alignments. The Cactus [41] software also follows a progressive strategy to reconstruct ancestral genomes by combining subalignments, although the performance of Cactus on plant species [42] may require new developments to deal with polyploidy and high sequence diversity. The ideal output of MGA is chromosome or **subgenome** alignments that can capture the relatedness between individuals. However, the features of many plant genomes mean that careful consideration of the desired output is essential before adopting these state-of-the-art genome alignment technologies.

Plant genomes organize chromatin differently from mammal genomes

Plant genomes vary extensively in size. The impact of differences in genome size is minimized by the folding and packaging of chromosomes into different regions of the nucleus. In plants with smaller genomes, actively transcribed genes are often found on euchromatic chromosome arms whereas transcriptionally repressed sequences are found in pericentric heterochromatin. This straightforward delimitation of chromosomes breaks down with larger genomes because genic euchromatin becomes structured into local territories within the nucleus.

This genome organization means that *cis*-regulatory elements do not necessarily regulate the nearest gene [43] and multiple neighboring genes can be coregulated [44]. Coexpressed genes are highly likely to be functionally related [44], and the conservation of collinear blocks across species is therefore an important signal that an aligner should recover. A plant genome aligner should thus identify collinear blocks of genes, thus allowing comparison of sequence

evolution for both genic sequences and intergenic sequences. This strategy has been implemented in AnchorWave [30] and is also proposed in NGSEP 4 [45]. Aligning collinear regions can also help to identify distal regulatory elements.

WGDs followed by fractionation are common in plants

When similar sequences are present at different dispersed locations in the genome, an aligner must decide at which of many potential positions to report an alignment. For many applications it is most parsimonious to assume one-to-one orthology between sequences – that each homologous sequence has only one matching position in the other genome. This assumption is not always a sufficient model for plant species with polyploidy in their history and requires WGD-aware genome alignment techniques.

In newly formed polyploids, each subgenome is present on a separate chromosome, and we can readily separate the parental contributions to each subgenome. Over deeper evolutionary time, polyploids begin to rediploidize as their parental chromosomes are reshuffled and rearranged through recombination and structural rearrangements. Duplicate genes can be redundant in a newly formed polyploid, but over time these genes can accrue inactivating mutations and be lost. Alternatively, these duplicates can be retained or gain differential or new functions through subfunctionalization or neofunctionalization.

Polyploidy events complicate WGA, and it is important to understand past WGD events when selecting an alignment approach. One simple way to detect unshared WGDs between assembled genomes is to align genes between two genomes and draw syntenic dot-plots [46]. If one lineage has undergone a WGD, from diploid to tetraploid, genes will align at two positions in the haploid genome. This syntenic dot-plot approach can break down when there are extensive chromosome fusions or rearrangements after the WGD. In these cases with little syntenic conservation, examining the distribution of synonymous substitutions per site (Ks) at pairs of genes within a genome is often used to detect very ancient WGDs [47].

Once an unshared WGD has been identified, we need to align multiple paralogous regions to the same orthologous segment. When aligning a recent polyploid to diploid relatives, we can conduct alignment for each subgenome separately by independently aligning each chromosome. For older polyploid species the subgenomes may not be easily separated before WGA, as is the case for WGA between maize and sorghum [48] or between soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*) [49]. To identify and separate these subgenomes, genome synteny technologies such as quota-alignment [50], MCScanX [51], and CoGe [52] use coding genes as anchors and identify blocks of genes in the same order in both genomes and subgenomes [51]. These approaches make it possible to identify chromosomal rearrangements and gene PAVs. The AnchorWave [30] program refines this idea to generate base-pair resolution WGAs within these subgenomes.

Progressive MGA uses phylogenetic trees to guide alignment [41], but the relationship between subgenomes can vary along the genome because of recombination, gene conversion, and **fractionation**. This means that, along a chromosome, ancestry may differ by region between allopolyploid progenitor subgenomes [53]. Methods have been developed to separate polyploid subgenomes, for example, GENESPACE, CoGe, POInT, SubPhaser [48,54–57]. Once these subgenomes have been identified, they can be used as separate individuals in an MGA. Despite a clear conceptual framework, there are no commonly used methods to perform those complex analyses automatically. Performing base-pair resolution MGA while incorporating the subgenomes of polyploids has rarely been performed and requires manual insights and curation.

Many size classes of indels cause polymorphisms between plant genomes

Indel variation has a greater impact than SNPs on base-pair differences between the genomes of different individuals of the same or different species [58], and can have a significant influence on traits. The distributions of indel lengths are not uniform in mammal and plant species, and multiple peaks can be seen in the density distributions of indel length [59,60]. Several mutational mechanisms can generate short indels, such as DNA polymerase errors or imperfect repair following DNA damage [61]. Long indels are thought mainly caused by TE activity [62] or non-allelic homologous recombination [63]. TE activity in most mammal species is low, and there is not too much TE variation between for example two human genomes or even between the human genome and the chimpanzee genome [6]. Lineage-specific amplification and potential deletion of TEs are common in plants, even among closely related species or different accessions of the same species [30,64,65]. Because TEs evolve faster than their host genome, it can be difficult to predict which TEs are present, or even to determine the length distribution of TEs in an individual genome. Further, common repeat mask approaches identify TEs, meaning that TE annotations often represent short fragments of TEs, and not the entire length of an indel.

One way to identify long indels is to identify positions where sequence alignment does not occur. For example, SyRI [33], Assemblytics [66], and other custom pipelines [67–69] consider each pair of consecutive alignments along a sequence and identify indels by considering the spacing and orientation between these alignments. These approaches often limit indel length to a maximum of 10 kb. To generate alignments for long indels and perform variant calling directly from sequence alignment, minimap2 [31] and AnchorWave [30] combine a global sequence alignment algorithm with a two-piece affine gap cost penalty. The computational time cost of affine gap cost is 1.67-fold higher than a linear gap cost, and the time cost of a two-piece affine gap cost is threefold higher than a linear gap cost (Table S2 and Figure 2C). Further improvements in long indel alignment could integrate the knowledge of length distributions of TE superfamilies [65] and optimize gap penalty parameters for the type of TE that is present in the underlying sequence. Because each additional gap cost increases the computational cost underlying the currently available dynamic programming algorithms, a more efficient algorithm is necessary to further optimize indel alignment.

Plant genomes have high nucleotide diversity

WGA is difficult when sequences diverge because there are fewer invariant base pairs that support the shared orthology of sequences. In general, plant genomes are considered to be more dynamic than the relatively stable animal genomes [6]. Previous studies suggest that the nucleotide diversity between different maize lines is even greater than that between humans and chimpanzees [70]. There is an additional 3.4% difference between humans and chimpanzees as a result of indels [71], but this is significantly less than that between two maize lines [30].

Moreover, the regulatory architecture of plant genomes appears to tolerate more variation than animal genomes. Studies of conserved noncoding sequences (CNSs) report that plant genomic sequences are more diverse than those of mammals [72–75]. CNSs are functionally constrained sequences that likely play roles in genome expression regulation [25,76]. One way to identify regulatory sequences is to focus on ultraconserved elements of at least 100 identical base pairs. There are several tenfold fewer ultraconserved elements in plant genomes than in animal genomes [77]. In animals, many CNSs are large (>100 bp) [78], whereas experimentally identified plant *cis*-elements are infrequently >30 bp, and have a median observed length of 8 bp [79]. The average size of a plant transcription factor binding site is only 6.8 bp [80,81]; this is smaller than the *k*-mer size used by the widely adopted seed-and-extend sequence alignment approach. In general, plant genome comparisons need a much higher sensitivity than comparisons of mammalian genomes.

Alignment parameters, nucleotide substitution scoring, and gap penalties have a substantial effect on the performance of sequence alignment applications. In contrast to simple match/mismatch scores, scoring matrices such as HOXD [82], and BISR [83] provide informative options to improve sensitivity and specificity. It is well known that different regions have different sequence diversity over the genome. For example, there are generally fewer SNPs and indels in genic regions than in intergenic regions. Whether this is a result of natural selection or is due to differences in mutation rate between regions of the genome remains an open question [84,85]. Nonetheless, these observed differences mean that using a single parameter set will be suboptimal. A recent test of different parameters using Poaceae genomes suggested that the optimal parameters differ between genomic regions [28]. How to automatically optimize scoring parameters efficiently and precisely for different genomic regions remains to be explored.

Multiallelic variants are common among plant genomes

When more than two alleles are present at a locus, methods for understanding genetic variation become complex. Population genetic models are built on assumptions of biallelic sites. Even classic cases of multiallelic variation such as human ABO blood groups do not simply comprise three segregating alleles but are underlain by dozens of haplotypic variants segregating in populations [86]. In plants, such allelic heterogeneity is common; for example, the loss-of-function alleles of *RDO5* [87] in *Arabidopsis* populations. Although it is most common to consider that SNPs determine genetic variation between different individuals, other types of variants can affect a much larger fraction of the genome, at least in plant populations. For example, a 3.2 Mb indel has been validated between two maize individuals (B73 and Mo17) [88], and ~30% of their genomes differ by indels or translocations [30]. Inversions reported between maize genomes often span multiple megabases [89,90]. A similar pattern has also been reported in rice [91]. Owing to the high prevalence of indels and structural variation in plant genomes, a large proportion of SNPs occur at positions that overlap these regions and are thus multiallelic (Figure 3A).

Multiallelic variants have always been simplified as biallelic when using short-read sequencing for plant population genotyping [92]. For population resequencing studies, each genomic variant is typically identified relative to the reference allele by using **short-read alignment**. When an individual sample lacks read coverage at a specific variant site, this may reflect a structural variation

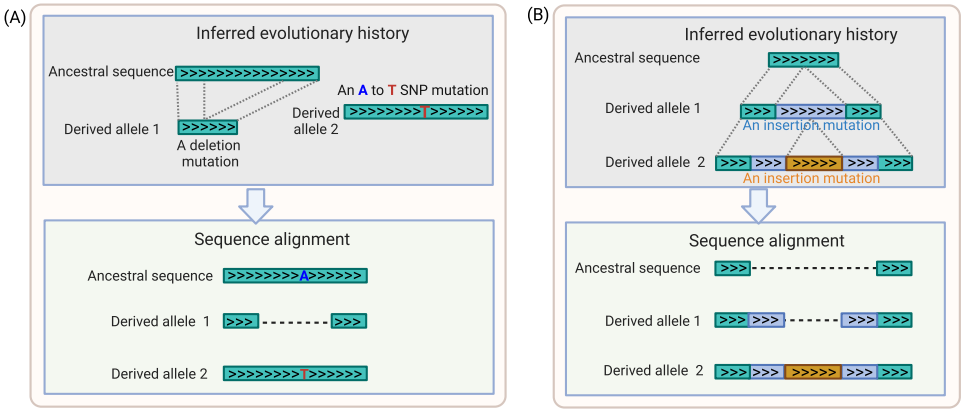

Figure 3. Multiallelic variants are common among plant genomes, and multiple genome alignment is essential to uncover their evolution. (A) A deletion can overlap with a single-nucleotide polymorphism (SNP). (B) An insertion can be followed by another insertion.

such that the region is absent. We often lose this information when imputation is applied to assign a reference allele or alternative allele to the missing site based on linkage disequilibrium.

In natural plant populations, indels, inversions, and translocations of diverse lengths frequently overlap. For example, TEs often insert into pre-existing TE sequences [65,93], generating nested TEs [65] (Figure 3B). As the inserted sequences, translocated sequences, and inversion sequences continue to accumulate, all types of mutations can take place within those regions. It is reasonable to assume that nested SNPs or short indels have different functional impacts compared to unnested variants. Given that long indels and inversions affect a large proportion of the genomes in a population, nested variants are very common in plants. One of the advantages of *de novo* genome assembly and MGA over short-read variant calling is the ability to call nested variants.

Solutions to represent and utilize those multiallelic variants can come from well-designed reference-free MGA and graph genomes. An MGA used for plant genomes should be able to cope with WGD followed by chromosome fusion, high sequence diversity, and high TE activity. Using a single linear reference genotype makes it difficult to report nested multiallelic variants. The graph genome representations of allelic variation will bring us closer to the goal of connecting genotype to phenotype and identifying causal variants that genome editing can repair.

Graph genome methods encode genetic variants as nodes and edges, and preserve the contiguity of the sequence and structural variation between individuals. The graph model has proved to be a powerful tool for dealing with complexity of genome-scale sequence alignment, and this data structure has been implemented in a wide range of MGA tools [41,94]. In addition, graphs provide a straightforward way to represent similarities and changes between genomes, and thus can visualize alignments as well as compute them in parallel. Given the decreasing cost of genome assemblies, graph data structures are increasingly essential for performing population-scale MGA assemblies efficiently and accurately, especially for plant genomes because of their high complexity. A more comprehensive review of the graph data structure for WGA is given in [95]. In the recent decade graph pangenomes have been constructed for key species to provide a reference for population-scale short-read mapping. With carefully selected accessions for genome assembly and graph genome construction, this technology gives short reads the ability to call complex or long variants approximate to alignment using *de novo* assembled genomes. The graph reference genome technology is expected to be commonly used in the near future once the cost of genome assembly falls sufficiently to replace the short-read technology.

Concluding remarks and future perspectives

When WGA was initially described, obtaining the primary sequences constituted the bottleneck. With rapid improvements in genome sequencing and assembly technologies, the bottleneck has now shifted to WGA. More species and larger genomes have been generated with high continuity, low error rates, and few gaps [14], including multiple to dozens of individuals of arabidopsis (*Arabidopsis thaliana)* [94] and crop species [95–99]. Given these rapid advances, the final goal of the Earth BioGenome Project – 'to sequence and annotate the genomes of all currently known eukaryotic species in 10 years' [100–102] – is likely to be achieved for plants [103]. The comprehensive sample of 300 000 plant species [104] are spread across the globe and have adapted to numerous environments. The growing reality of phylogenetic saturation of sequenced taxa is pushing evolutionary analysis to hundreds or thousands of phylogenetically close species. WGA is crucial to interpret these genomes.

WGA elucidates the evolution of the sequences we are comparing, but until recently no alignment tools were available that could recover the processes that have dominated plant evolution.

Outstanding questions

What is the tradeoff between close taxonomic sampling and deeper divergences? Does genome divergence co-occur with speciation?

How best to perform multiple alignments for related plant species?

What is the best way to represent sequence variation and to use multiallelic variants for population or quantitative genetics research?

How should we use pangenomes, and how best to map short reads to diverse reference genomes? Do we pick the most closely related individual or species for which we have a reference genome and proceed via WGA?

Polyploidy – when we have two copies that align to a haploid reference, how do we interpret variants and phase across chromosomes in our alignment files, and how do we interpret subgenomes?

Affine gap costs – can we do more? Should we use the TE landscape to decide what indel lengths are likely? Alternatively, do we assign alignment parameters based on TE annotation?

How to utilize multi-allelic variants for quantitative genetics analysis to narrow down the missing heritability? How to adjust the population genetics models to uncover the evolution and natural selection acting on multi-allelic variants?

Aligners that take this underlying biological variation into account should improve the sensitivity and accuracy of genome alignments. Features ranging from chromatin structure to indel abundance to nucleotide diversity all raise profound technological challenges for plant WGA. In many plant genomes TEs are heavily methylated [105], which leads to an extreme transition/transversion bias likely arising from deamination of cytosines [106]. One strategy may be to use a different substitution matrix or gap penalties so as to use the TE annotation of one or both genomes to guide alignment parameters. In the field of genome comparison, deep learning has been implemented in the latest version of commonly used variant-calling pipelines based on short reads. Although deep learning has considerable potential to improve the alignment of *de novo* assembled genomes, it has not yet been well explored.

Some conceptual challenges remain, such as the representation of nested and highly complex variation that accompanies WGD variation. Recent advances such as graph-based pangenome analysis have the potential to represent complex variants. However, graph genomes and Cactus have not been easily applicable plant genomes to date. This may be due to the higher diversity of plant genomes and the complexity introduced by dispersed repeats and larger repeated blocks arising from polyploidy. These limitations apply to implementation and software design, and not to the concepts underlying these approaches. For example, graph-based genomes have memory costs that grow with sequence diversity. Implementations can limit graph nodes to shorter haplotype blocks such as sequences conserved across individuals or genic sequences [107]. Graphs may need to ignore SNP variants in nodes that represent a particular TE that is found thousands of times in the genome. Similarly, MGAs may need to arbitrarily pick an artificial subgenome assignment to allow multiple allelic and subgenome alignments much like separate individual genome assemblies. This is straightforward when genome assemblies are haplotype resolved, but becomes more complex when the genomic contigs remain unphased. Switches in ancestry between subgenomes can be difficult to encode in an MGA framework. Similar problems in population genetics have been approached with ancestral recombination graphs (ARGs) that allow the reconstruction of local trees of related haplotypes along the genome. Implementations that take advantage of this tree sequence encoding should be extended to interspecific comparisons and may help to answer interesting questions in polyploids, such as how often gene conversion occurs between subgenomic copies, or how often crossovers occur between subgenomes in a recently formed polyploid.

The continued development of plant genome alignment tools will have a profound impact on the accurate identification of all variants and can help to identify causal variants – and thus truly bridge genotype and phenotype. The coming genome assemblies of all the plant species and large populations should be coupled with further development of genome alignment technology to advance our understanding of plant genetics and evolution (see Outstanding questions).

Acknowledgments

This project is supported by the National Natural Science Foundation of China (grant number 31900486), Shandong Provincial Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) (grant number 2023HWYQ-109) and the National Science Foundation (grant number 1822330). M.C.S. is supported by an National Science Foundation Postdoctoral Research Fellowship in Biology (grant number 1907343). We thank Merritt Khapho-Burch and other members of the laboratory of E.S.B. (Cornell University) as well as members of the laboratory of B.S. (Peking University Institute of Advanced Agricultural Sciences) for helpful discussions.

Declaration of interests

The authors declare no competing interests.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tplants.2023.08.013

References

1. Kille, B. *et al.* (2022) Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biol.* 23, 182
2. Schwartz, S. *et al.* (2003) Human–mouse alignments with BLASTZ. *Genome Res.* 13, 103–107
3. Marçais, G. *et al.* (2018) MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* 14, e1005944
4. Kiełbasa, S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493
5. Earl, D. *et al.* (2014) Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Res.* 24, 2077–2089
6. Murat, F. *et al.* (2012) Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* 4, 917–928
7. Fleischmann, A. *et al.* (2014) Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* 114, 1651–1663
8. Pellicer, J. *et al.* (2010) The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* 164, 10–15
9. Wood, T.E. *et al.* (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U. S. A.* 106, 13875–13879
10. Kidwell, M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115, 49–63
11. Bourque, G. *et al.* (2018) Ten things you should know about transposable elements. *Genome Biol.* 19, 199
12. Lisch, D. (2012) How important are transposons for plant evolution? *Nat. Rev. Genet.* 14, 49–61
13. Chen, J. *et al.* (2017) Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol. Biol. Evol.* 34, 1417–1428
14. Varshney, R.K. *et al.* (2021) Designing future crops: genomics-assisted breeding comes of age. *Trends Plant Sci.* 26, 631–649
15. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453
16. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197
17. Chao, K.M. *et al.* (1992) Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* 8, 481–487
18. Hirschberg, D.S. (1975) A linear space algorithm for computing maximal common subsequences. *Commun. ACM* 18, 341–343
19. Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. *Bioinformatics* 4, 11–17
20. Farrar, M. (2007) Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* 23, 156–161
21. Suzuki, H. and Kasahara, M. (2018) Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinforma.* 19, 45
22. João Jr., M. *et al.* (2019) On the parallelization of Hirschberg's algorithm for multi-core and many-core systems. *Concurr. Comput.* 31, e5174
23. Marco-Sola, S. *et al.* (2020) Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* 37, 456–463
24. Clausen, P.T.L.C. *et al.* (2018) Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinforma.* 19, 307
25. Song, B. *et al.* (2021) Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Res.* 31, 1245–1257
26. Ebel, M. *et al.* (2022) Global, highly specific and fast filtering of alignment seeds. *BMC Bioinforma.* 23, 225
27. Sun, Y. and Buhler, J. (2006) Choosing the best heuristic for seeded alignment of DNA sequences. *BMC Bioinforma.* 7, 133
28. Wu, Y. *et al.* (2022) A multiple alignment workflow shows the effect of repeat masking and parameter tuning on alignment in plants. *Plant Genome* 15, e20204
29. Haas, B.J. *et al.* (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643–3646
30. Song, B. *et al.* (2022) AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2113075119
31. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100
32. Pevzner, P. and Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.* 13, 37–45
33. Goel, M. *et al.* (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20, 277
34. Katoh, K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066
35. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797
36. Notredame, C. *et al.* (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217
37. Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
38. Feng, D.-F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *J. Mol. Evol.* 25, 351–360
39. Song, B. *et al.* (2018) Recovery of novel association loci in *Arabidopsis thaliana* and *Drosophila melanogaster* through leveraging INDELs association and integrated burden test. *PLoS Genet.* 14, e1007699
40. Blanchette, M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715
41. Armstrong, J. *et al.* (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 587, 246–251
42. Wu, Y. *et al.* (2023) Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* 186, 2313–2328
43. Salvi, S. *et al.* (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11376–11381
44. Michalak, P. (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* 91, 243–248
45. Tello, D. *et al.* (2023) NGSEP 4: efficient and accurate identification of orthogroups and whole-genome alignment. *Mol. Ecol. Resour.* 23, 712–724
46. Lyons, E. *et al.* (2008) The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop. Plant Biol.* 1, 181–190
47. Tiley, G.P. *et al.* (2018) Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol. Evol.* 10, 2882–2898
48. Schnable, J.C. *et al.* (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4069–4074
49. Schmutz, J. *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713
50. Tang, H. *et al.* (2011) Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinforma.* 12, 102
51. Wang, Y. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49
52. Lyons, E. *et al.* (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 148, 1772–1781

53. Estep, M.C. *et al.* (2014) Allopolyploidy, diversification, and the Miocene grassland expansion. *Proc. Natl. Acad. Sci. U. S. A.* 111, 15149–15154
54. Lyons, E. and Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673
55. Emery, M. *et al.* (2018) Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet.* 14, e1007267
56. Lovell, J.T. *et al.* (2022) GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* 11, e78526
57. Zhang, R.-G. *et al.* (2023) Subgenome-aware analyses suggest a reticulate allopolyploidization origin in three *Papaver* genomes. *Nat. Commun.* 14, 2204
58. Huddleston, J. *et al.* (2017) Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685
59. Bennetzen, J.L. *et al.* (2005) Mechanisms of recent genome size variation in flowering plants. *Ann. Bot.* 95, 127–132
60. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87
61. Garcia-Diaz, M. and Kunkel, T.A. (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem. Sci.* 31, 206–214
62. Mun, S. *et al.* (2021) A study of transposable element-associated structural variations (TASVs) using a de novo-assembled Korean genome. *Exp. Mol. Med.* 53, 615–630
63. Parks, M.M. *et al.* (2015) Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biol.* 16, 72
64. Jedlicka, P. *et al.* (2020) What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front. Plant Sci.* 11, 644
65. Stitzer, M.C. *et al.* (2021) The genomic ecosystem of transposable elements in maize. *PLoS Genet.* 17, e1009768
66. Nattestad, M. and Schatz, M.C. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023
67. Anderson, S.N. *et al.* (2019) Transposable elements contribute to dynamic genome content in maize. *Plant J.* 100, 1052–1065
68. Stuart, T. *et al.* (2016) Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *Elife* 5, e20777
69. Chakraborty, M. *et al.* (2018) Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* 50, 20–25
70. Buckler, E.S. and Stevens, N.M. (2006) Maize origins, domestication, and selection. In *Darwin's Harvest* (Motley, T.J. *et al.*, eds), pp. 67–90, Columbia University Press
71. Britten, R.J. (2002) Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13633–13635
72. Thomas, B.C. *et al.* (2007) *Arabidopsis* intragenomic conserved noncoding sequence. *Proc. Natl. Acad. Sci. U. S. A.* 104, 3348–3353
73. Baxter, L. *et al.* (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell* 24, 3949–3965
74. Haudry, A. *et al.* (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* 45, 891–898
75. Turco, G. *et al.* (2013) Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. *Front. Plant Sci.* 4, 170
76. Yocca, A.E. *et al.* (2021) Evolution of conserved noncoding sequences in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 38, 2692–2703
77. Reneker, J. *et al.* (2012) Long identical multispecies elements in plant and animal genomes. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1183–E1191
78. Stephen, S. *et al.* (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.* 25, 402–408
79. Haberer, G. *et al.* (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between *Arabidopsis* and *Brassica oleracea*. *Plant Physiol.* 142, 1589–1602
80. Tu, X. *et al.* (2020) Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.* 11, 5089
81. O'Malley, R.C. *et al.* (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* 165, 1280–1292
82. Frith, M.C. *et al.* (2010) Parameters for accurate genome alignment. *BMC Bioinforma.* 11, 80
83. Frith, M.C. *et al.* (2012) A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res.* 40, e100
84. Charlesworth, B. and Jensen, J.D. (2023) Population genetic considerations regarding evidence for biased mutation rates in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 40, msac275
85. Monroe, J.G. *et al.* (2022) Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* 602, 101–105
86. Yip, S.P. (2002) Sequence variation at the human ABO locus. *Ann. Hum. Genet.* 66, 1–27
87. Xiang, Y. *et al.* (2016) Sequence polymorphisms at the REDUCED DORMANCY5 pseudophosphatase underlie natural variation in *Arabidopsis* dormancy. *Plant Physiol.* 171, 2659–2670
88. Huang, Y. *et al.* (2021) Megabase-scale presence-absence variation with Tripsacum origin was under selection during maize domestication and adaptation. *Genome Biol.* 22, 237
89. Liu, J. *et al.* (2020) Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* 21, 121
90. Yang, N. *et al.* (2019) Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* 51, 1052–1059
91. Fuentes, R.R. *et al.* (2019) Structural variants in 3000 rice genomes. *Genome Res.* 29, 870–880
92. Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893–903
93. Fedoroff, N.V. (2012) Presidential address. Transposable elements, epigenetics, and genome evolution. *Science* 338, 758–767
94. Jiao, W.-B. and Schneeberger, K. (2020) Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11, 989
95. Hufford, M.B. *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* 373, 655–662
96. Zhou, Y. *et al.* (2022) Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606, 527–534
97. Shang, L. *et al.* (2022) A super pan-genomic landscape of rice. *Cell Res.* 32, 878–896
98. Liu, Y. *et al.* (2020) Pan-genome of wild and cultivated soybeans. *Cell* 182, 162–176
99. He, Q. *et al.* (2023) A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* 55, 1232–1242
100. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* 115, 4325–4333
101. Exposito-Alonso, M. *et al.* (2020) The Earth BioGenome Project: opportunities and challenges for plant genomics and conservation. *Plant J.* 102, 222–229
102. Lewin, H.A. *et al.* (2022) The Earth BioGenome Project 2020: starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115635118
103. Kress, W.J. *et al.* (2022) Green plant genomes: what we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115640118
104. Christenhusz, M.J.M. and Byng, J.W. (2016) The number of known plants species in the world and its annual increase. *Phytotaxa* 261, 201–217
105. Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476

106. Carpenter, M. *et al.* (2006) Sequence-dependent enhancement of hydrolytic deamination of cytosines in DNA by the restriction enzyme PspGI. *Nucleic Acids Res.* 34, 3762–3770
107. Bradbury, P.J. *et al.* (2022) The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. *Bioinformatics* 38, 3698–3702
108. Olson, N.D. *et al.* (2023) Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* 24, 464–483
109. Chaisson, M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1784
110. Valls-Margarit, J. *et al.* (2022) GCATlPanel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *Nucleic Acids Res.* 50, 2464–2479
111. Mahmoud, M. *et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biol.* 20, 246
112. Karakoc, E. *et al.* (2011) Detection of structural variants and indels within exome data. *Nat. Methods* 9, 176–178
113. Gardner, E.J. *et al.* (2021) Detecting cryptic clinically relevant structural variation in exome-sequencing data increases diagnostic yield for developmental disorders. *Am. J. Hum. Genet.* 108, 2186–2194
114. Guan, P. and Sung, W.-K. (2016) Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* 102, 36–49
115. Heller, D. and Vingron, M. (2019) SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35, 2907–2915
116. Sedlazeck, F.J. *et al.* (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15, 461–468
117. Miga, K.H. *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84
118. Logsdon, G.A. *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature* 593, 101–107
119. Song, J.-M. *et al.* (2021) Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* 14, 1757–1767
120. Zhang, Y. *et al.* (2022) The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. *Plant Biotechnol. J.* 20, 1642–1644
121. Chen, J. *et al.* (2023) A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* 55, 1221–1231
122. Deng, Y. *et al.* (2022) A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* 15, 1268–1284
123. Wang, B. *et al.* (2021) High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads. *Genomics Proteomics Bioinforma.* 20, 4–13
124. Schmalenbach, I. *et al.* (2014) Functional analysis of the *Landsberg erecta* allele of FRIGIDA. *BMC Plant Biol.* 14, 218