Genomics Proteomics Bioinformatics 21 (2023) 427–439

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Recent Advances in Assembly of Complex Plant Genomes

Weilong Kong, Yibin Wang, Shengcheng Zhang, Jiaxin Yu, Xingtan Zhang *

Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

Received 18 March 2023; revised 18 March 2023; accepted 7 April 2023
Available online 25 April 2023

Handled by Kai Ye

KEYWORDS

Complex plant genome;
Assembly algorithm;
Telomere-to-telomere genome;
Haplotype-resolved assembly;
Sequencing technology

Abstract Over the past 20 years, tremendous advances in sequencing technologies and computational algorithms have spurred plant genomic research into a thriving era with hundreds of genomes decoded already, ranging from those of nonvascular plants to those of flowering plants. However, **complex plant genome** assembly is still challenging and remains difficult to fully resolve with conventional sequencing and assembly methods due to high heterozygosity, highly repetitive sequences, or high ploidy characteristics of complex genomes. Herein, we summarize the challenges of and advances in complex plant genome assembly, including feasible experimental strategies, upgrades to **sequencing technology**, existing assembly methods, and different phasing algorithms. Moreover, we list actual cases of complex genome projects for readers to refer to and draw upon to solve future problems related to complex genomes. Finally, we expect that the accurate, gapless, telomere-to-telomere, and fully phased assembly of complex plant genomes could soon become routine.

Introduction

High-quality genome assembly establishes a reference for exploring the evolutionary history and genetic mechanisms of complex traits and facilitates molecular breeding and genomics studies. Fast-growing sequencing technologies and algorithm innovations have promoted breakthroughs in both animal and plant genomes to date. However, assemblies of plant genomes are much more challenging than those of animal genomes, because most animal genomes are diploid and contain fewer repetitive sequences compared to plant genomes [1,2]. In contrast, plant genomes span several orders of magnitude in size, vary in ploidy and heterozygosity levels, and contain a large number of different types of repeats (35%–90% of the genome) [3]. Since the first plant genome release for the model plant *Arabidopsis thaliana* in 2000 [4], more than 800 plant genomes have been published to date, including the genomes of eudicots, monocots, gymnosperms, ferns, lycophytes, bryophytes, charophytes, and chlorophytes [5]. However, most of these published plant genomes are simple genomes characterized by $< 0.8\%$ heterozygosity, and/or $< 60\%$ repetitive sequences, whereas chromosome-scale assemblies of plant genomes with highly repetitive sequences,

* Corresponding author.
E-mail: zhangxingtan@caas.cn (Zhang X).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2023.04.004

high heterozygosity, or polyploid genomes are incredibly scarce.

Resolving highly repetitive sequences is meaningful for understanding genome evolution and for mining functional elements. For instance, many plants are dioecious with a newly evolved Y chromosome. The suppressed recombination region in the Y chromosome accumulates a large number of mobile elements, which could account for more than 90% of the examined region [6–8]. Although some sex determination factors have been identified in a few plant species, including papaya [9], poplar [10], and fig trees [11], the assembly of highly repetitive sequences poses notable challenges, hindering the discovery of sex determination mechanisms in a massive number of dioecious plants. In addition, nearly 80% of plants have undergone whole-genome duplication(s), and many of them still maintain polyploidy with a high level of heterozygosity among haplotypes [1]. Polyploidy is considered the main force of plant evolution [2], contributing to many well-known crops that humans rely on for survival, including wheat (*Triticum aestivum*), rape (*Brassica napus*), upland cotton (*Gossypium hirsutum*), peanut (*Arachis hypogaea*), strawberry (*Fragaria ananassa*), potato (*Solanum tuberosum*), banana (*Musa* spp.), and sugarcane (*Saccharum officinarum*).

Given the importance of these plant species, complex plant genome sequencing has been an emerging frontier in the genomics field. Recently, single-molecule sequencing (SMS) technologies and haplotype assembly algorithms have efficiently generated chromosome-scale and haplotype-phased complex genome assemblies for a few species, including potato [12–16], sugarcane [17,18], and alfalfa [19]. Herein, we summarize the challenges of complex genome assembly, the advantages of SMS platforms, and newly developed assembly algorithms in complex plant genome assembly, aiming to provide a comprehensive reference facilitating future genome projects.

Challenges of complex genome assembly

High repetitive sequence content

Repetitive sequences that are similar or identical to sequences elsewhere in the genome represent an important and pervasive part of the dark matter of genomes [20,21]. Many plant genomes are filled with repetitive sequences, including various satellites, rDNA, short interspersed nuclear elements, long interspersed nuclear elements, long terminal repeat (LTR) retrotransposons, and DNA transposons [22]. For instance, the total repetitive sequences account for ~ 85% in maize genome [23] and in the wheat genome [24], and 74%–80% in the tea plant genome [25]. These different types of repetitive sequences contain anywhere from two copies to millions of copies, ranging from 1–2 bases (mono- and dinucleotide repeats) to millions of bases [20,26]. These repeat-rich regions usually involve many important genetic functional regions, namely, telomeres, centromeres, multicopy genes, and non-recombining and highly heterochromatic chromosomes such as the Y and W sex chromosomes [11,27]. Given that these regions play essential roles in the function and evolution of the genome [28–31], the need to precisely assemble them has become a hurdle in complex genome studies.

However, repetitive sequences with hundreds or thousands of repeat units are widely distributed in genomes and cover an ultralong genomic region (such as nest LTRs, which can span 20–200 kb) that cannot be spanned by even long reads generated by SMS. In assemblies with short reads (35–800 bp) as input, nearly identical tandem repeats usually fold into fewer copies (*i.e.*, collapsed assembly, **Figure 1**A), making it difficult to determine the true number of copies. Similarly, unzipping two identical interspersed repeat units from the assembly graph can produce false joins with flanking unique sequences, leading to chimeric and fragmented contigs (Figure 1B). Due to improvements in the length of SMS reads, many repeat regions can now be well resolved, except for the extremely long and complex repeat regions. However, the high level of sequencing errors in SMS reads poses challenges to accurately distinguish minor variants among repetitive sequences from sequencing errors. On the one hand, low-frequency genetic variants with frequencies lower than the sequencing error rate may be mistaken for sequencing error, leading to the underestimation of low-frequency genetic variations. On the other hand, the sequencing error will be mistaken for the genetic variants, resulting in misassembly when sequencing errors are not corrected (Figure 1C).

High heterozygosity

In low-heterozygosity species, the variations between the two haplotypes are mainly small-scale. These small-scale variations enable accurate alignment during assembly, resulting in consensus sequences (**Figure 2**A). However, genomes with high heterozygosity contain many large-scale structural variations between the two haplotypes, leading to assembly 'bubbles' that represent redundant allelic sequences (Figure 2B). Many plants have high genome heterozygosity due to distant hybridization and self-incompatibility. Therefore, the assembly of these highly heterozygous genomes usually generates a larger size than the estimated size of the haploid genome.

Polyploidy

Plant polyploidizations originate either from whole-genome duplication of a single species (autopolyploidy) or interspecific hybridization followed by chromosome doubling (allopolyploidy) [32,33]. The assembly of allopolyploid genomes is less complex, and the first wave of polyploid genome assemblies has mainly involved allopolyploid crops, such as rape (*B. napus*), cotton (*G. hirsutum*), and peanut (*A. hypogaea*). It is relatively easy to distinguish subgenomes originating from different ancestral species because they have maintained a large proportion of genetic variations during their long evolutionary history. However, autopolyploid organisms consisting of more than two homologous sets of chromosomes pose significant challenges in genome assembly and haplotype phasing due to the high similarity between homologous chromosomes [34].

For example, the autotetraploid genome has four similar haplotypes that contain a large proportion of nearly identical sequences. Linking these identical sequences has a tendency to generate chimeric contigs with switch errors or false duplications (**Figure 3**A). These chimeric contigs confuse high-throughput/resolution chromosome conformation capture (Hi-C) signals, resulting in erroneous scaffolds that mess up sequences from different haplotypes (Figure 3B). In addition, nearly identical homologous sequences between different hap-

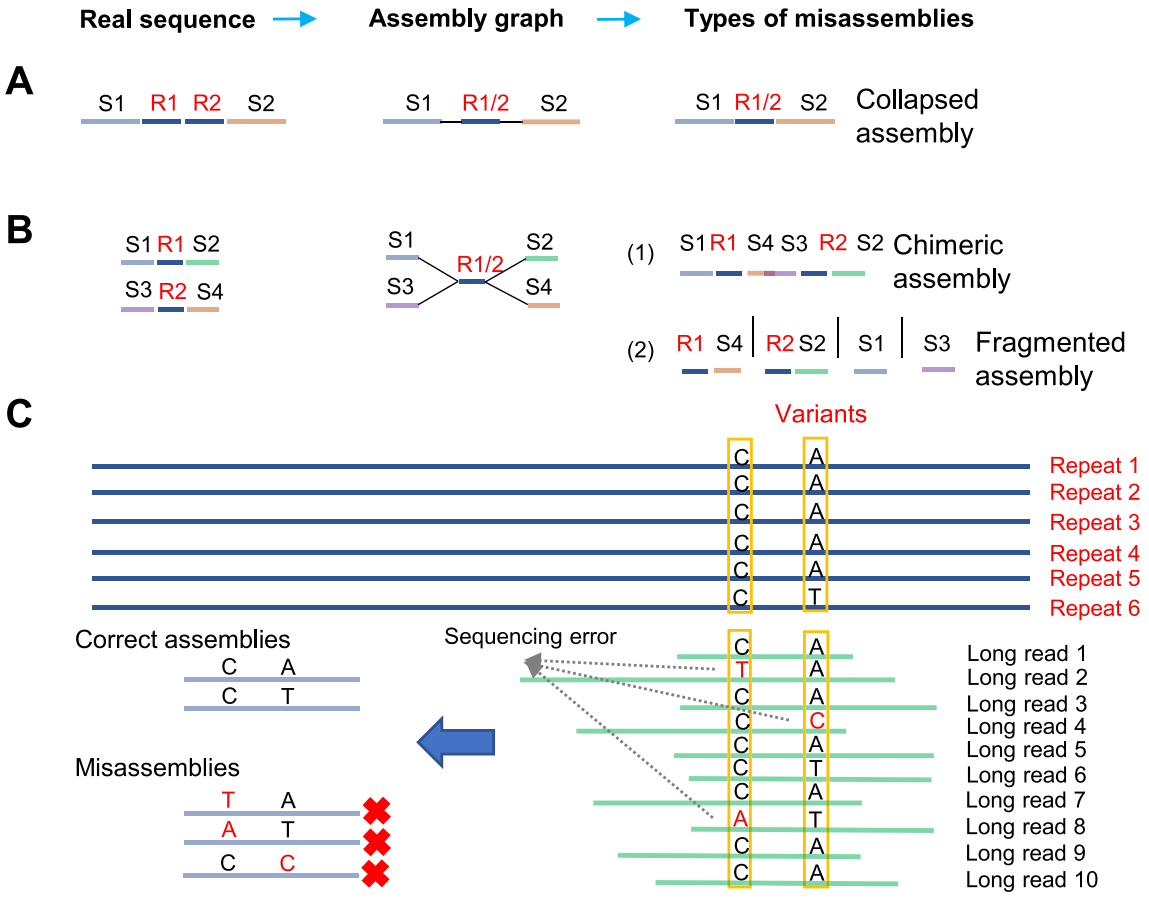

Figure 1 The assembly of highly repetitive sequences

A. A collapsed assembly error example in tandem repeats. A tandem repeat containing two copies (R1 and R2) separates unique sequences S1 and S2. **B.** Chimeric or fragmented assembly errors in long segmental repeats among different chromosomal regions. S1, S2, S3, and S4 indicate unique sequences, and R1 and R2 represent two identical long segmental repeats. **C.** The impact of sequencing errors on the assembly of highly similar repeats.

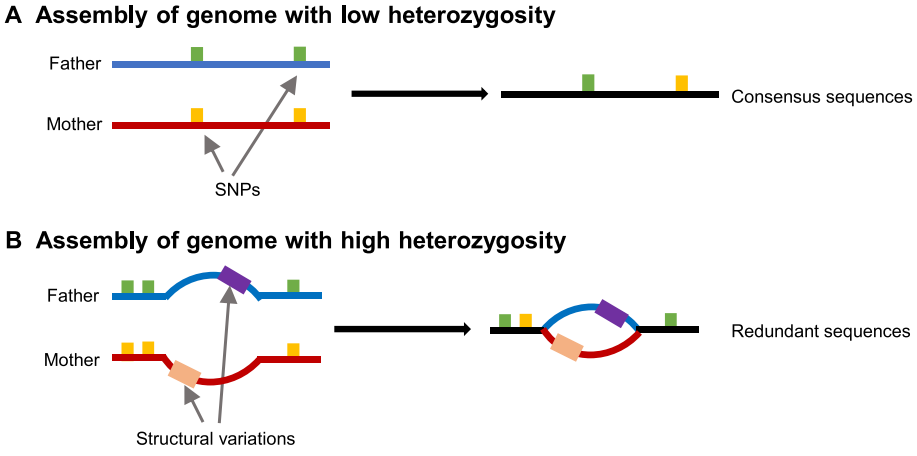

Figure 2 The impact of high heterozygosity on genome assembly

A. Consensus sequence assembly of a low-heterozygosity genome. Small-scale variations (such as SNPs) in different haplotype sequences can be aligned during assembly and then assembled into consensus sequences. **B.** Bubble structures of highly heterozygous genomes. Large-scale structural variations from different haplotype sequences affect the sequence alignment to form 'bubbles' representing redundant allelic sequences and fail to form consensus sequences. SNP, single-nucleotide polymorphism.

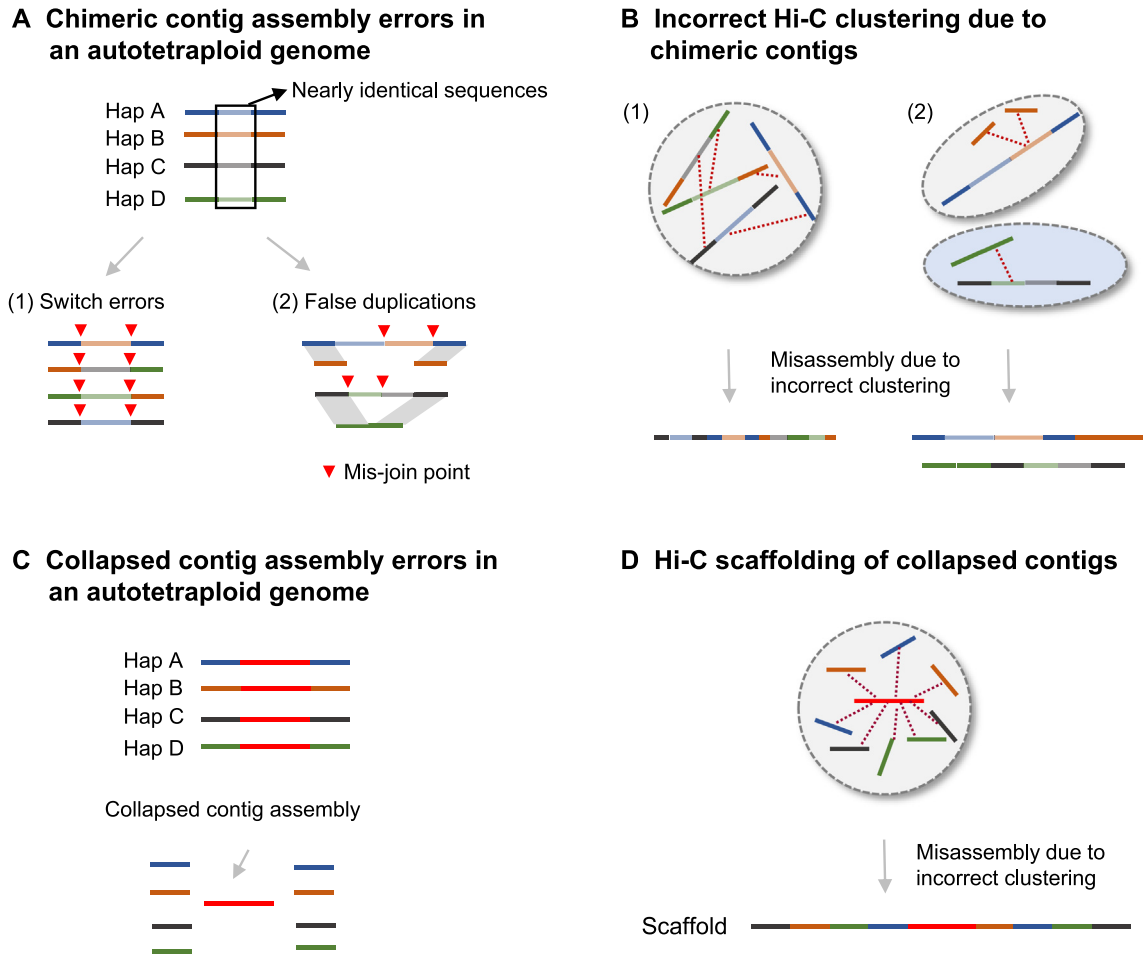

Figure 3 Challenges of polyploid genome assembly

A. Illustration of chimeric contig assembly errors in an autotetraploid genome, including switch errors and false duplications. **B.** Incorrect Hi-C clustering of chimeric contigs leads to multiple misassemblies. **C.** Illustration of collapsed contig assembly errors in an autotetraploid genome. **D.** The collapsed contig generates Hi-C links with all contigs belonging to four haplotypes, resulting in a superlong and erroneous scaffold. Hap, haplotype; Hi-C, high-throughput/resolution chromosome conformation capture.

lotypes cannot be accurately distinguished, leaving many collapsed contigs (Figure 3C). Furthermore, these collapsed contigs generate Hi-C links with phased contigs belonging to different haplotypes, resulting in superlong and erroneous scaffolds (Figure 3D).

Technical innovations in complex genome assembly

Evolution of sequencing platforms

Earlier plant genome assemblies were generated using Sanger sequencing and next-generation sequencing (NGS) technologies combined with the minimum tiling path, the overlap layout consensus, or *de Bruijn* graph approaches [35–38] for species such as *A. thaliana* [4], *Oryza sativa* [39,40], *Carica papaya* [41], *G. max* [42], and *Populus trichocarpa* [43]. Although widely used in many genome projects, these sequencing technologies have limited power to overcome assembly challenges in complex genomes due to the limited length of short reads (< 1000 bp) and inevitable GC bias. For instance, sequencing a 2.3-Gb maize genome relied on the construction of 16,848 bacterial artificial chromosome (BAC) libraries. This process was highly labor intensive and generated a fragmented assembly with more than 10% of genomic sequences missing [23].

The subsequent SMS technologies advanced by Pacific Biosciences (*i.e.*, PacBio) and Oxford Nanopore Technology (ONT) companies were able to generate long reads that could span the kilobase- or even megabase-level repetitive regions along chromosomes. The first plant genome (*Oropetium thomaeum*) assembled based on only PacBio long reads demonstrated the ability of genome assembly in terms of contiguity and completeness [37,44]. In addition, the maize B73 genome assembled by PacBio data had a 52-fold increase in contig continuity with reduced assembly errors in the centromeric region compared with the previous version [23] and greatly facilitated the annotation of functional genes and the evolutionary analysis of transposons [45]. Although SMS technologies have made revolutionary advancements in the assembly of complex plant genomes, they suffer from a higher sequencing error rate, ranging from 5% to 20% [46]. To address this issue, PacBio adopts the circular consensus sequencing model to generate long high-fidelity (HiFi) reads by reading multiple passes of a single template molecule [47]. This strategy achieves a read accuracy of more than 99.8% but at the cost of read length.

Experimental approaches for genome scaffolding

The reconstruction of chromosomes is an ultimate goal in genome assembly, aiming to determine the orientation and orders of contigs globally. This step, called scaffolding, is vital for many downstream analyses and applied tasks, including the identification of genome-wide genotype–phenotype associations, marker-assisted breeding, and chromosome evolution analysis. Genetic maps were widely applied to early genome projects for genome scaffolding, such as *Arabidopsis* [4] and rice genomes [39,40]. It successfully solved some complex genome assemblies, including that of the hexaploid bread wheat genome [48].

During the past decade, fruitful achievements have been made in experimental approaches for genome scaffolding, including BioNano optical maps using a light microscope-based technique to capture the physical locations of selected enzymes and a chromatin conformation capture technique (Hi-C) based on proximity ligation of chromatin. These two novel scaffolding approaches can quickly and accurately reconstruct the chromosomes for some complex plant genomes. Despite the limitation of sparse enzyme sites and the requirement of extraction of long DNA molecules [49], BioNano technology has shown its power in chromosomal-scale genome assembly in some plant genome projects, such as that of sorghum [50]. Hi-C technology can construct linkage information between contigs by detecting long-distance DNA interactions and has become routine for most genome projects. Applying Hi-C has resulted in the successful assembly of dozens or even hundreds of genomes, especially some complex polyploid genomes, such as those of sugarcane [17] and alfalfa [51].

Strategies for monoploid assembly in diploid genomes

Most diploid genome projects aim to generate 'consensus' sequences (*i.e.*, monoploid assembly) to represent a reference genome for a given species (Figure 2A). This goal can be easily achieved for some plant genomes with extremely low heterozygosity, such as those of *Arabidopsis* and rice. However, the assembly of heterozygous genomes requires additional processes to solve 'bubbles' representing redundant sequences in initial contigs, which contain a large proportion of allelic contigs that originate from homologous chromosomes. Thus, three main strategies have been designed to classify these redundant contigs: read depth (RD)-, whole genome alignment comparison (WGAC)-, and *K*-mer-based strategies.

The RD-based strategy identifies collapsed and redundant sequences by investigating the sequencing depth of mapped reads. The RDplot of the initial contigs in a highly heterozygous genome usually shows a bimodal distribution. Suppose collapsed or haplotype-fused contigs have a 1× RD. In that case, redundant contigs will only have approximately 0.5× RD, because redundant sequences will evenly distribute total reads due to the sequence similarity of two redundant contigs (**Figure 4**A). As a typical example, purge_haplotigs software [52] utilizes the RD-based strategy for identifying and removing these redundant sequences from a heterozygous assembly and eventually retains primary contigs to construct the monoploid genome (Figure 4A). The RD-based strategy has been successfully applied to several highly heterozygous genomes, namely, golden buckwheat (*Fagopyrum dibotrys*) [53], red clover (*Trifolium pratense* L.) [54], lilacs (*Syringa oblata* L.) [55], and carnation (*Dianthus caryophyllus*) [56]. However, the RD-based strategy consumes time and money due to the need for large-scale global alignment of genome sequences.

In contrast, several software programs, such as Pseudohaploid [19], purge_dups [57], and Redundans [58], implement the WGAC-based strategy to identify allelic contigs that have a high level of similarity and overlapping sequences. The long alignment chains detected by pairwise comparison between assembled contigs are considered redundant homologous regions. Only one copy of these homologous regions with a longer size was eventually retained as a representative haplotype (Figure 4B). However, the WGAC-based strategy is also time-consuming due to the global contig pairwise comparison.

To efficiently detect redundant contigs in complex genomes, a *K*-mer-based strategy named Khaper was proposed [46]. The basic concept of Khaper is to search for common low- and medium-frequency *K*-mers via pairwise comparison between contigs and to identify potential allelic contigs if they share a high proportion of low- and medium-frequency *K*-mers (Figure 4C). Because it does not rely on genome-wide sequence alignment, Khaper significantly saves central processing unit (CPU) time and solves the problems of time consumption and overutilization of computational resources in removing the redundancy process of large genomes with high heterozygosity [46].

Toward haplotype-resolved assembly

Most reference genomes for diploid and polyploid organisms stay at the 'monoploid' level, which represents 'mosaic' sequences that mix two or more homologous chromosomes. Unzipping accurate haplotypes in a polyploid genome is beset with difficulties. In the case of an *n*-ploid organism, $n-1$ haplotypes must be computed before the haplotype of interest can be inferred. For a pair of single-nucleotide polymorphisms (SNPs) in a polyploid, there are theoretically $n!$ connection possibilities. Recently, upgraded sequencing technologies and innovations in algorithms and strategies have provided the basis for assembling highly heterozygous diploid and polyploid genomes at the haplotype level rather than at the monoploid level.

To date, the built-in heuristic algorithms of multiple reference-based or *de novo* phasing tools have been systematically summarized [34,59,60]. However, many have been developed for the human genome and have not been applied well to plant genome assembly. Here, to explore effectual phased tools and strategies for assembling complex plant genomes, we have summarized the recently published haplotype-resolved assemblies in plant genomes. We have further divided these assembly approaches into two strategies: reference-based variant phasing and *de novo* assembly-based haplotype phasing (**Table 1**) [12,14–18,51,61–90].

Reference-based variant phasing

In reference-based variant phasing, a high-quality genome is required as the reference to distinguish different haplotypes

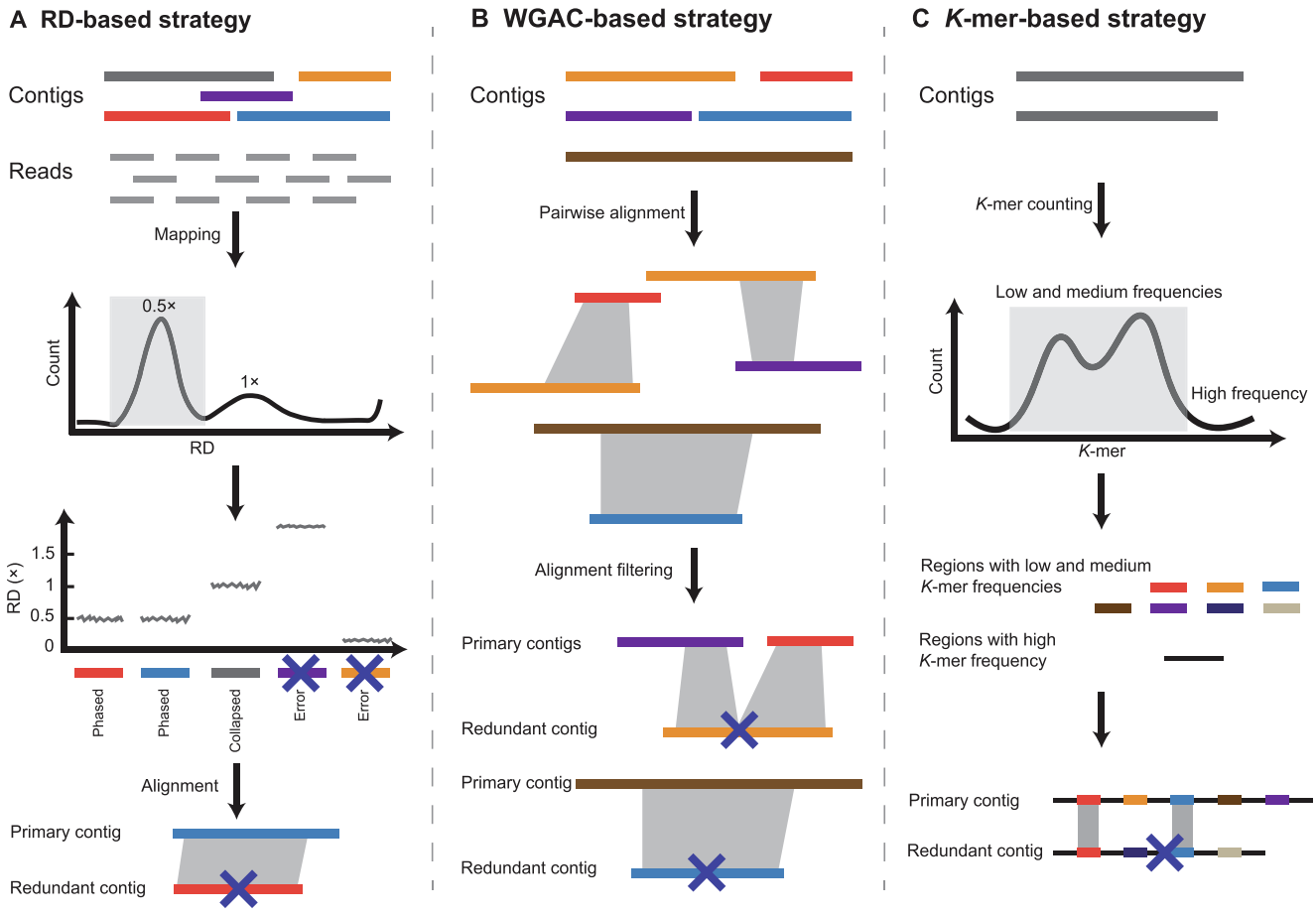

Figure 4 Three strategies for identifying redundant contigs

A. With the RD-based strategy, redundant or phased contigs are approximately one-half of the mapped RD of collapsed or haplotype-fused contigs due to the bisected RD and the extreme similarity between redundant contigs. Based on the RD of contigs, phased contigs and collapsed contigs can be accurately identified, and the redundant phased contigs will be filtered. **B.** With WGAC-based strategy, contigs with long-scale alignment are identified as redundant contigs, and only the longer one is selected to leave in the monoploid genome. **C.** In the *K*-mer-based strategy, more than 40× Illumina or BGI short reads are first used to build the *K*-mer data pool. Then, low- and medium-frequency *K*-mers are mapped to assembled contigs. Redundant contigs share a high proportion of low- and medium-frequency *K*-mers, and relatively long contigs are finally selected to leave in the monoploid genome. RD, read depth; WGAC, whole genome alignment comparison.

based on long-range linked allelic variants through alignments of sequencing reads against the reference genome using different phasing algorithms, including minimum error correction [91], weighted minimum letter flip [59], maximum fragment cut [92], and polyploid balanced optimal partition [93] as different frameworks [34,59,60]. More than 20 reference-based variant phasing tools have been developed (Table S1). However, only HapCUT2 [94] and Ranbow [95] were effectively used in haplotype-resolved genome assemblies of *Litchi chinensis* (diploid) [61] and *Ipomoea batatas* (hexaploid) [62] (Table 1). HapCUT2 utilizes advanced hybrid phasing programs and can handle chromosome-scale phasing using multiple types of sequencing data, including HiFi, 10X Genomics linked reads, and Hi-C reads [94]. In contrast, Ranbow is designed for haplotype reconstruction of the polyploid genome using a graph-based algorithm and can integrate all types of small variants in bi- and multiallelic sites to reconstruct haplotypes [95]. Although the reference-based variant phasing strategy shows its ability with less computational consumption, it also has drawbacks that limit its application to a wide range of complex genomes. The accuracy of this strategy is affected by a series of factors, including the quality of the reference genome, read length, sequencing depth, sequencing errors, and repeats. For instance, most plant genomes contain a large proportion of repetitive sequences, leading to ambiguous read mapping and inaccurate identification of variants. In addition, reference-based variant phasing tools mostly ignore large-scale allelic variants due to the inefficacy of identifying structural variations based on read mapping.

De novo assembly-based haplotype phasing

In contrast to reference-based variant phasing, which mainly retains single-nucleotide allelic variants, *de novo* assembly-based haplotype phasing tends to be more comprehensive. It can produce a noncollapsed haplotype-phased assembly, covering large types of allelic variants, such as indels and structural variants [34].

Table 1 Summary of available haplotype-resolved plant genome assemblies

Phasing strategy	Species	Karyotype	Sequencing platform	Tool or strategy	Ref.
Reference-based variant phasing					
	Litchi chinensis	$2n = 2x = 30$	Illumina + PacBio + 10X Genomics	HapCUT2	[61]
	Ipomoea batatas	$2n = 6x = 90$	Illumina	Ranbow	[62]
***De novo* assembly-based haplotype phasing**					
De novo phased contig tools + Hi-C scaffolding	*Bletilla striata*	$2n = 2x = 32$	HiFi + Hi-C	HiFiasm + LACHESIS	[63]
	Bupleurum chinense	$2n = 2x = 12$	Illumina + HiFi + Hi-C	HiFiasm + 3D-DNA	[64]
	Suaeda glauca	$2n = 2x = 18$	HiFi + Hi-C	HiFiasm + 3D-DNA	[65]
	Cynodon dactylon	$2n = 4x = 36$	Illumina + PacBio + Bionano + Hi-C	HiFiasm + 3D-DNA	[66]
	Populus tomentosa	$2n = 3x = 57$	Illumina + HiFi + Hi-C	HiFiasm + 3D-DNA	[67]
	Malus domestica	$2n = 2x = 34$	Illumina + 10X Genomics + HiFi	HiFiasm + DeNovoMAGIC	[68]
	Manihot esculenta	$2n = 2x = 36$	Illumina + HiFi + Hi-C	HiFiasm + ALLHiC	[69]
	Pogostemon cablin	$2n = 4x = 64$	Illumina + PacBio + Hi-C	Canu + 3D-DNA	[70]
	Manihot esculenta	$2n = 2x = 36$	Illumina + PacBio + Hi-C	FALCON + FALCON_unzip + FALCON-Phase	[71]
	Humulus lupulus	$2n = 2x = 20$	Illumina + PacBio + Hi-C	FALCON + FALCON-unzip	[72]
	Vanilla planifolia	$2n = 2x = 28$	Illumina + ONT	Miniasm + FALCON-Phase + LACHESIS	[73]
	Hydrangea macrophylla	$2n = 2x = 36$	Illumina + PacBio + Hi-C	FALCON + FALCON_unzip + FALCON-Phase	[74]
	Zingiber officinale	$2n = 2x = 22$	Illumina + PacBio + Hi-C	FALCON-Phase + 3D-DNA	[75]
	Saccharum spontaneum	$1n = 4x = 32$	Illumina + BACs + PacBio + Hi-C	Canu + ALLHiC	[17]
	Saccharum spontaneum	$2n = 4x = 40$	HiFi + Hi-C	Canu + ALLHiC + 3D-DNA	[18]
	Camellia sinensis	$2n = 2x = 30$	HiFi + Hi-C	HiFiasm + ALLHiC	[76]
	Solanum tuberosum	$2n = 4x = 48$	HiFi + Hi-C	HiFiasm + ALLHiC	[15]
	Camellia sinensis	$2n = 2x = 30$	Illumina + PacBio + Hi-C	Canu + ALLHiC	[77]
	Manihot esculenta	$2n = 2x = 36$	PacBio + Hi-C	Canu + Wtdbg + ALLHiC	[78]
	Dendrocalamus latiflorus	$2n = 6x = 70$	Illumina + PacBio + Hi-C	Falcon + ALLHiC	[79]
	Medicago sativa	$2n = 4x = 32$	Illumina + HiFi + Hi-C	Canu + ALLHiC	[51]
	Medicago sativa	$2n = 4x = 32$	PacBio + Bionano + Hi-C	Canu + MECAT + ALLHiC	[80]
	Medicago sativa	$2n = 4x = 32$	PacBio + Hi-C	Canu + ALLHiC	[81]
	Artemisia annua	$2n = 2x = 18$	Illumina + PacBio + Bionano + Hi-C	Canu + HiFiasm + FALCON + LACHESIS	[82]
Trio-binning-based *de novo* phasing	*Ananas comosus*	$2n = 2x = 50$	PacBio + ONT + Hi-C + Illumina	Trio-binning	[83]
	Cerasus × *kanzakura*	$2n = 2x = 16$	Illumina + PacBio	Trio-binning	[84]
	Cerasus × *yedoensis*	$2n = 2x = 16$	Illumina + PacBio	Trio-binning	[85]
Genetic map-based *de novo* phasing	*Pyrus bretschneideri*	$2n = 2x = 34$	BACs + SCS	Gamete binning	[86]
	Solanum tuberosum	$2n = 4x = 48$	HiFi + SCS	Gamete binning	[16]
	Camellia sinensis	$2n = 2x = 30$	SCS	Gamete binning	[87]
	Prunus armeniaca	$2n = 2x = 16$	PacBio + SCS	Gamete binning	[88]
	Solanum tuberosum	$2n = 2x = 24$	ONT + 10X Genomics + HiFi + Hi-C + Illumina for self-population	Population resequencing	[12]
	Solanum tuberosum	$2n = 4x = 48$	HiFi + Hi-C + Illumina for self-population	Population resequencing	[14]
	Vitis riparia	$2n = 2x = 38$	Illumina + PacBio + 10X Genomics + GBS data	Population resequencing	[89]
	Zoysia japonica	$2n = 4x = 20$	PacBio + GBS data	Population resequencing (PolyGembler)	[90]

Note: Hi-C, high-throughput/resolution chromosome conformation capture; PacBio, Pacific Biosciences; ONT, Oxford Nanopore Technologies; GBS, genotyping-by-sequencing; BAC, bacterial artificial chromosome; SCS, single-cell sequencing; HiFi, high fidelity.

Most of the phased plant genomes published to date were completed by relying on *de novo* phased contigs followed by Hi-C scaffolding (*i.e.*, *de novo* phased contig tools + Hi-C scaffolding). Briefly, allelic contigs are initially assembled and phased by allele-aware algorithms implemented in PacBio assemblers (*e.g.*, Hifiasm [96,97], Canu [98], and FALCON-Unzip [99]). The phased contigs are subsequently subjected to Hi-C scaffolding tools (such as LACHESIS [100], 3D-DNA [101], FALCON-Phase [102], and ALLHiC [103]), achieving haplotype construction at the chromosome level (Table 1). Hifiasm and Canu use haplotype-aware graphs with reads as nodes and read overlaps as edges to assemble all contigs from different haplotypes [97,98]. In heterozygous diploid genomes, Hifiasm can solve haplotype-aware graphs based on Hi-C reads that provide long-range links between contigs, in which step allelic contigs are fully separated into two haplotypes [96,97]. In contrast, Canu requires postprocessing to assign contigs to haplotypes with tools such as Purge_dups [57], FALCON-Phase [102], and ALLHiC [103] to split contigs into different haplotypes [98]. The widely used Hi-C scaffolding programs in haplotype-resolved genome assembly include 3D-DNA and ALLHiC. Benefiting from the fully separated allelic contigs in Hifiasm, 3D-DNA takes contigs from each haplotype as inputs and implements scaffolding algorithms in highly homozygous diploid genomes. However, it has limited power to work with the assembled allelic contigs that are not separated into haplotypes in the polyploid genomes. ALLHiC uses a novel pruning step to remove Hi-C links between phased contigs and collapsed regions as well as allelic Hi-C signals based on a customized allelic contig table [103]. Removing the interference of Hi-C links allows the phased contigs to be accurately partitioned according to the strength of the Hi-C links. However, ALLHiC depends heavily on the initial assembly quality, a phenomenon that is known as "garbage in, garbage out" [103].

Recently, trio-binning-based diploid phasing algorithms for trio sequencing data have been developed, including TrioCanu [104], Hifiasm + trio [97], and WHdenovo [105]. The long sequencing reads of an F_1 hybrid with a high level of heterozygosity are first partitioned into paternal and maternal read sets based on the unique parental Kmers [104]. The two read sets are assembled separately into two haploid genomes, with each representing a parental genome. Although trio-binning-based algorithms perform exceptionally well in terms of continuity and accuracy of phased contigs, they have limited application to complex plant genomes due to a lack of parental data. In addition, genomic regions that are heterozygous in both parents cannot be phased [34].

Genetic maps have been widely used in early genome projects for chromosome construction. Additionally, they demonstrate an ability to carry out haplotype phasing by resequencing hundreds of individuals in a derived population (*e.g.*, a selfing population). In a heterozygous diploid potato (RH), all contigs were assembled from high-quality long reads and 10X Genomics linked reads. Then, each contig was regarded as a molecular marker, and the copy number (0, 1, 2) of the contig in each progeny was inferred based on the distribution of each individual read number, corresponding to the genotype (aa, Aa, AA). The genotype matrix of all contigs in the selfing population allowed the contigs to be divided into 24 linkage groups corresponding to 12 chromosome pairs using traditional genetic mapping strategies. Finally, the long reads and 10X Genomics linked reads for each linkage group were retrieved and reassembled to generate an improved scaffold assembly [12]. Bao et al. recently introduced this approach into the tetraploid potato genome and assembled the haplotype-resolved genome of a tetraploid cultivated potato [14]. In addition, gamete binning, a method based on single-cell sequencing of hundreds of haploid gamete genomes, enables the separation of long sequencing reads (such as HiFi reads) into two haploid-specific read sets. After the independent assembly of reads for each haplotype, contigs were scaffolded to the chromosomal level using a genetic map derived from recombination patterns within the same gamete genomes [88]. Gamete binning has been efficiently used to infer genome-wide haplotypes in diploid pear, apricot tree, tea plant, and tetraploid potato [86–88,106]. However, the construction of the genetic map relies on extensive meiotic recombination, which often means genotyping hundreds of recombined genomes, leading to doubling of sequencing costs relative to other strategies. Moreover, the separation of gametes is severely limited by the level of the experimental technique and by specific seasons. In addition, developing derived populations is time-consuming and costly and may pose significant challenges if the individuals show long juvenility or sterility [17].

Implications of haplotype-resolved genome assembly

In the era of NGS, the compromise method is to use sequence-derived haploid materials or to tolerate chimeric heterozygous regions to construct a reference genome. Therefore, most reference genomes for diploid and polyploid organisms stay at the 'monoploid' level, which represents 'mosaic' sequences from more than two homologous chromosomes yet fails to capture allelic variants that are diploid or polyploid in nature and that may be associated with compound heterozygotes, dosage effects, homeolog silencing, heterosis, population genetics, and species evolution [34]. The functional and evolutionary study of polyploids would require a full dissection of the different allele sequences. Accurate haplotype-resolved genome assembly is essential for analyzing haplotypic structural variants and allele-specific expression for complex traits, such as heterosis and genomic imprinting. Additionally, a better understanding of haplotypic variations is key to designing advanced breeding strategies, especially for overcoming severe inbreeding depression or for improving crop yield. Recently, several research groups have generated haplotype-resolved genome assemblies for several important plant species, including sugarcane [17], banyan tree [11], tea plant [77], and potato [12–16]. These studies not only established references for the assembly of complex genomes but also provided new insights into genome evolution and biological questions concerning these horticulture or crop plants.

The application of the most advanced SMS sequencing and Hi-C technologies has successfully anchored an autotetraploid sugarcane genome onto 32 chromosomes [17]. Based on a high-quality phased genome, a syntenic analysis confirmed two rounds of whole-genome duplication events in the sugarcane species. The reduction in chromosome bases from 10 to 8 in *Saccharum spontaneum* compared with sorghum has resulted from two chromosome fissions and two fusions.

The Tieguanyin tea cultivar has been cultivated for approximately 300 years, and its genome accumulates a large number

of somatic mutations, including deleterious mutations, during the long-term asexual reproduction process. This process increases the genetic load and reduces adaptability. However, our knowledge of dealing with genetic load in the context of vegetatively propagated crops is limited. A fully phased assembly genome provides two sets of alleles that allow the precise study of the allele-specific expression pattern in multiple tissues. The authors found that asexually propagated individuals prefer ancestral or beneficial alleles rather than deleterious mutations to maintain plant growth and development as well as adaptability to the environment [77].

Potato (*S. tuberosum* L.) is one of the most important tuber crops, but its genetic improvement is slow due to tetrasomic inheritance and clonal propagation [16]. Asexual propagation through tubers is prone to the accumulation of deleterious mutations and has a higher cost than seed propagation [107]. However, the decline in selfing caused by deleterious mutations is an obstacle in potato seed propagation. To address this problem, Zhou et al. identified dispersed deleterious mutations and differentially expressed alleles based on a phase diploid genome, which provides operational targets for eradicating harmful alleles or for the accumulation of beneficial alleles through recombination [12]. This study has subsequently resulted in a breakthrough in potato breeding, leading to vigorous inbred-line-based F_1 hybrids with strong heterosis [107]. Based on the haplotype-resolved genome of a tetraploid potato cultivar (Otava), Sun et al. found that only 53.6% of the genes have all four haplotypes, and some of the four haplotypes for one gene are identical. Thus, there are only 3.2 haplotypes and 1.9 distinct alleles per gene, suggesting that potato yield and resistance can still be further improved by increasing the allelic diversity of the tetraploid genome because heterosis itself is based on nonadditive interactions of different alleles [16]. In addition, there were benefits from the phased assembly of tetraploid potato; deleterious mutations between homologous chromosomes were systematically identified; and the mutual shielding of deleterious mutations and functional gene complementation between parents were further reported [14]. In another phased assembly report on tetraploid potato, researchers analyzed the number and roles of deleterious and dysfunctional genes in the four haplotypes across six tetraploid cultivated potatoes. The autotetraploid potato is a clonally propagated species that undergoes limited meiosis. Its dysfunctional and harmful alleles are not eliminated, which significantly increases the difficulty of breeding. Using phased deleterious and dysfunctional gene information will help breeders create the best allele combination in the F_1 potato generation, thereby improving potato yield and quality [13].

Ficus hispida is a dioecious plant species, but its sex determination mechanism is a mystery owing to the lack of assembly of sex chromosomes, making it impossible to directly compare the sequence differences between the X and Y chromosomes. To investigate sex determination, our previous study on *Ficus* genomes proposed a novel pipeline, sex phase, to separate X and Y chromosomes by utilizing resequencing reads from individuals of known sex [11]. The comparative analysis of the phased sex chromosomes highlighted important structural variations between the X and Y chromosomes, including chromosome size (22.6 Mb in Y *vs.* 21.9 Mb in X) and a genomic inversion between 0.61 Mb and 1.57 Mb on the Y chromosome. Importantly, one protein-coding gene, *AGAMOUS 2* (*AG2*), in the sex-determining region, whose ortholog is essential for the development of the stamen and the carpel in *Arabidopsis* [108], is present in male individuals but absent in females. Furthermore, PCR amplification of three representative species of the dioecious subgenera confirmed that the *AG2* gene is likely a dominant sex-determining factor across the *Ficus* genus.

The era of telomere-to-telomere assembly

A more challenging task is to generate gapless or telomere-to-telomere (T2T) assemblies of complex plant genomes. Recently, the goal of near-T2T assemblies has been achieved in rice [27,109,110], *Arabidopsis* [111–113], banana [114], and watermelon [115]. These near-T2T assemblies not only provide the opportunity to update the knowledge of megabase-scale tandemly repeated satellite arrays and epigenetic organization in centromeres, but also indicate that the plant genome has entered the T2T era [27,111]. In sequencing techniques and assembly strategies, these near-T2T genomes all used recent assembly algorithms such as Hicanu and Hifiasm to complete primary contig assembly and fill gaps with high-precision HiFi reads. The remaining gaps were then filled twice with the superlong ONT reads or BAC reads, and the gap sequences were finally corrected and polished with HiFi reads. Alternatively, Rautiainen et al. developed a hybrid genome assembly pipeline (Verkko) for T2T assembly by integrating HiFi and ultralong ONT reads, showing its power to generate the T2T assembly for human genomes [116]. However, the accumulated experience in these model-like plants to complete the T2T assemblies of the complex plant genomes remains a long way to go before completely solving the high heterozygosity, high repeat sequence content, and high ploidy problems of complex plant genomes.

Discussion and future prospects

Deciphering complex plant genomes is of significance in understanding the basic biological mechanisms, including the discovery of key genes or structural variants related to resistance [117,118] and sex determination [11]. Additionally, benefitting from the development of phasing algorithms, comparative analysis between/among haplotypes in highly heterozygous or polyploid genomes identifies abundant allelic variations, which provide a genetic basis for studying fundamental biological questions about heterosis and allelic imbalance.

In this review, we describe some examples of complex plant genome projects and many tools being used to address the complexity of genome assembly. We recommend that $\sim 30\times$ HiFi and $\sim 100\times$ Hi-C sequencing are necessary for high-quality assemblies of complex genomes. Contigs with high continuity and accuracy can be assembled and phased by a widely used HiFi assembler, Hifiasm [96,97]. Chromosomal-scale assembly can be achieved using a series of Hi-C scaffolders, including 3D-DNA for diploid genomes and ALLHiC for polyploid genomes. In addition, a significantly improved gapless or T2T assembly requires additional ultralong ONT reads

based on several successful cases [27,109–113]. If parents or derived population materials exist, resequencing of these materials can obviously improve haplotype results. However, there is no all-powerful method applicable to all genomes and the optimal genome assembly sometimes needs testing of different pipelines.

Indeed, dozens of plant species that are economically important have ultracomplex genomes, leaving their genomic sequences under ongoing development. A typical example is modern hybrid sugarcane (*Saccharum* spp. hybrids), a widely cultivated crop of sugar and bioenergy production [119]. The nuclear genome of modern hybrid sugarcane is composed of subgenomes originally from *S. officinarum* (an octoploid species, with a basic chromosome number of 10, $2n = 80$) and *S. spontaneum* (varied ploidy levels between 5× and 16×, a basic chromosome number of 8, $2n = 40$–128). The hybrid sugarcane genome is much more complicated due to the uneven inheritance of genetic materials from its progenitors through interspecific crosses and one or more subsequent backcrosses [120,121]. Modern hybrid sugarcane has a basic chromosome number of 10. However, its complexity resides in the mixture of aneuploid and homo(eo)logous chromosomes, which results in 10 uneven homo(eo)logous chromosome groups of the modern hybrid sugarcane genome carrying a total number of chromosomes ranging from 100 to 130 [122,123]. It means that there are 8–14 homo(eo)logous copies for most genes in the hybrid sugarcane genome [123,124]. The state-of-the-art Hi-C scaffolding technology loses its power on ultracomplex genomes mostly owing to an extremely low level of uniquely mapped short reads. The recently proposed Pore-C method, which integrates single-molecule long-read sequencing and three-dimensional chromatin conformation capture technology, is able to detect multiway interactions among different genomic regions and distinguish highly similar genomic sequences [125]. The experimental innovation promises an effective approach to avoid multiple alignment in polyploid genomes, likely solving the ultracomplex sugarcane genome assembly.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Weilong Kong: Investigation, Methodology, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Yibin Wang:** Formal analysis, Visualization. **Shengcheng Zhang:** Formal analysis, Visualization. **Jiaxin Yu:** Formal analysis. **Xingtan Zhang:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision. All authors have read and approved the final manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 32222019) and the National Key R&D Program of China (Grant No. 2021YFF1000900). We would like to thank Profs. Jue Ruan and Tao Zhao, who provided many valuable ideas in manuscript preparation.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2023.04.004.

ORCID

ORCID 0000-0002-6066-0508 (Weilong Kong)
ORCID 0000-0002-0781-3966 (Yibin Wang)
ORCID 0000-0002-3949-4269 (Shengcheng Zhang)
ORCID 0000-0003-3374-8675 (Jiaxin Yu)
ORCID 0000-0002-5207-0882 (Xingtan Zhang)

References

[1] Meyers LA, Levin DA. On the abundance of polyploids in flowering plants. Evolution 2006;60:1198–206.
[2] Kyriakidou M, Tai HH, Anglin NL, Ellis D, Stromvik MV. Current strategies of polyploid plant genome sequence assembly. Front Plant Sci 2018;9:1660.
[3] Wang PP, Moore BM, Panchy NL, Meng FR, Lehti-Shiu MD, Shiu SH. Factors influencing gene family size variation among related species in a plant family, Solanaceae. Genome Biol Evol 2018;10:2596–613.
[4] Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, et al. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 2000;408:796–815.
[5] Sun YQ, Shang LG, Zhu QH, Fan LJ, Guo LB. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci 2022;27:391–401.
[6] Ming R, Bendahmane A, Renner SS. Sex chromosomes in land plants. Annu Rev Plant Biol 2011;62:485–514.
[7] Carey SB, Lovell JT, Jenkins J, Leebens-Mack J, Schmutz J, Wilson MA, et al. Representing sex chromosomes in genome assemblies. Cell Genom 2022;2:100132.
[8] Renner SS, Muller NA. Plant sex chromosomes defy evolutionary models of expanding recombination suppression and genetic degeneration. Nat Plants 2021;7:392–402.
[9] Wang JP, Na JK, Yu QY, Gschwend AR, Han J, Zeng FC, et al. Sequencing papaya X and Y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci U S A 2012;109:13710–5.
[10] Muller NA, Kersten B, Montalvao APL, Mahler N, Bernhardsson C, Brautigam K, et al. A single gene underlies the dynamic evolution of poplar sex determination. Nat Plants 2020;6:630–7.
[11] Zhang XT, Wang G, Zhang SC, Chen S, Wang YB, Wen P, et al. Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. Cell 2020;183:875–89.
[12] Zhou Q, Tang D, Huang W, Yang ZM, Zhang Y, Hamilton JP, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Genet 2020;52:1018–23.
[13] Hoopes G, Meng X, Hamilton JP, Achakkagari SR, de Alves Freitas Guesdes F, Bolger ME, et al. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. Mol Plant 2022;15:520–36.
[14] Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. Mol Plant 2022;15:1211–26.

[15] Wang F, Xia Z, Zou M, Zhao L, Jiang S, Zhou Y, et al. The autotetraploid potato genome provides insights into highly heterozygous species. Plant Biotechnol J 2022;20:1996–2005.
[16] Sun HQ, Jiao WB, Campoy JA, Krause K, Goel M, Folz-Donahue K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet 2022;54:342–8.
[17] Zhang JS, Zhang XT, Tang HB, Zhang Q, Hua XT, Ma XK, et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat Genet 2018;50:1565–73.
[18] Zhang Q, Qi YY, Pan HR, Tang HB, Wang G, Hua XT, et al. Genomic insights into the recent chromosome reduction of autopolyploid sugarcane *Saccharum spontaneum*. Nat Genet 2022;54:885–96.
[19] Chen LY, VanBuren R, Paris M, Zhou HY, Zhang XT, Wai CM, et al. The bracteatus pineapple genome and domestication of clonally propagated crops. Nat Genet 2019;51:1549–58.
[20] Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 2012;13:36–46.
[21] Du HL, Liang CZ. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nat Commun 2019;10:5360.
[22] Biscotti MA, Olmo E, Heslop-Harrison JS. Repetitive DNA in eukaryotic genomes. Chromosome Res 2015;23:415–23.
[23] Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science 2009;326:1112–5.
[24] Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 2018;361:eaar719.
[25] Jiao WB, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 2017;36:64–70.
[26] Sun SL, Zhou YS, Chen J, Shi JP, Zhao HM, Zhao HN, et al. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet 2018;50:1289–95.
[27] Song JM, Xie WZ, Wang S, Guo YX, Koo DH, Kudrna D, et al. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol Plant 2021;14:1757–67.
[28] Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 2007;8:241–59.
[29] Britten RJ. Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A 2010;107:19945–8.
[30] Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res 2009;19:243–54.
[31] Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature 2009;457:551–6.
[32] Soltis PS, Soltis DE. The role of hybridization in plant speciation. Annu Rev Plant Biol 2009;60:561–88.
[33] Spoelhof JP, Soltis PS, Soltis DE. Pure polyploidy: closing the gaps in autopolyploid research. J Syst Evol 2017;55:340–52.
[34] Zhang XT, Wu RX, Wang YB, Yu JX, Tang HB. Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J 2020;18:66–72.
[35] Michael TP, Jackson S. The first 50 plant genomes. Plant Genome 2013;6:1.
[36] Rizzi R, Beretta S, Patterson M, Pirola Y, Previtali M, Della Vedova G, et al. Overlap graphs and *de Bruijn* graphs: data structures for *de novo* genome assembly in the big data era. Quant Biol 2019;7:278–92.
[37] Michael TP, VanBuren R. Building near-complete plant genomes. Curr Opin Plant Biol 2020;54:26–33.
[38] Miklos GLG, Rubin GM. The role of the genome project in determining gene function: insights from model organisms. Cell 1996;86:521–9.
[39] Goff SA, Ricke D, Lan TH, Presting G, Wang RL, Dunn M, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 2002;296:92–100.
[40] Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science 2002;296:79–92.
[41] Ming R, Hou SB, Feng Y, Yu QY, Dionne-Laporte A, Saw JH, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 2008;452:991–6.
[42] Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature 2010;463:178–463.
[43] Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 2006;313:1596–604.
[44] VanBuren R, Bryant D, Edger PP, Tang HB, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature 2015;527:508–11.
[45] Jiao YP, Peluso P, Shi JH, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature 2017;546:524–7.
[46] Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 2018;15:461–8.
[47] Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol 2019;37:1155–62.
[48] Chapman JA, Mascher M, Buluc A, Barry K, Georganas E, Session A, et al. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol 2015;16:26.
[49] Yuan YX, Chung CYL, Chan TF. Advances in optical mapping for genomic research. Comput Struct Biotechnol J 2020;18:2051–62.
[50] Deschamps S, Zhang Y, Llaca V, Ye L, Sanyal A, King M, et al. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat Commun 2018;9:4844.
[51] Chen HT, Zeng Y, Yang YZ, Huang LL, Tang BL, Zhang H, et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun 2020;11:2494.
[52] Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 2018;19:460.
[53] He M, He Y, Zhang K, Lu X, Zhang X, Gao B, et al. Comparison of buckwheat genomes reveals the genetic basis of metabolomic divergence and ecotype differentiation. New Phytol 2022;235:1927–43.
[54] Yan ZF, Sang LJ, Ma Y, He Y, Sun J, Ma LC, et al. A *de novo* assembled high-quality chromosome-scale *Trifolium pratense* genome and fine-scale phylogenetic analysis. BMC Plant Biol 2022;22:332.
[55] Wang Y, Lu L, Li J, Li H, You Y, Zang S, et al. A chromosome-level genome of *Syringa oblata* provides new insights into chromosome formation in Oleaceae and evolutionary history of lilacs. Plant J 2022;111:836–48.
[56] Zhang XN, Lin SN, Peng D, Wu QS, Liao XZ, Xiang KL, et al. Integrated multi-omic data and analyses reveal the pathways

underlying key ornamental traits in carnation flowers. Plant Biotechnol J 2022;20:1182–96.
[57] Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 2020;36:2896–8.
[58] Pryszcz LP, Gabaldon T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 2016;44:e113.
[59] Rhee JK, Li H, Joung JG, Hwang KB, Zhang BT, Shin SY. Survey of computational haplotype determination methods for single individual. Genes Genomics 2016;38:1–12.
[60] Abou Saada O, Friedrich A, Schacherer J. Towards accurate, contiguous and complete alignment-based polyploid phasing algorithms. Genomics 2022;114:110369.
[61] Hu GB, Feng JT, Xiang X, Wang JB, Salojarvi J, Liu CM, et al. Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. Nat Genet 2022;54:73–83.
[62] Yang J, Moeinzadeh MH, Kuhl H, Helmuth J, Xiao P, Haas S, et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nat Plants 2017;3:696–703.
[63] Jiang L, Lin MF, Wang H, Song H, Zhang L, Huang QY, et al. Haplotype-resolved genome assembly of *Bletilla striata* (Thunb.) Reichb.f. to elucidate medicinal value. Plant J 2022;111:1340–53.
[64] Zhang QF, Li M, Chen XY, Liu GX, Zhang Z, Tan QQ, et al. Chromosome-level genome assembly of *Bupleurum chinense* DC provides insights into the saikosaponin biosynthesis. Front Genet 2022;13:878431.
[65] Yi LX, Sa R, Zhao SW, Zhang XM, Lu XD, Mu YN, et al. Chromosome-scale, haplotype-resolved genome assembly of *Suaeda Glauca*. Front Genet 2022;13:884081.
[66] Zhang B, Chen S, Liu JX, Yan YB, Chen JB, Li DD, et al. A high-quality haplotype-resolved genome of common bermudagrass (*Cynodon dactylon* L.) provides insights into polyploid genome stability and prostrate growth. Front Plant Sci 2022;13:890980.
[67] Tong SF, Wang YB, Chen NN, Wang DY, Liu B, Wang WW, et al. PtoNF-YC9-SRMT-PtoRD26 module regulates the high saline tolerance of a triploid poplar. Genome Biol 2022;23:148.
[68] Sun XP, Jiao C, Schwaninger H, Chao CT, Ma YM, Duan NB, et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat Genet 2020;52:1423–32.
[69] Qi W, Lim YW, Patrignani A, Schlapfer P, Bratus-Neuenschwander A, Gruter S, et al. The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. Gigascience 2022;11:giac028.
[70] Shen YT, Li WY, Zeng Y, Li ZP, Chen YQ, Zhang JX, et al. Chromosome-level and haplotype-resolved genome provides insight into the tetraploid hybrid origin of patchouli. Nat Commun 2022;13:3511.
[71] Mansfeld BN, Boyher A, Berry JC, Wilson M, Ou S, Polydore S, et al. Large structural variations in the haplotype-resolved African cassava genome. Plant J 2021;108:1830–48.
[72] Padgitt-Cobb LK, Kingan SB, Wells J, Elser J, Kronmiller B, Moore D, et al. A draft phased assembly of the diploid Cascade hop (*Humulus lupulus*) genome. Plant Genome 2021;14:e20072.
[73] Hasing T, Tang HB, Brym M, Khazi F, Huang TF, Chambers AH. A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. Nat Food 2020;1:811–9.
[74] Nashima K, Shirasawa K, Ghelfi A, Hirakawa H, Isobe S, Suyama T, et al. Genome sequence of *Hydrangea macrophylla* and its application in analysis of the double flower phenotype. DNA Res 2021;28:dsaa026.
[75] Li HL, Wu L, Dong ZM, Jiang YS, Jiang SJ, Xing HT, et al. Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. Hortic Res 2021;8:189.
[76] Wang PJ, Yu JX, Jin S, Chen S, Yue C, Wang WL, et al. Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. Hortic Res 2021;8:107.
[77] Zhang XT, Chen S, Shi LQ, Gong DP, Zhang SC, Zhao Q, et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. Nat Genet 2021;53:1250–9.
[78] Hu W, Ji CM, Shi HT, Liang Z, Ding ZH, Ye JQ, et al. Allele-defined genome reveals biallelic differentiation during cassava evolution. Mol Plant 2021;14:851–4.
[79] Zheng YS, Yang DM, Rong JD, Chen LG, Zhu Q, He TY, et al. Allele-aware chromosome-scale assembly of the allopolyploid genome of hexaploid Ma bamboo (*Dendrocalamus latiflorus* Munro). J Integr Plant Biol 2022;64:649–70.
[80] Shen C, Du HL, Chen Z, Lu HW, Zhu FG, Chen H, et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. Mol Plant 2020;13:1250–61.
[81] Long R, Zhang F, Zhang Z, Li M, Chen L, Wang X, et al. Genome assembly of alfalfa cultivar zhongmu-4 and identification of SNPs associated with agronomic traits. Genomics Proteomics Bioinformatics 2022;20:14–28.
[82] Liao B, Shen X, Xiang L, Guo S, Chen S, Meng Y, et al. Allele-aware chromosome-level genome assembly of *Artemisia annua* reveals the correlation between ADS expansion and artemisinin yield. Mol Plant 2022;15:1310–28.
[83] Nashima K, Shirasawa K, Isobe S, Urasaki N, Tarora K, Irei A, et al. Gene prediction for leaf margin phenotype and fruit flesh color in pineapple (*Ananas comosus*) using haplotype-resolved genome sequencing. Plant J 2022;110:720–34.
[84] Shirasawa K, Itai A, Isobe S. Genome sequencing and analysis of two early-flowering cherry (*Cerasus* × *kanzakura*) varieties, 'Kawazu-zakura' and 'Atami-zakura'. DNA Res 2021;28:dsab026.
[85] Shirasawa K, Esumi T, Hirakawa H, Tanaka H, Itai A, Ghelfi A, et al. Phased genome sequence of an interspecific hybrid flowering cherry, 'Somei-Yoshino' (*Cerasus* × *yedoensis*). DNA Res 2019;26:379–89.
[86] Shi DQ, Wu J, Tang HB, Yin H, Wang HT, Wang R, et al. Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. Genome Res 2019;29:1889–99.
[87] Zhang W, Luo C, Scossa F, Zhang Q, Usadel B, Fernie AR, et al. A phased genome based on single sperm sequencing reveals crossover pattern and complex relatedness in tea plants. Plant J 2021;105:197–208.
[88] Campoy JA, Sun HQ, Goel M, Jiao WB, Folz-Donahue K, Wang N, et al. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. Genome Biol 2020;21:306.
[89] Girollet N, Rubio B, Bert PF. *De novo* phased assembly of the *Vitis riparia* grape genome. Sci Data 2019;6:127.
[90] Zhou CX, Olukolu B, Gemenet DC, Wu S, Gruneberg W, Cao MD, et al. Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations. Nat Genet 2020;52:1256–64.
[91] Bonizzoni P, Dondi R, Klau GW, Pirola Y, Pisanti N, Zaccaria S. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. J Comput Biol 2016;23:718–36.
[92] Duitama J, Huebsch T, McEwen G, Suk EK, Hoehe MR. ReFHap: a reliable and fast algorithm for single individual haplotyping. Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology 2010:160–9.

[93] Xie MZ, Wu Q, Wang JX, Jiang T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics 2016;32:3735–44.
[94] Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res 2017;27:801–12.
[95] Moeinzadeh MH, Yang J, Muzychenko E, Gallone G, Heller D, Reinert K, et al. Ranbow: a fast and accurate method for polyploid haplotype reconstruction. PloS Comput Biol 2020;16:e1007843.
[96] Cheng HY, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol 2022;40:1332–5.
[97] Cheng HY, Concepcion GT, Feng XW, Zhang HW, Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods 2021;18:170–5.
[98] Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 2020;30:1291–305.
[99] Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016;13:1050–4.
[100] Burton JN, Adey A, Patwardhan RP, Qiu RL, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nature Biotechnol 2013;31:1119–25.
[101] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.
[102] Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, et al. Extended haplotype-phasing of long-read *de novo* genome assemblies using Hi-C. Nat Commun 2021;12:1935.
[103] Zhang XT, Zhang SC, Zhao Q, Ming R, Tang HB. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants 2019;5:833–45.
[104] Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. *De novo* assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol 2018;36:1174–82.
[105] Garg S, Aach J, Li H, Sebenius I, Durbin R, Church G. A haplotype-aware *de novo* assembly of related individuals using pedigree sequence graph. Bioinformatics 2020;36:2385–92.
[106] Guk JY, Jang MJ, Choi JW, Lee YM, Kim S. *De novo* phasing resolves haplotype sequences in complex plant genomes. Plant Biotechnol J 2022;20:1031–41.
[107] Zhang CZ, Yang ZM, Tang D, Zhu YH, Wang P, Li DW, et al. Genome design of hybrid potato. Cell 2021;184:3873–83.
[108] Ito T, Ng KH, Lim TS, Yu H, Meyerowitz EM. The homeotic protein AGAMOUS controls late stamen development by regulating a jasmonate biosynthetic gene in *Arabidopsis*. Plant Cell 2007;19:3516–29.
[109] Li K, Jiang WK, Hui YY, Kong MJ, Feng LY, Gao LZ, et al. Gapless *indica* rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol Plant 2021;14:1745–56.
[110] Zhang YL, Fu J, Wang K, Han X, Yan TZ, Su YN, et al. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol J 2022;20:1642–4.
[111] Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science 2021;374:6588.
[112] Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, et al. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. Genomics Proteomics Bioinformatics 2022;20:4–13.
[113] Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. A near-complete assembly of an *Arabidopsis thaliana* genome. Mol Plant 2022;15:1247–50.
[114] Belser C, Baurens FC, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biol 2021;4:1047.
[115] Deng Y, Liu S, Zhang Y, Tan J, Li X, Chu X, et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Mol Plant 2022:1268–84.
[116] Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol 2023;41:1474–82.
[117] Deng YW, Zhai KR, Xie Z, Yang DY, Zhu XD, Liu JZ, et al. Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. Science 2017;355:962–5.
[118] Wang HW, Sun SL, Ge WY, Zhao LF, Hou BQ, Wang K, et al. Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science 2020;368:844.
[119] Lam E, Shine J, Da Silva J, Lawton M, Bonos S, Calvino M, et al. Improving sugarcane for biofuel: engineering for an even better feedstock. Glob Change Biol Bioenergy 2009;1:251–5.
[120] D'Hont A, Glaszmann JC. Sugarcane genome analysis with molecular markers: a first decade of research. Proceedings of the XXIV Congress 2001;2:556–9.
[121] Zhang JS, Nagai C, Yu QY, Pan YB, Ayala-Silva T, Schnell RJ, et al. Genome size variation in three *Saccharum* species. Euphytica 2012;185:511–9.
[122] Grivet L, Dhont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC. RFLP mapping in cultivated sugarcane (*Saccharum* spp): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 1996;142:987–1000.
[123] Thirugnanasambandam PP, Hoang NV, Henry RJ. The challenge of analyzing the sugarcane genome. Front Plant Sci 2018;9:616.
[124] Souza GM, Berges H, Bocs S, Casu R, D'Hont A, Ferreira JE, et al. The sugarcane genome challenge: strategies for sequencing a highly complex genome. Trop Plant Biol 2011;4:145–56.
[125] Deshpande AS, Ulahannan N, Pendleton M, Dai XG, Ly L, Behr JM, et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat Biotechnol 2022;40:1488–99.