Tropical
Plants

REVIEW

Open Access

https://doi.org/10.48130/TP-2022-0007

Tropical Plants **2022**, 1:7

De novo assembly of plant complete genomes

Yuhan Zhou[1,2], Ji Zhang[1,3], Xianghui Xiong[1,3], Zong-Ming Cheng[2], and Fei Chen[1,3*]

[1] *Hainan Yazhou Bay Seed Laboratory & Sanya Nanfan Research Institute from Hainan University, Sanya 572025, China*
[2] *College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China*
[3] *College of Tropical Crops, Hainan University, Haikou 570228, China*
* Corresponding author, E-mail: feichen@hainanu.edu.cn

Abstract

Plant genomes encode the mysteries of how plants cope with complex environments over long evolutionary histories. Over the past 20 years, rapidly developing technologies have allowed the decoding of hundreds of plant draft or reference genomes. The diversity, polyploidy and heterozygosity of plants make it technically challenging and time-consuming to generate high-quality plant genome assemblies. Recently invented ultra-long read sequencing technologies have achieved a milestone where several plant genomes have been gapless and assembled into telomere to telomere. Telomere-to-telomere (T2T) genome refers to a high-quality complete genome with high genomic accuracy, high continuity, and high integrity. With the release of the completed human genome and *Arabidopsis thaliana* genome, the era of complete T2T species genome has arrived. In this review, we summarize the history leading up to the gap free plant genomes based on emerging ultra-long read sequencing technologies. We discuss to close gaps relying on targeted genome sequencing and assembling technologies. However, there are still quite a lot of challenges in super large, polyploidy, and unstable genomes. Nevertheless, these complete genomes have already provided unprecedented information, which will certainly deepen our understanding of plant genomes and the exploration of more functional sequences. By taking advantage of the complete genomes, a series of important genes could be annotated, which will help achieve the goal of genome design in crop species.

Citation: Zhou Y, Zhang J, Xiong X, Cheng Z, Chen F. 2022. *De novo* assembly of plant complete genomes. *Tropical Plants* 1:7 https://doi.org/10.48130/TP-2022-0007

Introduction

In recent years, more and more high-quality genomes of plants have been deciphered. As of August 1, 2022, we have collected a total of 300 plant genomes assembled at the chromosome level with contig N50 greater than 1 Mb from PLAZA (https://plabipd.de/plant_genomes_pa.ep), and classified them according to whether the sequencing technology are ultra-long reads (> 50 kb) (Fig. 1). Contig N50 increased from 99.5 ± 48.1 kb in 2010 to 3,395.2 ± 735.4 kb in 2020[1]. There are seven genomes assembled from ultra-long reads, including *Arabidopsis thaliana*[2], *Lolium perenne*[3], *Papaver rhoeas, Papaver somniferum, Papaver setigerum*[4], *Citrullus lanatus*[5] and *Oryza sativa*[6]. From Fig. 1, we can find that the genome size of most sequenced species is between 100 Mb and 10 Gb. The contig N50 of maize is the highest, about 162 Mb[7], and the *Ginkgo biloba*[8] genome is the largest, up to 9.87 GB. In the long reads sequencing, there are also many plant genomes with high sequencing and assembly levels, such as the high-quality reference genome HFTH1 of apple[9]. It has certain reference significance for identifying structural variation, integrating phenotypic and genotypic association, analyzing the pattern and speed of genome evolution and clarifying the genetic structure of important traits.

Ultra-long DNA library preparation

Nowadays, based on different shearing methods and Nanopore sequencing kits, there are two ways to construct ultra-long DNA libraries: one is based on mechanical shearing, and the other is based on transposase. Mechanical fragmentation refers to the physical method for breaking DNA molecules into varying sizes which is considered the gold standard, including ultrasonic, spraying and hydrodynamic shearing methods. The N50 produced by the former is between 50 and 70 kb, and the construction of the library takes about 8 h. The N50 produced by the latter is between 90 and 100 kb, and it only takes 90 min to build the library[10]. However, the scheme of inputting the same DNA mechanical shearing can make it possible to obtain more productive fragments.

Oxford Nanopore and Circulomicsde Ultra-long DNA Sequencing Kit have supported the reading of sequences up to 4.2 Mb to maximize the number of ultra-long fragments. The kit is based on transposase chemistry: the transposase simultaneously cleaves template molecules and attaches tags to the cleaved ends. Its consumption of fuel during a sequencing run is reduced significantly. Combined with the Nanobind Kit from Circulomics, the Oxford Nanopore Ultra-long DNA Sequencing Kit can maximize the number of ultra-long read lengths, and has supported the continuous sequencing of single DNA fragments up to 3+ Mb (user data) and up to 4+ Mb (internal data)[11]. At first, the transposase of the Ultra-long DNA Sequencing Kit will simultaneously cleave the template molecule and attach the molecular marker to the cleaved end, then add the rapid sequencing linker to the labeled end, and finally elute the DNA library overnight with a Circulomics Nanobind disk (5 mm) to remove the free linker and short DNA fragments. With this method, it is possible for users to obtain

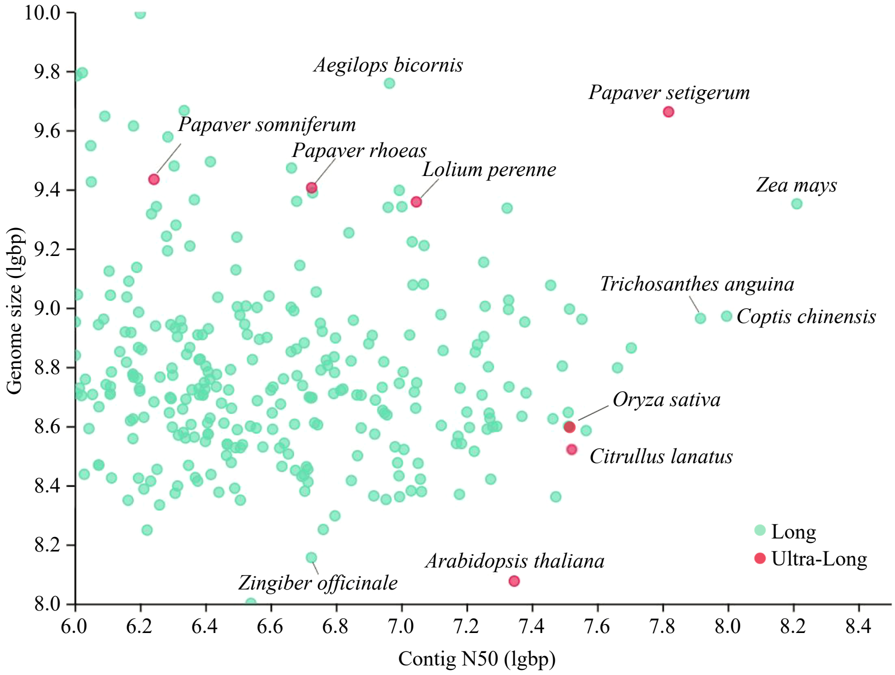

Fig. 1 Collection of the high-quality genome of 300 plants from PLAZA (https://plabipd.de/plant_genomes_pa.ep). Ultra-long sequencing dots are shown in red, and the rest are shown in green.

more than 100 read-length sequences larger than 1 Mb by running a PromethION sequencing chip. For example, the sequencing reading length N50 of *Lolium perenne* by this method is as long as 62 kb[3].

Ultra-long DNA sequencing technologies

Next-generation sequencing including Roche 454 and Illumina is a technology of sequencing while synthesizing, but its reading length is less than 200 bp, and time consuming. ONT (Oxford Nanopore Technologies) single molecule sequencing and Pacbio (Pacific Biosciences) HiFi sequencing are two current mainstream technologies. First of all, it is necessary to complete the preliminary assembly of the genome: survey, PacBio HiFi and ONT ultra-long sequencing of the genome to be tested are completed by using DNBSEQ short-length sequencing technology. And ONT PacBio single molecule real-time sequencing (SMRT) makes the single molecule read length exceed 10 kb, which is beyond the length of most simple sequence repeats.

Both of the two sequencing technologies have their own merits. With the comparison of HiFi, ONT ultra-long reads delivered higher contiguity. However, the ultra-long fragment obtained by ONT still has a relatively high base error rate before error correction. PacBio HiFi sequencing is a sequencing technology based on circular consensus sequencing (CCS). Its accuracy is as high as 99.8%, and the average length of generated HiFi reads is as long as 13.5 kb[12]. The quality of HiFi data is relatively stable in regions with different GC content and the repeatability is better. In this sequencing mode, it still has the same or even longer enzyme reading length (over 100 kb) as CLR (Continuous long reads) sequencing mode, but the inserted fragment is only 10−20 kb[13], which is far less than the reading length of the enzyme. Therefore, when sequencing, the enzyme will perform rolling circle sequencing around the DNA template, that is, the insert will be sequenced many times. In this way, random sequencing errors caused by single sequencing can be self-corrected by the algorithm, and finally high-accuracy HiFi Reads can be obtained. Because the amount of data used for assembly is small, and there is no need for data self-correction, the required computing resources in the assembly process are less than those in the traditional CLR mode, and the assembly cost is saved.

Combining ONT with HiFi can realize ultra-long sequencing. Finally, by combining Hi-C technology to obtain the relative position information of genes on chromosomes, the genome chromosome level assembly is completed, and the complex region is manually adjusted, and the T2T reference genome sequence is obtained (Fig. 2). PacBio HiFi sequencing depth is about 60×, and ONT ultra-long sequencing depth is about 30−200×.

Tools for genome assembly of ultra-long reads

The huge amounts of data of the second-generation sequencing increases the computational complexity of the assembly. And it is difficult to distinguish repetitive sequences of the genome which will produce incomplete assembly. Then, with the development of the third-generation sequencing technology, researchers combined the second and third-generation sequencing technology, plus optical mapping, genetic mapping or Hi-C technology to assemble the genome, which greatly improved the quality of assembly. Recent advances in optical mapping have allowed the genome comparison and identification of large-scale structural variations which also enables construction of improved genome assemblies with greater contiguity[14]. There are a large number of repetitive sequencing in the genome, which interferes with assembling, so genetic maps are required. These maps have their own advantages and disadvantages, and they need to be integrated and corrected with each other.

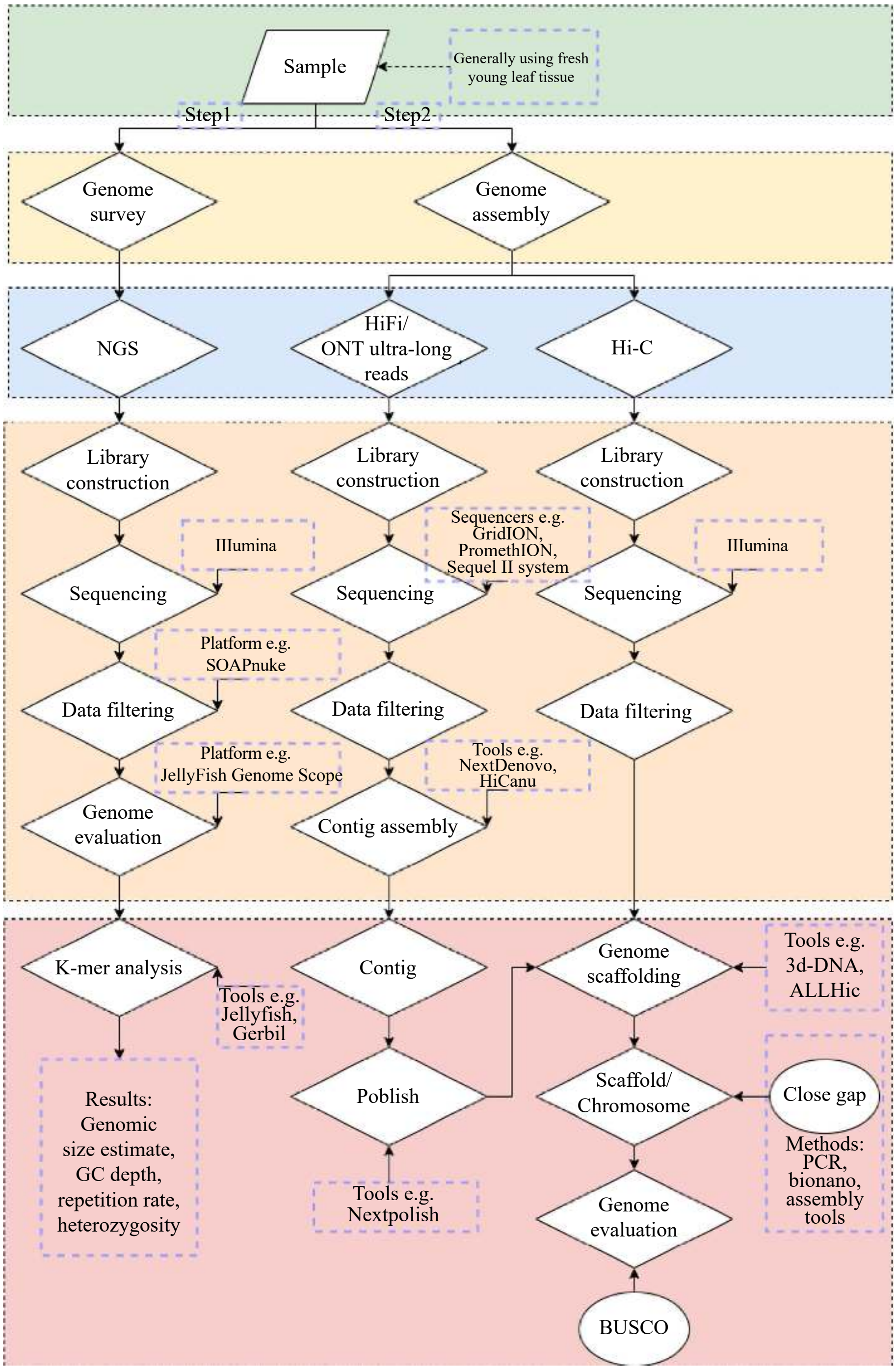

Fig. 2 Process of *de novo* assembly of the complete plant genomes. Sequencing platform, software, precautions, etc are noted in the figure alongside the steps.

To achieve the complete genome, firstly, researchers use DNBSEQ short-read sequencing technology to complete Survey. Then, using ONT and HiFi makes the assembling and polishing process simpler and the results more accurate. Finally, obtaining the relative position information of genes on chromosomes by combining Hi-C technology could accomplish the chromosomal level of the genome. Subsequently, haplotype genomes were constructed by combining Hi-C data and short-read sequencing data from the parents. Tools for assembling long reads are constantly emerging and updated, and the assembly efficiency and accuracy are getting higher and higher, such as Falcon[15−17], miniasm[18], Flye[19], Hinge[20], CANU[21], wtdbg[22], Shasta[23] and Wengan[24], etc (Table 1). But with the emergence of HiFi reads, HiCanu[25] and hifiasm[26] have become the most important assembly tools chosen by researchers. Most assembly results are based on multiple software, and it is necessary to try different sorts of assembly software constantly, so that the reliability of the results obtained is often the highest, and few assemblies only use a single specific assembly software for assembly. In the assembly of Hi-C reads, the LACHESIS[27] tool has not been updated, and 3D-DNA[28] and ALLHiC[29,30] gradually take its place. ALLHiC uses signal density to remove the link between alleles, which makes it easier for homologous chromosomes to be separated, so it can be used to solve the problem of plant polyploid assembly.

Some scientists are interested in reference-guided genome assembly, that is, how to mount newly assembled genomes to the homology of genome chromosomes of related or the same species. RaGOO[31] is a fast and reliable reference-guided scaffolding method.

After obtaining the genome sequence, it is of great importance to assess assembly quality. There are always some conserved sequences among similar species, and BUSCO (Benchmarking Universal Single-copy Orthologs) uses these conserved sequences to compare with the assembly results[32]. To identify whether the assembly results contains these sequences, so the integrity of the assembly can be obtained.

Targeted genome sequencing and PCR sequencing could close gaps

In view of the availability of whole genome assembly from scratch, especially those from long-reads sequencing data, gap closure sequences can be determined. There will be more gaps in the middle of the scaffold. In order to make the assembled sequence more complete, it is necessary to connect contigs again by using the pairing relationship between the sequenced double-ended data, and fill the holes between contigs by using the covering relationship between the sequenced data and the

Table 1. Tools for assembling long reads and Hi-C reads.

Tools	Features	URL
PacBio and ONT Assemblage relevant		
NextDenovo[33] (V2.5.0)	String graph-based de novo assembler for long reads which uses a 'correct-then-assemble' strategy, but requires significantly less computing resources and storage.	https://github.com/Nextomics/NextDenovo
Hifiasm[26] (V0.16.1)	Fast haplotype-resolved *de novo* assembler for PacBio HiFi reads.	https://github.com/chhylp123/hifiasm
SMARTdenovo[34]	A *de novo* assembler for PacBio and Oxford Nanopore (ONT) data. It produces an assembly from all-*vs*-all raw read alignments without an error correction stage.	https://github.com/ruanjue/smartdenovo
NGMLR[35] (V0.2.7)	Long-read mapper designed to align PacBio or Oxford Nanopore (standard and ultra-long) to a reference genome with a focus on reads that span structural variations.	https://github.com/philres/ngmlr
CentroFlye[36] (V0.8.3)	Algorithm for centromere assembly using long error-prone reads.	https://github.com/seryrzu/centroFlye_paper_scripts
Canu[21] (V2.2)	Fork of the Celera Assembler, designed for high-noise single-molecule sequencing (such as the PacBio RS II/Sequel or Oxford Nanopore MinION).	https://github.com/marbl/canu
Wtdbg2[22] (V2.5)	*De novo* sequence assembler for long noisy reads produced by PacBio or ONT which assembles raw reads without error correction and then builds the consensus from intermediate assembly output.	https://github.com/ruanjue/wtdbg2
HiCanu[25]	Modification of the Canu assembler designed to leverage the full potential of HiFi reads *via* homopolymer compression, overlap-based error correction, and aggressive false overlap filtering.	https://github.com/marbl/canu
HINGE[20]	Long read assembler based on OLC (Overlap-Layout-Consensus).	https://github.com/HingeAssembler/HINGE
Peregrine	Fast genome assembler for accurate long reads (length > 10 kb, accuracy > 99%).	https://github.com/cschin/Peregrine
Flye[19] (V2.9)	*De novo* assembler for single-molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies designed for a wide range of datasets.	https://github.com/fenderglass/Flye
Shasta[23] (V0.9)	The goal of the Shasta long read assembler is to rapidly produce accurate assembled sequence using DNA reads generated by Oxford Nanopore flow cells as input.	https://github.com/chanzuckerberg/shasta
NECAT[37] (V0.01)	Error correction and *de novo* assembly tool for Nanopore long noisy reads.	https://github.com/xiaochuanle/NECAT
Wengan[24] (V0.2)	Wengan avoids entirely the all-*vs*-all read comparison. The key idea behind Wengan is that long-read alignments can be inferred by building paths on a sequence graph.	https://github.com/adigenova/wengan
NextPolish[38] (V1.4)	NextPolish is used to fix base errors (SNV/Indel) in the genome generated by noisy long reads, it can be used with short read data only or long read data only or a combination of both.	https://github.com/Nextomics/NextPolish
Miniasm[18] (V0.3)	Very fast OLC-based *de novo* assembler for noisy long reads.	https://github.com/lh3/miniasm
Falcon[15−17] (V0.3)	Experimental PacBio diploid assembler	https://github.com/PacificBiosciences/FALCON
Raven[39] (V1.7)	*De novo* genome assembler for long uncorrected reads.	https://github.com/lbcb-sci/raven
HERA[40]	Local assembly tool using assembled contigs and self-corrected long reads as input which can generate ultra-long, even chromosome-scale, contigs.	https://github.com/liangclab/HERA
Hi-C scaffolding relevant		
HiC-Pro[41] (V3.1)	HiC-Pro was designed to process Hi-C data, from raw fastq files (paired-end Illumina data) to normalized contact maps.	https://github.com/nservant/HiC-Pro
SALSA[42] (V2.3)	A tool to scaffold long read assemblies with Hi-C.	https://github.com/marbl/SALSA
3D-DNA[28] (V201008)	3D *de novo* assembly (3D DNA) pipeline.	https://github.com/aidenlab/3d-dna
ALLHiC[29, 30]	Phasing and scaffolding polyploid genomes based on Hi-C data.	https://github.com/tangerzhang/ALLHiC
LACHESIS[27]	First tool to use Hi-C data to assist genome assembly.	https://github.com/shendurelab/LACHESIS

assembled contigs, so as to extend the contigs. The length of contigs after filling holes is generally increased by 2−7 times compared with that before filling holes. GapFiller is the software for filling holes[43], TGS-GapCloser[44], GAPPadder[45], PBJelly[46] and so on. Similarly, it is necessary to try different software and keep trying to choose the best solution.

Among them, Bionano can make complete single DNA molecules arranged in parallel in nanochannels through its unique chip technology, and take photos and images, so that the genome structure can be fully displayed[47]. Therefore, it can assist genome assembly. Five contigs were obtained from the human MHC region through NGS. Through Bionano, their positions in the genome and the size of gaps can be accurately determined. By comparing the genome map with sequencing fragments, it is easy to determine the positions of gaps and two contigs separated by 36.4 kb on the chromosome[48]. It can also read through repetitive sequence information. There are many repetitive fragments in the human genome. The copy number of repetitive fragments can be accurately determined through the Bionano system, which can read through long-chain single-molecule DNA fragments, so that this result can be clearly presented.

Researchers also introduced BAC (bacterial artificial chromosome)-anchor strategy to fill the remaining gaps in ONT-HiC assemblies which used ONT-generated ultra-long reads[49]. In short, for each gap, BAC sequences used to replace ONT contigs with HiFi contigs because HiFi contigs enjoy better continuity. All the BACs used share more than 99.9% sequence identity with their target contigs.

Complete genomes provide complete information

The genome assembled by the second-generation sequencing technology can only be regarded as a draft genome. With the continuous development of sequencing technology, the third-generation sequencing technology takes into account the advantages of the first-generation sequencing technology and the second-generation sequencing technology in terms of length and high-throughput, and can obtain a longer sequence, thus obtaining a reference-level genome. Therefore, more genetic information can be obtained through the third-generation sequencing technology[50].

However, due to the highly repetitive sequences such as telomeres and centromeres in the genome, almost all the genomes obtained today have a relatively high number of gaps, which are usually expressed by N or n. There are two reasons for the gap. One is that the gap is generated because of the restriction of the reading length. For example, if sequencing only has a reading length of 150 bp at both ends, and one fragment has 350 bp, the remaining 50 bp will not be known, so the longer the reading length, the smaller the gap will be. Second, assembly technology restricts the generation of gaps, such as comparing the sequenced reads with the contigs, and assembling contigs into scaffolding by using the connection relationship between the reads and the size information of inserted fragments (Mate-Pair), in which the undetermined region in scaffolding sequence. At present, many complete chromosomes have been sequenced and assembled in human beings[51], rice[6], *Arabidopsis thaliana*[2], banana (*Musa acuminata*)[52]. T2T genome refers to the high-precision, high-continuity, high-integrity genome assembly from telomere to telomere. It is realized by combining a variety of sequencing technologies, which is helpful to clarify the complex structure of highly repetitive regions in the genome, such as the detailed study of the variation characteristics and evolution patterns of centromeres and telomeres. T2T Alliance researchers have assembled and published the first completed picture of the human X chromosome[53]. Autosomal completion diagram[54] is a complete map of the human genome. Therefore, the satellite array in the centromere region, telomeres, large genome duplication and important genetic information in the rRNA region have been uncovered, among which there are many genes related to human aging and diseases. In human X chromosome sequencing, more than half of the reads obtained are over 70 kb, and the longest one is 1.04 Mb. The centromere contains a highly repetitive DNA region, which spans 3.1 million base pairs. In 2021, on the occasion of the 20th anniversary of the release of the draft human genome sequence, T2T Alliance released the latest complete human genome sequence CHM13 v1.1, which not only contains all the unresolved data, but also corrects the original assembly errors. In this completed picture of the human genome, researchers newly added or corrected 238 Mb of sequences, of which 182 Mb is a brand new sequence, and annotated 2,226 new genes, which is the most complete human genome ever. In the future, scientists will perform pan-genome sequencing on individuals of different races, so as to understand the genetic diversity of different races and individuals, and provide greater help for the future goal of precision medicine.

Due to the high similarity of homologous chromosomes in diploid species, genome assembly usually does not distinguish homologous chromosomes, and only assembles genomes with mixed parental genetic information. However, this assembly method will lead to inaccurate genetic annotation, which is not conducive to some biological studies that distinguish the genetic information of parents. Therefore, obtaining two haplotype genomes from parents provides important reference information for further study of allele mutation and understanding of species genetic relationship and evolutionary history. Furthermore, the sex chromosomes of a species usually carry important sex-determining genetic information, determine the development of reproductive organs, and show many completely different genomic characteristics and evolutionary patterns from them. However, the highly repetitive sequence and heterochromatin of sex chromosomes make them difficult to assemble. By sequencing and assembling the complete sex chromosomes, we can deeply analyze the specific differences of different sex individuals in species.

For plants, as early as 1999, the American Genome Research Institute (TIGR) assembled chromosome 2 of *Arabidopsis thaliana* without a gap, and the chromosome length obtained was 19.6 Mb[55]. It includes centromere and nucleolar organizer regions, but the function of nearly half of the genes on this chromosome is unknown. Presently, three high-quality assemblies of *A. thaliana*, Col-CEN[56], Col-XJTU[49] and Col-PEK[2], were deciphered and their quality was progressively improved. The rice[6] genome is the first complete gap-free plant genome published so far, and it fills the gap of 223 (ZS97RS1) and 167 (MH63RS1) between the two genomes.

In the past 100 years or so, 60% of the plants on the earth have been wiped out[57]. There are many wild varieties with excellent genes and traits, which is a great loss for those engaged in agricultural production and scientific research. The goal of Vertebrate Genome Project (VGP)[58], is to collect, sequence and assemble at least one high-quality, error-free, nearly gap-free, haplotype staged and annotated reference genome of all existing vertebrates, and to use these genomes to solve basic problems in biology, diseases and ecological protection. The minimum expected measurement values are contig N50 > 1 Mb, scaffold N50 > 10 Mb[59], and 90% of the assembly is located at the chromosome level. Tropical plants include 2/3 of higher plant species which have extremely rich genetic diversities. To protect endangered tropical wild plants, we could launch a tropical plant genome project and we hope that researchers can make use of the latest genome sequencing and assembly technology to assemble as complete and gapless as possible, which is more conducive to us to identify and classify a large amount of genetic information contained in excellent species.

High-definition genome provides complete gene sequences and complete repetitive sequences, which can help us understand the composition of centrioles and telomeres, promote the development of comparative genomics and evolutionary biology, and better modify the genome[60], providing genome data for genetic breeding and domestication, which in turn further promotes the development of the three major omics.

Challenges and prospects

Nowadays, for plants, the complete genome can only be assembled at the level of a single chromosome or simple species. Sequencing and assembling of the large genome (5 GB ≤ genome size < 10 GB) or very large genome (Genome Size ≥ 10 Gb) is still a worldwide problem. For example, the genome size of *Fritillaria pallida* exceeds 40 Gb[61]. Therefore, it is even more difficult to obtain its complete genome. The amount of data required for the assembly of very large genomes often reaches the Tb level. To obtain sequencing data quickly, the sequencing platform must have ultra-high throughput, and its computing cost and occupied server resources are huge. However, species with very large genomes often have a large number of repetitive regions, and short reading and long sequencing technology are difficult to span. At the same time, a large number of short clips lead to extremely complicated genome assembly, and it is difficult to get ideal results. For example, *Paris japonica*[62] has the largest genome of any plant yet assayed, about 150 Gb which is 50 times larger than that of a human haploid genome. The genome size of *Paris polyphylla*[63], the same genus as former, is about 82.55 Gb, making it the largest genome draft to date. The difference in genome size of plants belonging to the family *Nigellaceae* is as high as 230 times, which is an ideal model for studying the change of genome size. Deciphering its genome is of great significance for studying the evolution in genome size and biosynthesis pathway of *Paris polyphylla* saponins. In this study, 10.25 Tb of sequenced reads were assembled by using SOAPdenovo2[64] to get a genome sketch of 70.18 Gb, but it was not assembled by sequencing with the third generation technology, and it was not mounted on the chromosome with Hi-C. The genome of gymnosperm *Pinus tabulaeformis* was assembled to the chromosome level, reaching 25.4 Gb[65]. The genome contains a large number of super-long introns, the average length of which is 10 kb. It takes 1.3 million CPU hours to assemble the whole genome by WDL (Workflow Description Language)-Canu. Therefore, an ultra-high throughput sequencer, a method of obtaining ultra-long fragments, a set of resource-saving assembly algorithms and a powerful CPU are urgently needed to overcome the last field of genome sequencing.

Sequencing and assembling the genome of polyploid plants are also a future trend. Polyploidy mainly occurs in angiosperms, and many polyploid plants are of great value in agricultural production. It can be divided into two types: autopolyploidy from whole genome doubling, like *Medicago sativa*[66]; allopolyploidy whose chromosome doubles after interspecific or intraspecific hybridization, such as allohexaploid *Triticum aestivum*[67] (AABBDD, 2n = 6x = 42) and allotetraploid *Arachis hypogaea*[68] (AABB，2n = 4x = 40). The relationship between phenotypes and genotypes of polyploids is more complicated. For example, they need complicated regulation to ensure the consistent expression of homologous genes. In genome assembly, autopolyploid is more difficult than allopolyploid, because the whole genome doubling event will result in highly similar segments. Therefore, one of the challenges in genome assembly is that similar fragments in two subgenomes can't be assembled by mistake.

The whole genome of rice gap-free was evaluated by gene BUSCO. ZS97RS3 and MH63RS3 both covered 99.88% of the reference gene sets, but there was no gap in the reference genome, but BUSCO still didn't reach 100%. BUSCO may not be the best method to evaluate genome integrity in the future, and whether the genome is gap-free may replace BUSCO to evaluate gene integrity. In the near future, we can predict that simple diploid plants will be sequenced and assembled into gap-free genomes, and the data quality of large plant genomes and polyploid plant genomes will be greatly optimized.

Acknowledgments

The authors acknowledge the funding from National Natural Science Foundation of China (32172614) and a startup fund from Hainan University to Fei Chen.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 1 August 2022; Accepted 13 September 2022; Published online 26 September 2022

REFERENCES

1. Marks RA, Hotaling S, Frandsen PB, VanBuren R. 2021. Representation and participation across 20 years of plant genome sequencing. *Nature Plants* 7:1571−78
2. Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. 2022. A near-complete assembly of an *Arabidopsis thaliana* genome. *Molecular Plant* 15:1247−50
3. Frei D, Veekman E, Grogg D, Stoffel-Studer I, Morishima A, et al. 2021. Ultralong Oxford nanopore reads enable the development of a reference-grade perennial ryegrass genome assembly. *Genome Biology and Evolution* 13:evab159

4. Yang X, Gao S, Guo L, Wang B, Jia Y, et al. 2021. Three chromosome-scale *Papaver* genomes reveal punctuated patchwork evolution of the morphinan and noscapine biosynthesis pathway. *Nature Communications* 12:6030
5. Deng Y, Liu S, Zhang Y, Tan J, Li X, 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. *Molecular Plant* 15:1268−84
6. Song J, Xie W, Wang S, Guo Y, Koo DH, et al. 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant* 14:1757−67
7. Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, et al. 2020. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology* 21:121
8. Liu H, Wang X, Wang G, Cui P, Wu S, et al. 2021. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nature Plants* 7:748−56
9. Zhang L, Hu J, Han X, Li J, Gao Y, et al. 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature Communications* 10:1494
10. Gong L, Wong CH, Idol J, Ngan CY, Wei CL. 2019. Ultra-long read sequencing for whole genomic dna analysis. *Journal of Visualized Experiments* 145:e58954
11. Prall TM, Neumann EK, Karl JA, Shortreed CG, Baker DA, et al. 2021. Consistent ultra-long DNA sequencing with automated slow pipetting. *BMC Genomics* 22:182
12. Lang D, Zhang S, Ren P, Liang F, Sun Z, et al. 2020. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* 9:giaa123
13. Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics* 21:597−614
14. Yuan Y, Chung CY, Chan TF. 2020. Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal* 18:2051−62
15. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* 37:1155−62
16. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* 13:1050−54
17. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10:563−69
18. Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32:2103−10
19. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37:540−46
20. Kamath GM, Shomorony I, Xia F, Courtade TA, Tse DN. 2017. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research* 27:747−56
21. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. 2017. Canu: scalable and accurate long-read assembly via adaptivek *k*-mer weighting and repeat separation. *Genome Research* 27:722−36
22. Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods* 17:155−58
23. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology* 38:1044−53
24. Di Genova A, Buena-Atienza E, Ossowski S, Sagot MF. 2021. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nature Biotechnology* 39:422−30
25. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, et al. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* 30:1291−305
26. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* 18:170−75
27. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* 31:1119−25
28. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356:92−95
29. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* 5:833−45
30. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, et al. 2018. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nature Genetics* 50:1565−73
31. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, et al. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology* 20:224
32. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. In Gene Prediction. Methods in Molecular Biology, ed. Kollmar M. vol 1962. New York: Humana, New York. pp. 227–45 https://doi.org/10.1007/978-1-4939-9173-0_14
33. Wick RR, Holt KE. 2019. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* 8:2138
34. Liu H, Wu S, Li A, Ruan J. 2021. SMARTdenovo: A *de novo* Assembler Using Long Noisy Reads. *Gigabyte* 2021:1−9
35. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, et al. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 15:461−68
36. Bzikadze AV, Pevzner PA. 2020. Automated assembly of centromeres from ultra-long error-prone reads. *Nature Biotechnology* 38:1309−16
37. Chen Y, Nie F, Xie S, Zheng Y, Dai Q, et al. 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* 12:60
38. Hu J, Fan J, Sun Z, Liu S. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36:2253−55
39. Vaser R, Šikić M. 2021. Time- and memory-efficient genome assembly with Raven. *Nature Computational Science* 1:332−36
40. Du H, Liang C. 2019. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nature Communications* 10:5360
41. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C, et al. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* 16:259
42. Ghurye J, Pop M, Koren S, Bickhart D, Chin CS. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18:527
43. Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* 13:R56
44. Xu M, Guo L, Gu S, Wang O, Zhang R, et al. 2020. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* 9:giaa094
45. Chu C, Li X, Wu Y. 2019. GAPPadder: a sensitive approach for closing gaps on draft genomes with short sequence reads. *BMC Genomics* 20:426
46. English AC, Richards S, Han Y, Wang M, Vee V, et al. 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* 7:e47768

47. Chen P, Jing X, Ren J, Cao H, Hao P, et al. 2018. Modelling BioNano optical data and simulation study of genome map assembly. *Bioinformatics* 34:3966−74
48. Michaeli Y, Ebenstein Y. 2012. Channeling DNA for optical mapping. *Nature Biotechnology* 30:762−63
49. Wang B, Yang X, Jia Y, Xu Y, Jia P, et al. 2021. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *Genomics, Proteomics & Bioinformatics* 20:4−13
50. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* 9:1660
51. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, et al. 2022. The complete sequence of a human genome. *Science* 376:44−53
52. Belser C, Baurens FC, Noel B, Martin G, Cruaud C, et al. 2021. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Communications Biology* 4:1047
53. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585:79−84
54. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* 593:101−7
55. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402:761−68
56. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, et al. 2021. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* 374:eabi7489
57. Chen F, Song Y, Li X, Chen J, Mo L, et al. 2019. Genome sequences of horticultural plants: past, present, and future. *Horticulture Research* 6:112
58. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Horticulture Research* 592:737−46
59. Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, et al. 2022. Green plant genomes: What we know in an era of rapidly expanding opportunities. *PNAS* 119:e2115640118
60. Di Marco M, Harwood TD, Hoskins AJ, Ware C, Hill SLL, et al. 2019. Projecting impacts of global climate and land-use scenarios on plant biodiversity using compositional-turnover modelling. *Global Change Biology* 25:2763−78
61. Dodsworth S, Leitch AR, Leitch IJ. 2015. Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development* 35:73−8
62. McGrath CL, Katz LA. 2004. Genome diversity in microbial eukaryotes. *Trends in Ecology & Evolution* 19:32−38
63. Li J, Lv M, Du L, Yunga A, Hao S, et al. 2020. An enormous *Paris polyphylla* genome sheds light on genome size evolution and polyphyllin biogenesis. *bioRxiv* Preprint
64. Luo R, Liu B, Xie Y, Li Z, Huang W, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1:2047-217X-1-18
65. Niu S, Li J, Bo W, Yang W, Zuccolo A, et al. 2022. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* 185:204−217.E14
66. Chen H, Zeng Y, Yang Y, Huang L, Tang B, et al. 2020. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature Communications* 11:2494
67. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, et al. 2020. Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588:277−83
68. Zhuang W, Chen H, Yang M, Wang J, Pandey MK, et al. 2019. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nature Genetics* 51:865−76