Technology-enabled great leap in deciphering plant genomes

Longjiang Fan

fanlj@zju.edu.cn

Zhejiang University https://orcid.org/0000-0003-4846-0500

Xiaojiao Gong

Zhejiang University

Lingjuan Xie

Institute of Crop Science & Institute of Bioinformatics, Zhejiang University

Kun Yang

Zhejiang University

Leti Shen

Zhejiang University

Yujie Huang

Institute of Crop Science, Zhejiang University

Shiyu Zhang

Zhejiang University

Yanqing Sun

Zhejiang University

Dongya Wu

Zhejiang University https://orcid.org/0000-0003-1967-2264

Chuyu Ye

Zhejiang University https://orcid.org/0000-0003-0903-0356

Qian-Hao Zhu

CSIRO, Agriculture and Food https://orcid.org/0000-0002-6505-7417

Article

Keywords: plant genome, sequencing technology, assembly algorithm, T2T genome, pan-genome

Posted Date: October 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3376742/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Plants on March 20th, 2024. See the published version at https://doi.org/10.1038/s41477-024-01655-6.

Abstract

Plant genomes provide essential and vital basic resources to study multifarious aspects of plant biology and applications (e.g. breeding). From 2000 to 2020, 1,031 genomes of 788 plant species were sequenced. From 2021 to 2023, 1,857 genomes of 841 plant species, including 622 newly sequenced species, were assembled, showing a great leap. The 1,857 newly assembled genomes, with some being telomere-to-telomere (T2T) assembles, cover the major phylogenetic clades and many of them have a high quality. The achievement is mainly attributed to significant advances in both sequencing technologies and assembly software. A database named N3: plants, genomes, technologies was developed to accommodate the metadata associated with the sequenced plant genomes. In the end, we discussed the challenges involved in building huge size gap-free and single-cell genomes as well as the future opportunities in plant genomic studies.

Introduction

Most plant species have a complex genome, including large size, extremely high repeat content, high heterozygosity, and polyploidy[1,2]. Genome sequencing plays an important role in providing unparalleled genomic resources for deeper understanding the biology and evolution of plants[3]. Since the publication of the first plant reference genome of *Arabidopsis thaliana* in 2000[4], benefiting from notable improvements in sequencing technology and computational algorithms for genome assembly, a great many of plant genomes have been sequenced and assembled from 2000 to 2020, including 1,031 *de novo* genomes of 788 different plant species[5]. The genomes provide a myriad of high-quality genomic resources for studies of plant functional genomics, population genetics, evolution, and breeding. However, given the number of green plant species (Viridiplantae) being approximately 500,000[6], the number of plant species that have been sequenced so far is just a tip of the iceberg. More plant genomes need to be decoded and deciphered in order to have a comprehensive understanding of plant biology and evolution and so to better serve humanity in the long run, which largely relies on advancing in sequencing and assembling technologies.

At the beginning of plant genome sequencing, the predominate approach applied was the Sanger technology or the first-generation sequencing technology[7], and gradually, the next-generation sequencing (NGS) technologies (e.g., Illumina) overtook the first-generation technology, especially when doing large-scale genome resequencing[8]. With further technological innovations, we are now in the third-generation sequencing (TGS) era, characterized by its capability of generating very long reads to overcome the drawback of short reads of NGS[9]. The TGS technologies include the Pacific Biosciences (PacBio) continuous long reads (CLR) / circular consensus sequencing (CCS), the Oxford Nanopore Technologies (ONT) sequencing platforms, and so on[10,11]. Although the CLR method has played an important role in PacBio sequencing for several years, high fidelity (HiFi) reads generated by CCS have attracted more and more attention due to significant improvement in sequencing accuracy[12]. The accuracy of the ONT sequencing platforms is still not as high as that of the PacBio platforms, but ONT has the advantage of

being able to generate ultra-long reads, which can resolve the issue of large repeats or compensate for HiFi's coverage dropouts[13]. Meanwhile, multiple software developed using different algorithms offer us the opportunity to create more complete assembly[14]. From Velvet[15], ABySS[16], AllPath-LG[17], and SOAPdenovo[18] contig assemblers by short reads to Canu[19], Mecat2[20], Necat[21], Falcon[22] by TGS technologies, more high-quality genomes have been assembled. Hifiasm, a software using HiFi reads as input for contig assembly, has been applied in generating haplotype-resolved *de novo* assembly, even telomere-to-telomere (T2T) assembly when combining HiFi reads with ultra-long ONT reads[5,23].

Here, we updated the progress of plant genome sequencing and found that the number of plant genomes (1,857 genomes from 841 plant species) released in the past two years (from January 2021 to June 2023) is more than that released in the first twenty years (2000 – 2020). We further analyzed the technologies used in plant genome sequencing and assembly, shedding lights on technology enabled great leap in plant genome research and identifying the challenges faced by the plant community to further advance genome sequencing and assembly in the years to come.

Results

A rapid increase of sequenced plant genomes in the past two years

During the past several years, we have witnessed rapid and great progresses in sequencing technology and computing algorithm, making it relatively easier for genome assembly and for achieving high-quality assembly, compared to ~10 or so years ago. Compared with the 1,031 genomes of 788 plant species published from 2000 to 2020[5], we have observed a great leap in *de novo* genomes in the past two years, with a total of newly generated 1,857 genomes representing 841 plant species (including 622 newly sequenced species). As a result, a total of 2,836 genomes of 1,410 plant species are available up to now (Figure 1A). In addition to the surge in quantity, the quality of plant genome assemblies has also increased rapidly, with 72.0% of chromosome level assemblies in the past two years compared to 44.1% in the first twenty years (Figure 1A). There were not significant changes in the size of the sequenced genomes in the first 20 years and the last two years, with a minimum, maximum and median value of 12.40 Mb, 27.60 Gb and 522.60 Mb, and 16.70 Mb, 26.45 Gb and 475.00 Mb, respectively. However, contig N50 and scaffold N50 has increased significantly, from 1.36±3.48 Mb and 20.58±76.51 Mb (2000-2020) to 10.37±13.63 Mb and 68.71±106.67 Mb (2021-June 2023) (P 0.01), respectively (Figure 1B). Moreover, the innovation of technology provides the impetus for the sequencing and assembly of the plant pan-genome. Among these *de novo* genomes, the number of genomes assembled by the pan-genome project accounted for a considerable part, contributing to the surge in the plant genome sequencing in the past two years (Figure 1A).

Compared to the plant species in the first 20 years[5], the plant species sequenced in the last two years showed similar diversity in Viridiplantae, but taxonomic species gaps persist. The plant species from several new clades, such as Acorales, Buxus, Chloranthales, and Cycads, were covered in the recent two years (Figure 2). However, no species of the ANA (Amborellales, Nymphaeales and Austrobaileyales)

clade have been sequenced in past two years. In both the previous two decades and the last two years, the most sequenced species are angiosperms, including monocot Commlinids and eudicot Asterids and Rosids (Figure 2) (see Table S1 for the details of all the sequenced species).

To eliminate the effect of the number of pan-genomes on the analysis results, we reanalyzed all the sequenced plant species since 2000 by excluding the pan-genomes. We found that plant species from a total of 91 orders have been assembled so far and among these sequenced genomes, most were from Poales (Commlinids), Brassicales (Rosids), Fabales (Rosids), Rosales (Rosids), and Lamiales (Asterids) (Supplementary Table 1). Although the number of published genomes varies among the other orders, the Poales (near 130 and 190 genomes in the first twenty years and the past two years, respectively) was consistently the most studied. Meanwhile, Poaceae, which includes major food crops (such as rice, wheat, and maize), was the main target and had many more sequenced (near 129 and 160 genomes in the first twenty years and the past two years, respectively) than other families in Poales in both time periods (Supplementary Figure 1). Two countries (China and the United States) contributed the most to plant *de novo* assembly in both time periods, with China's contribution to the sequenced genomes rising from about 30% to 60% (Supplementary Figure 2; Supplementary Table 1). The plant genomes released in the first 20 years[5] were reported by 767 articles in near 100 different journals, with *Nature Genetics, Nature Communications, GigaScience, The Plant Journal,* and *DNA Research* being the top five journals in terms of the number of plant genomes published, while the genomes reported in the past two years were published in 829 articles in near 100 different journals (Supplementary Table 1), with *Horticulture Research, Nature Communications, Frontiers in Plant Science, The Plant Journal,* and *G3-Genes Genomes Genetics* becoming the top journals (Supplementary Table 1; Supplementary Figure 2). Based on the word cloud analysis for the titles of the recent 829 articles, the sequenced genomes contributed a diverse of study topics, including the origin, domestication, and evolution of plant species, biosynthetic pathways of key substances, resistances, molecular mechanisms regulating plant biological processes, agronomic traits of crops, and molecular breeding (Supplementary Figure 3). Compared with the previous 20 years, the biggest change is that “chromosome level” appeared more frequently, while “draft” appeared less frequently (Supplementary Figure 3).

Pan-genomes, haplotype-resolved and T2T assembles

During the past two years, in addition to generating more *de novo* reference genomes of plant species that had not been sequenced previously, for the plant species already with a reference, genomic studies shifted to generating of pan-genomes, haplotype-resolved genomes and telomere-to-telomere (T2T) assembles. Pan-genome consists of the core genome and the unique genome of all individuals of a species[24]. In the recent two years, generation of plant pan-genomes has been attempted in soybean[25], Arabidopsis[26], barley[27], wheat[28], cotton[29], tomato[30], potato[31], and rapeseed[32], especially in the main crops rice[33,34] and maize[35,36]. Overall, *de novo* genomes of plant species from seven families, including Poaceae, Brassicaceae, Solanaceae, Fabaceae, Malvaceae, Cucurbitaceae, and Rutaceae have been

assembled and used in pan-genome construction (Table 1). As for the sequencing platforms applied in construction of pan-genomes, and TGS was used for 97% (12% by HiFi and 85% by other TGS platforms) of plant assemblies, while 3% of the genomes were assembled by only NGS reads. No ONT ultra-long reads have been used yet in pan-genome assembly (Supplementary Figure 4). While the quality of the pan-genomes assembled in the first half of 2021 is relatively low, the mean contig N50 of pan-genomes assembled afterwards is generally in the range of 4 Mb to 30 Mb and the genomes have a relatively high continuity (Supplementary Figure 4). Currently, the typical number of *de novo* genomes used in assembly of pan-genomes is 10-15 (Supplementary Figure 4). The most used software for contig assembly of pan-genome is Canu, Falcon, and Hifiasm (Supplementary Table 1). Pilon dominates in the polishing tool used. 3D-DNA, LACHESIS, ALLMAPS, and BioNano Solve are the most used for scaffolding (Supplementary Table 1).

Haplotype-resolved or phased genome separates heterozygous genomic regions and helps genome annotation, mutation detection, evolutionary analysis, gene function characterization, comparative genomics, and so on[1,37]. Haplotype-resolved genomes are current available for 17 orders, with the top five being Poales, Solanales, Vitales, Malvales, and Sapindales (Supplementary Table 1). Most haplotype-resolved assemblies utilize TGS and 37% of them used HiFi reads. The most commonly used software for assembling haplotype-resolved genomes includes Hifiasm, Falcon, and Canu, with Hifiasm being utilized 2-3 times more often than Falcon and Canu (Supplementary Table 1).

T2T genome refers to a high-quality, fully complete genome assembly, including all centromeres and repeating regions, with high precision, continuity, and integrity[13]. Among the assembled T2T assemblies, the quality is variable due to different criteria used in defining a T2T genome. In this study, only those with at least half of the total number of chromosomes with both ends, i.e., telomere, assembled were considered as T2T genome. With this criterion, a total of 30 genomes from 20 species have been assembled to the T2T level. All of these assemblies belong to angiosperms, including Poales, Brassicales, Rosales, Solanales, and so on. The majority (≥22) of the 30 T2T assemblies is diploid (Supplementary Table 1). Most T2T assembles adopted a combination of PacBio, ONT, and NGS sequencing technologies and nearly 90% of these T2T genomes used HiFi reads. In contig assembling, in addition to Hifiasm, other contig assembling software were also used for an integrated assembling version.

The innovation of sequencing technologies and assembly software/algorithm

Regarding the sequencing strategies used in the assembly of plant genomes reported during the last two years, the TGS technology was the dominated one (91.2%), only 8.8% of the sequenced genomes were solely based on the NGS technology (Figure 3A). No apparent relationship was observed between the genome sizes and their contig N50 sizes, i.e., big genomes also can have a high continuity of assembly (Figure 3A). Although the CLR method integrated with Sequel and RS II platforms has played an important role in PacBio sequencing in the past two years, HiFi reads generated by the CCS method

attracted more and more attention (used by 17.4% of the sequenced genomes) due to its accuracy improvement (Figure 3A). HiFi reads were adopted by the assembles of almost all genome sizes, generating a significantly higher contig N50 size compared to the traditional TGS reads, indicating the advantages of HiFi reads for genome assembly (Figure 3A). The use of HiFi reads was significant increased in around the beginning of 2022, with 21.1% of utilization frequency in 2022 and 28.5% in 2023, compared with 5.1% in 2021 (Supplementary Figure 5; Supplementary Table 1). And the number of genomes assembled by ONT ultra-long reads (reads N50 size ≥100kb) is low.

Generating a genome usually involves three stages, including contig assembly, polishing, and scaffolding. In the past two years, the *de novo* assemblers including Canu, Falcon, Hifiasm, Nextdenovo, and wtdbg were used in contig assembly of most sequenced genomes, particularly Hifiasm, which showed an a significantly increase trend (Figure 3B; Figure 3C). Polishing of assembly (except pan-genomes) usually used NGS reads generated by Ilumina's HiSeq, Novaseq, and MiSeq series, and BGI's DNBSEQ and MGISEQ series and typically requires multiple iterations. Approximately 23% assemblies were polished by using Pilon, following by a combination of Racon and Pilon (17%), Arrow and Pilon (11%) (Supplementary Table 1). The top five software used in polishing genomes includes Pilon, Racon, Arrow, NextPolish, and Quiver (Figure 3C). Five approaches, including physical linkage map, genetic map, mate pair linkage reads, high-throughput chromosome conformation capture (Hi-C) reads, and homology-based approaches, were generally used in the scaffolding stage. Currently the focus of this stage is chromosomal level assembly, with Hi-C reads being the major contributor, because 84% of chromosome-scale genomes used Hi-C reads, alone or in combination with others, such as genetic map (RagTag) and Bionano optical map (BioNano Sovle), to link contigs to scaffolds (Supplementary Table 1). The most common scaffolding software involving Hi-C reads in building chromosomes includes 3D-DNA, LACHESIS, and ALLHIC (Figure 3C).

The N3 database: a hub of sequenced plant genomes and assembly technologies

To provide the details of the plant genomes collected and analyzed in this study, such as the sequencing platforms and assembly tools used by each of the published plant reference and *de novo* genome, we built a database named *N3: plants, genomes, technologies (*http://ibi.zju.edu.cn/n3/) (Figure 4). It included all available 2,836 plant reference/*de novo* genomes collected by our previous study[5] and this study. The database mainly consists of four modules: Statistics, Search, Pan&T2T, and Links modules. The ‘Statistics’ module summarizes the total number of genome/species that have been *de novo* assembled so far and the technologies (sequencing and assembly) used in generating the genomes. In the ‘Search’ module, users can search any available plant genome information by choosing species names, sequencing platform and assembly software. The ‘Pan&T2T’ module presents the information about the available plant pan-genomes and T2T assemblies. In the ‘Links’ module, we provide the web links to the main sequencing platforms and assembly software.

Discussion

A total of 2,836 genomes from 1,410 plant species have been sequenced up to now. However, comparing to the total number of plant species, that is just a tip of the iceberg and many species, including the basal groups of angiosperms (e.g. Amborellales, Nymphaeales, and Austrobaileyales) are yet to be sequenced because of their negligible economic values. Even crop species with giant economic value were sequenced, the genomes of their closely related species are not taken seriously, although they are useful and essential for understanding crops' evolution trajectory and providing superior alleles for agronomic traits. Notwithstanding the quality of the genomes assembled during the past two years has improved significantly compared to those assembled in the previous two decades, however, the assembly quality of many genomes, particularly the giga-genomes (with a genome size >10 Gb), still remain to be improved, mainly due to the nature of massive repetitive sequences and high heterozygosity of plant genomes[38,39]. With the further advance in sequencing technologies, the heavy workload involved in sequencing large genomes can be partially addressed by the high volume of sequencing output and the repetitive sequences can be addressed by long and ultra-long sequence reads. It is thus expected a significant increase of plant giga-genomes in the near future.

Long read sequencing is a key to a high-quality genome assembly. In recent years, HiFi reads are preferred, especially in haplotype-resolved assembly and T2T assembly. Although the lengths of the ONT ultra-long reads can even reach mega-base level, the base accuracy is not as high as that of HiFi reads. Longer HiFi reads and more precise ONT ultra-long reads with lower sequencing costs are expected in the near future, which will accelerate the acquisition of more complete plant (pan-)genomes and provide opportunities in studying hidden regions along chromosomes (e.g. centromere). Centromere is one of the important components of chromosomes and plays a key role in mitosis and meiosis[40]. Unfortunately, due to highly repetitive sequences and complex structures, centromeric regions have been rarely explored[41]. The recently published *Actinidia chinensis* and *Daucus carota*, for example, used HiFi reads and ONT ultra-long reads to identify all chromosomal centromeres[14,42]. It is expected that more and more of plant genome with intact chromosomal centromeres will be available in future.

Single-cell genome sequencing can reveal cell heterogeneity in tissues and cell developmental trajectory. It has quickly become a new research hotspot in recent years and will continue to be one of the future research directions. Compared to the studies of single-cell genomics in human and microbiome[43–45], that in plants is still in its infancy. Single-cell genomics involves sequencing thousands or more genomes in parallel, so heavily depends on sequencing capacity, which is still a challenge for most sequencing platforms. To facilitate studies on single-cell genomics, in addition to enhancing sequencing throughput, computational algorithms and data storage that can handle increasing volume of sequencing data are to be improved.

Methods

With the aim to have an as complete as possible list of plant genome assemblies of the Viridiplantae clade, which consists of green algae and land plants, we checked and searched genome related papers in the Web of Science (not including BioRxiv), downloaded papers on genome assembly from Published Plant Genomes (https://www.plabipd.de/plant_genomes_pa.ep), and conducted an extended search based on the genomes mentioned in the published articles. The complete list of the genomes sequenced during the past two years and the approximately 830 papers describing the genomes are provided in Supplementary Table 1. All the 2,836 genomes of 1,410 plant species published so far and the 1,600 related papers (including our previous study[5]) were available at the N3 database (http://ibi.zju.edu.cn/n3/).

For the purpose of this study, we only used the genomes by *de novo* assembly with a contig N50 size over 10kb. For the same plant accession with multiple genome assembles, all assemblies were included in the Supplementary Table 1 and the N3 database, but only counted once; the updated genome and different haplotypes of the same accession also only counted once. The family and order of the plant species are based on the Taxonomy plate on NCBI (https://www.ncbi.nlm.nih.gov/) and Wikipedia (https://www.wikipedia.org/). Classification of species were classified according to the Angiosperm Phylogeny Group IV (APG IV) (http://www.mobot.org/MOBOT/research/APweb/) and Phylogenetic inferences of one thousand plant transcriptomes[46]. The estimated genome size of all species was based on *k*-mer analysis or flow cytometry. If the genome size of a species was not mentioned in the source article, we used the estimated size of the species in Published Plant Genomes. Country/institution refers to the region where the first publication was located. Different versions of a software used in genome assembly were considered as the same. For the pan-genome assemblies, we also counted only the *de novo* assembles and excluded those with a contig N50 size less than 10kb. In order to exclude the influence of pan-genome assembly on the analysis results, we analyzed separately the genome assembles by excluding pan-genomes and therefore generated two datasets, one with all genomes and another without pan-genomes.

We built the N3 database by SQLyog visual tool (https://webyog.com/product/sqlyog/) and processed and analyzed the data using Python and R scripts.

Data availability

All metadata associated with this study, including accession numbers of all genome assemblies, the quality statistics of the assembles and the technologies (sequencing platforms and assembly tools) used in generating the published plant genomes were available in the N3 database (http://ibi.zju.edu.cn/n3/).

Code availability

The primary code used in this study was provided at http://ibi.zju.edu.cn/n3//links.php.

References

1. Zhang, T. *et al.* Complex genome assembly based on long-read sequencing. *Brief. Bioinform.* **23**, bbac305 (2022).

2. Edwards, D., Batley, J. & Snowdon, R. J. Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11 (2013).

3. Marks, R. A., Hotaling, S., Frandsen, P. B. & VanBuren, R. Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* **7**, 1571–1578 (2021).

4. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).

5. Sun, Y., Shang, L., Zhu, Q.-H., Fan, L. & Guo, L. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* **27**, 391–401 (2022).

6. Nie, Y. *et al.* Accounting for uncertainty in the evolutionary timescale of green plants through clock-partitioning and fossil calibration strategies. *Syst. Biol.* **69**, 1–16 (2020).

7. Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264 (2008).

8. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).

9. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: an overview. *Hum. Immunol.* **82**, 801–811 (2021).

10. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100 (2017).

11. Lang, D. *et al.* 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 (2020).

12. Hon, T. *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* **7**, 399 (2020).

13. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

14. Yue, J. *et al.* Telomere-to-telomere and gap-free reference genome assembly of the kiwifruit *Actinidia chinensis*. *Hortic. Res.* **10**, uhac264 (2023).

15. Shirasawa, K. *et al.* An improved reference genome for *Trifolium subterraneum* L. provides insight into molecular diversity and intra-specific phylogeny. *Front. Plant Sci.* **14**, 1103857 (2023).

16. McLay, T. G. B. *et al.* A genome resource for *Acacia*, Australia's largest plant genus. *PLOS ONE* **17**, e0274267 (2022).

17. Alabi, N., Wu, Y., Bossdorf, O., Rieseberg, L. H. & Colautti, R. I. Genome report: a draft genome of *Alliaria petiolata* (garlic mustard) as a model system for invasion genetics. *G3 GenesGenomesGenetics* **11**, jkab339 (2021).

18. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).

19. Zheng, P. *et al.* Comparative analyses of American and Asian lotus genomes reveal insights into petal color, carpel thermogenesis and domestication. *Plant J.* **110**, 1498–1515 (2022).

20. Xiao, C.-L. *et al.* MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).

21. Zhang, B. *et al.* The chromosome-scale assembly of endive (*Cichorium endivia*) genome provides insights into the sesquiterpenoid biosynthesis. *Genomics* **114**, 110400 (2022).

22. Hale, I. *et al.* Genomic resources to guide improvement of the shea tree. *Front. Plant Sci.* **12**, 720670 (2021).

23. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

24. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci.* **102**, 13950–13955 (2005).

25. Li, Y. *et al. De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).

26. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).

27. Jayakodi, M. *et al.* The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).

28. Montenegro, J. D. *et al.* The pangenome of hexaploid bread wheat. *Plant J.* **90**, 1007–1013 (2017).

29. Li, J. *et al.* Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol.* **22**, 119 (2021).

30. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).

31. Tang, D. *et al.* Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).

32. Song, J.-M. *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45 (2020).

33. Qin, P. *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542-3558.e16 (2021).

34. Wu, D. *et al.* A syntelog-based pan-genome provides insights into rice domestication and de-domestication. *Genome Biol* 24, 179 (2023).

35. Hufford, M. B. *et al. De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).

36. Wang, B. *et al. De novo* genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nat. Genet.* **55**, 312–323 (2023).

37. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).

38. Neale, D. B. *et al.* Assembled and annotated 26.5 Gbp coast redwood genome: a resource for estimating evolutionary adaptive potential and investigating hexaploid origin. *G3 GenesGenomesGenetics* **12**, jkab380 (2022).

39. Niu, S. *et al.* The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204-217.e14 (2022).

40. Feng, C. *et al.* Recent advances in plant centromere biology. *Sci. China Life Sci.* **58**, 240–245 (2015).

41. Perumal, S. *et al.* A high-contiguity *Brassica nigra* genome localizes active centromeres and defines the ancestral *Brassica* genome. *Nat. Plants* **6**, 929–941 (2020).

42. Wang, Y.-H. *et al.* Telomere-to-telomere carrot (*Daucus carota*) genome assembly reveals carotenoid characteristics. *Hortic. Res.* **10**, uhad103 (2023).

43. Fan, X. *et al.* SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. *Genome Biol.* **22**, 195 (2021).

44. Zheng, W. *et al.* High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).

45. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

46. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).

Tables

Table 1. Summary of the studies on plant pan-genomes and T2T genomes in the last two years (from January, 2021 to June, 2023)

Species	Pan-genome	*de novo* genomes generated
Arabidopsis thaliana	Glick *et al.*, 2021, Molecular Ecology Resources	15
Oryza sativa;Oryza glaberrima	Qin *et al.*, 2021, Cell	31
Oryza barthii;Oryza glaberrima;Oryza rufipogon;Oryza sativa	Shang *et al.*, 2022, Cell Research	251
Oryza rufipogon;Oryza sativa	Zhang *et al.*, 2022, Genome Research	88
Raphanus raphanistrum;Raphanus raphanistrumm x Raphanus sativus;Raphanus sativus	Zhang *et al.*, 2021, Molecular Plant	11
Brassica rapa	Cai *et al.*, 2021, Genome Biology	16
Brassica napus	Cui *et al.*, 2023, Plant Communications	11
Cucumis sativus	Li *et al.*, 2022, Nature Communications	11
Cucumis melo	Oren *et al.*, 2022, The Plant Journal	25
Vigna unguiculata	Liang *et al.*, 2023, The Plant Genome	7
Pisum sativum	Yang *et al.*, 2022, Nature Genetics	1
Glycine cyrtoloba;Glycine dolichocarpa;Glycine falcata;Glycine stenophita;Glycine syndetika;Glycine tomentella	Zhuang *et al.*, 2022, Nature Plants	6
Gossypium anomalum;Gossypium bickii;Gossypium herbaceum;Gossypium longicalyx;Gossypium stocksii;Gossypium sturtianum	Wang *et al.*, 2022, Nature Genetics	7
Gossypium hirsutum	Li *et al.*, 2021, Genome Biology	10
Zea mays	Wang et al., 2023, Nature Genetics	12
Zea mays	Hufford *et al.*, 2021, Science	26
Setaria italica	He *et al.*, 2023, Nature Genetics	110
Pennisetum glaucum	Yan *et al.*, 2023, Nature Genetics	10
Sorghum bicolor;Sorghum propinquum	Tao *et al.*, 2021, Nature Plants	13
Citrus sinensis	Gao *et al.*, 2023, iScience	1
Solanum andreanum;Solanum tuberosum;Solanum boliviense;Solanum brevicaule;Solanum buesii;Solanum bulbocastanum;Solanum burkartii;Solanum cajamarquense;Solanum candolleanum;Solanum chacoense;Solanum chomatophilum;Solanum commersonii;Solanum etuberosum;Solanum jamesii;Solanum lignicaule;Solanum morelliforme;Solanum multiinterruptum;Solanum neorossii;Solanum palustre;Solanum paucissectum;Solanum pinnatisectum;Solanum piurae;Solanum sogarandinum;Solanum vernei	Tang *et al.*, 2022, Nature	46
Solanum tuberosum	Karetnikov *et al.*, 2023,	9

	International Journal of Molecular Sciences	
Solanum lycopersicum;Solanum pimpinellifolium	Zhou *et al.*, 2022, Nature	32
Solanum lycopersicoides;Solanum habrochaites;Solanum chilense;Solanum peruvianum;Solanum corneliomulleri;Solanum neorickii;Solanum chmielewskii;Solanum pimpinellifolium;Solanum galapagense;Solanum lycopersicum	Li *et al.*, 2023, Nature Genetics	11

Species	T2T assembly	T2T chromosome/basic number of chromosomes
Arabidopsis thaliana	Hou et al., 2022, Molecular Plant	4/5
Oryza sativa	Song *et al.*, 2021, Molecular Plant	7 and10/12
Oryza sativa	Zhang *et al.*, 2022, Plant Biotechnology Journal	7-10/12
Solanum tuberosum	Yang *et al.*, 2023, Molecular Plant	12/12
Actinidia chinensis	Han *et al.*, 2023, Molecular Plant	28/29
Actinidia latifolia	Han *et al.*, 2023, Molecular Plant	28/29
Actinidia chinensis	Yue *et al.*, 2022, Horticulture Research	29/30; 28/29
Brassica rapa	Zhang *et al.*, 2023, Plant Biotechnology Journal	8/10
Cenchrus fungigraminus	Zheng *et al.*, 2023, Plant Communications	11/14
Chlamydomonas reinhardtii	Payne *et al.*, 2023, Plant Communications	15/17
Citrullus lanatus	Deng *et al.*, 2022, Molecular Plant	11/11
Daucus carota	Wang *et al.*, 2023, Horticulture Research	6/9
Erianthus rufpilus	Wang *et al.*, 2023, Nature Plants	10/10
Fragaria vesca	Zhou *et al.*, 2023, Horticulture Research	7/7
Jasminum sambac	Xu *et al.*, 2023, Journal of Experimental Botany	9/13
Momordica charantia	Fu *et al.*, 2022, Horticulture Research	8/11
Morus notabilis	Ma *et al.*, 2023, Horticulture	4/6

	Research	
Musa acuminata	Belser *et al.*, 2021, Communications Biology	11/11
Rhodomyrtus tomentosa	Li *et al.*, 2023, Horticulture Research	11/11
Thalia dealbata	Tang *et al.*, 2023, Frontiers in Plant Science	6/6
Vitis vinifera	Shi *et al.*, 2023, Horticulture Research	17/19
Zea mays	Chen *et al.*, 2023, Nature Genetics	10/10

Figures

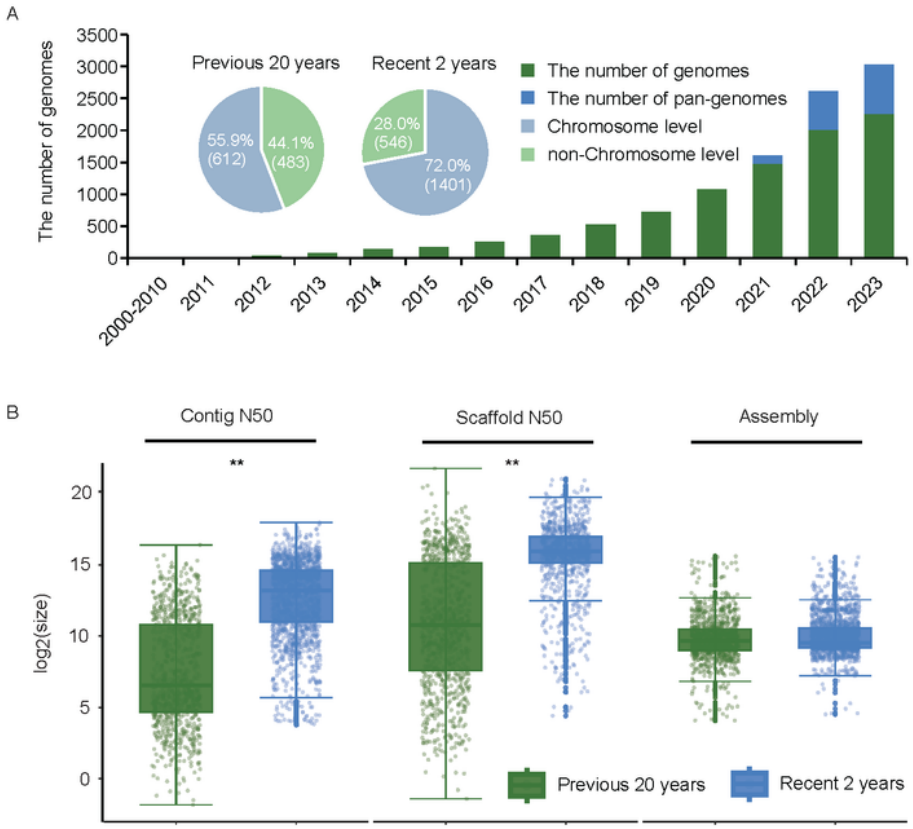

Figure 1

An overview of the published plant genomes. (A) The number of plant genomes sequenced up to now (June 2023) since the publication of the *Arabidopsis thaliana* genome in 2000. (B) Comparison of contig N50, scaffold N50, and the size of the genome assembles generated in the previous 20 years and recent two years.

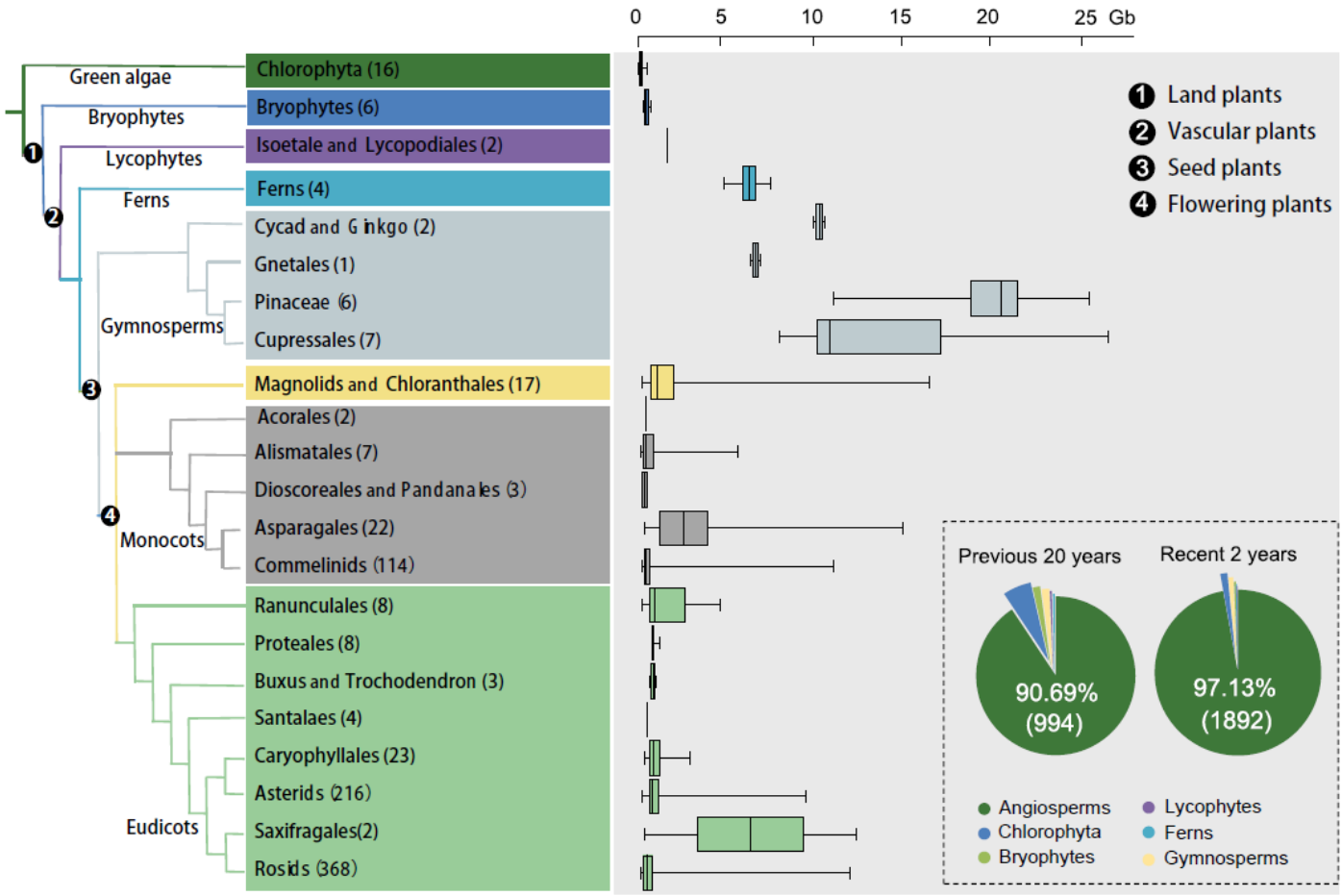

Figure 2

Phylogenetic and distribution of the 1,857 sequenced plant genomes. Phylogenetic clade (left) and genome size (right) distribution of the 1,857 sequenced plant genomes. Phylogenetic relationships and classification are based on the Angiosperm Phylogeny Group IV (APG IV) and phylogenetic inferences of one thousand plant transcriptomes[46]. The numbers in parentheses represent the number of sequenced genomes. The box plots on the right panel show the genome assembly size. The pie charts show the proportion of the sequenced plant species of different groups in the previous twenty years and in the last two years, respectively.

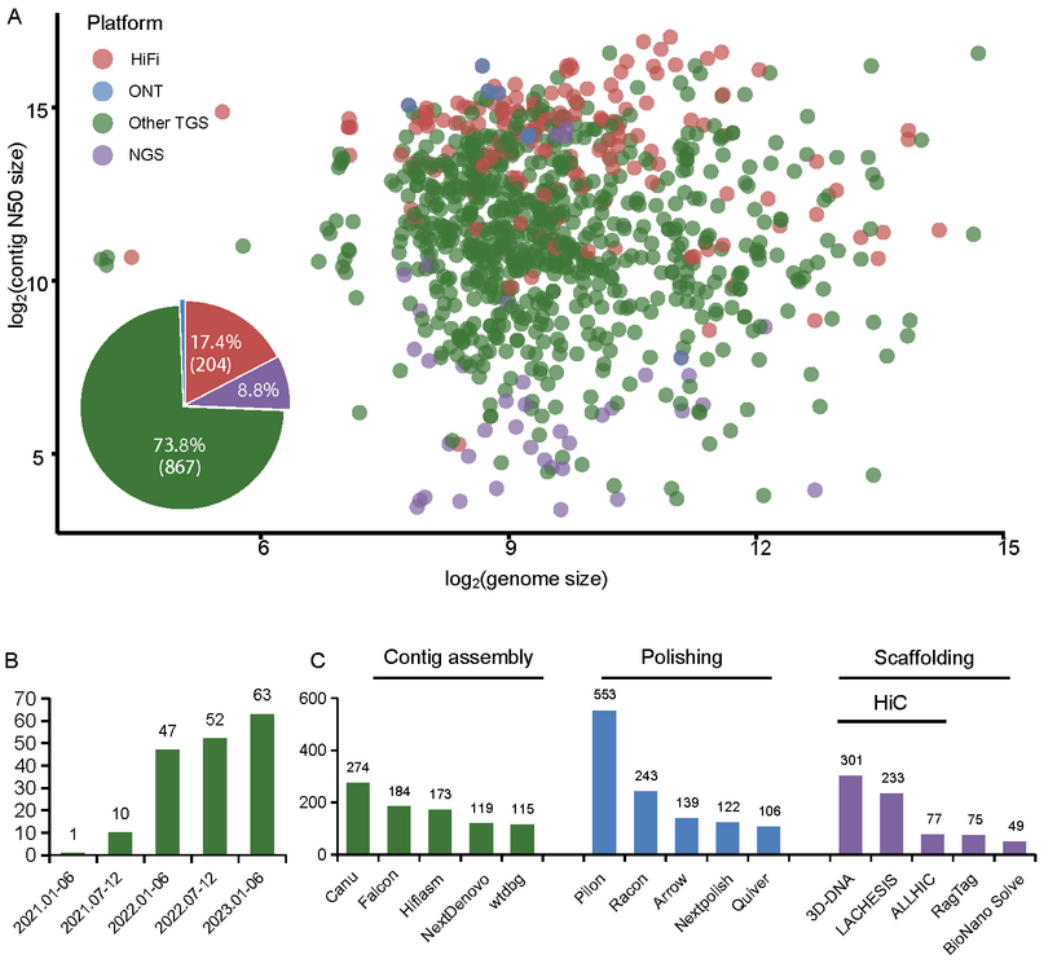

Figure 3

The sequencing platforms and the top software used in the assembly of plant genomes in the past two years. (A) The number of plant genomes generated by using different combination of sequencing platforms. “HiFi” means HiFi reads generated by PacBio sequencing platforms and “ONT” refers to ONT ultra-long reads (reads N50 size ≥100kb); “Other TGS” refers to sequencing technologies that incorporate the third-generation sequencing platforms (HiFi and ONT not included); “NGS” refers to the use of only

the second-generation sequencing platforms. The pie chart shows the proportion of the four sequencing platforms used. (B) The utilization frequency of Hifiasm in the past two and half years with a time slot of half year. (C) The top software used in different stages of plant genome assembly, i.e., contig assembly, polishing and scaffolding, in the past two years. For scaffolding, HiC data, genetic map (RagTag) and Bionano optical map (BioNano) were mainly used.

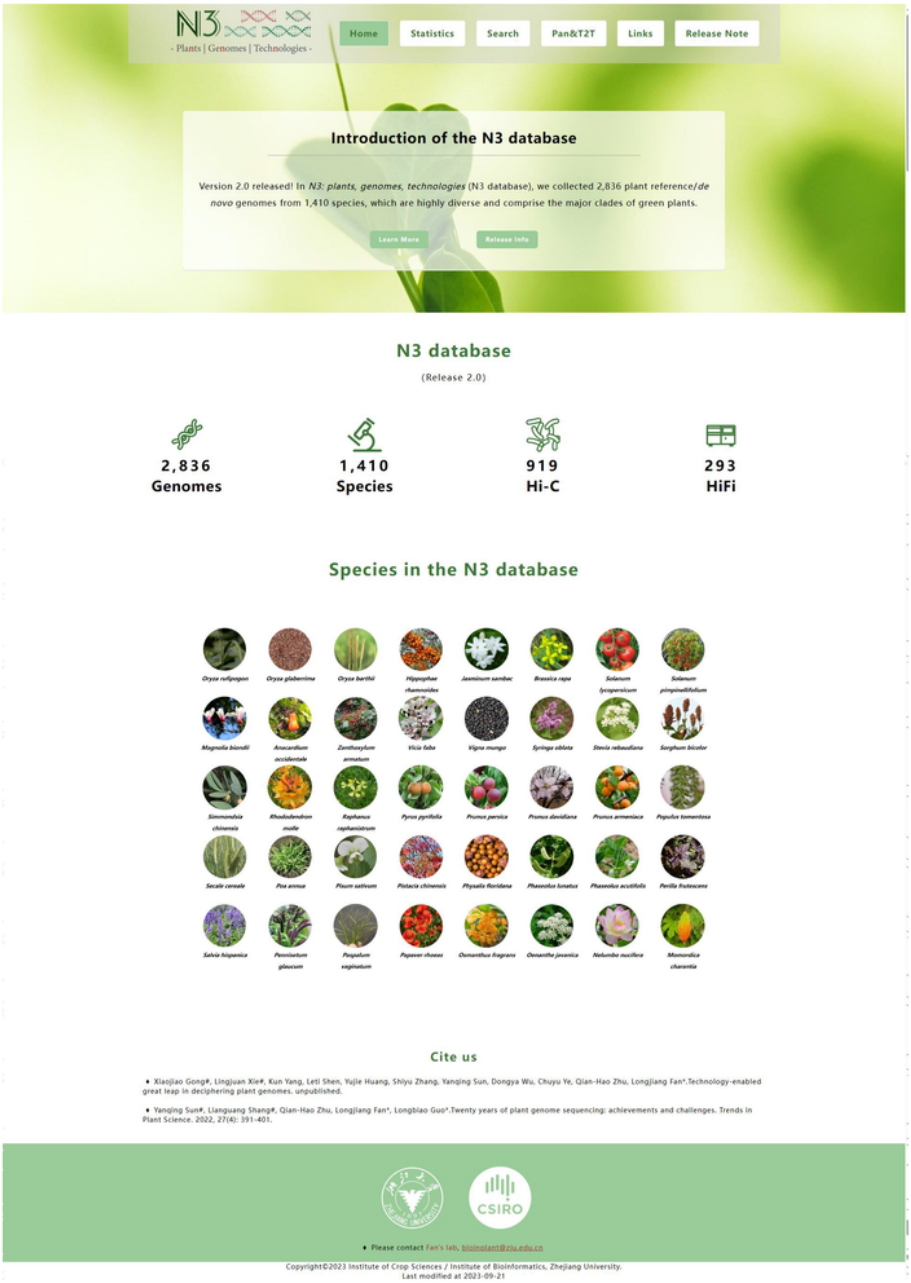

Figure 4

An overview of the N3 database. The homepage of the N3 database at http://ibi.zju.edu.cn/n3/.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx
- SupplementaryTable1.xlsx