scientific data

OPEN

DATA DESCRIPTOR

A chromosomal-level genome assembly of *Begonia fimbristipula* (Begoniaceae)

Tian-Wen Xiao[1,2], Zheng-Feng Wang[1,2,3,4] ✉ & Hai-Fei Yan[1,2,3,5] ✉

Begonia fimbristipula **Hance (Begoniaceae) is a valuable medicinal herb that is classified as a protected species in Guangdong Province, China. In this study, we present a chromosome-level genome assembly of *B. fimbristipula*, aiming to facilitate its conservation and utilization. The genome was assembled using a combination of Oxford Nanopore long-read data and Illumina short-read data. The assembled genome size of *B. fimbristipula* is 462.11 Mb, with a scaffold N50 of 38.22 Mb. A total of 91.96% (424.94 Mb) of the sequences were anchored to 11 pseudochromosomes using Hi-C technology. The genome assembly exhibits a BUSCO completeness of 90.3% and an LTR Assembly Index (LAI) of 17.73. Genome annotation revealed 25,563 protein-coding genes and 274 tRNA genes. The high-quality chromosome-level assembly and annotation provide valuable insights into the genomic characteristics of *B. fimbristipula*, thereby offering essential resources for its conservation and economic utilization.**

Background & Summary

The family Begoniaceae C. Agardh consists of two genera: *Hillebrandia* Oliv. and *Begonia* L. *Hillebrandia* is a monotypic genus, while *Begonia* is one of the ten largest angiosperm genera, comprising over 2,000 species[1]. Species within *Begonia* are perennial herbs widely distributed in moist tropical and subtropical regions worldwide, with some species extending into the warm temperate zone (e.g., *B. grandis* Dryand.)[2]. Members of this genus display a wide range of phenotypic diversity and possess significant ornamental value, with some species also having medicinal properties.

B. fimbristipula Hance (2n = 22) is indigenous to southeastern China, specifically in Zhejiang, Jiangxi, Hunan, Fujian, Guangdong, Guangxi, Hainan, and Hong Kong[3], with a subspecies endemic to Thailand (*B. fimbristipula* subsp. *siamensis* Phutthai & Radbouch.)[4]. This species typically has a solitary leaf and grows on rock or soil slopes in forest areas (Fig. 1). It holds considerable economic significance as both a medicinal and food source; for example, the essential oil derived from *B. fimbristipula* has shown inhibitory effects against *Streptococcus iniae* Pier in tilapia[5]. However, this species faces threats from climate change and human activities, particularly the local demand for its use in herbal tea. This species was designated as a protected wild species by the Guangdong Province, China, in 2023.

Despite the utilization of complete genomes for plant conservation over the past few decades, only four *Begonia* genomes have been published to date: *B. loranthoides* Hook.f., *B. masoniana* Irmsch. ex Ziesenh., *B. darthvaderiana* C.W.Lin & C.I Peng and *B. peltatifolia* Li[6]. Therefore, the generation of a high-quality genome of *B. fimbristipula* is essential for promoting its conservation and utilization, as well as for elucidating species relationships and evolutionary histories within this megadiverse genus.

In this study, we assembled and annotated the genome of *B. fimbristipula* using Oxford Nanopore Technology (ONT) reads, next-generation sequencing (NGS) reads, high-throughput chromosome conformation capture (Hi-C) reads, and RNA-seq reads. The assembled genome has a total size of 462.11 Mb and a scaffold N50 of

[1]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [2]South China National Botanical Garden, Guangzhou, 510650, China. [3]Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [4]Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [5]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. ✉e-mail: wzf@scbg.ac.cn; yanhaifei@scbg.ac.cn

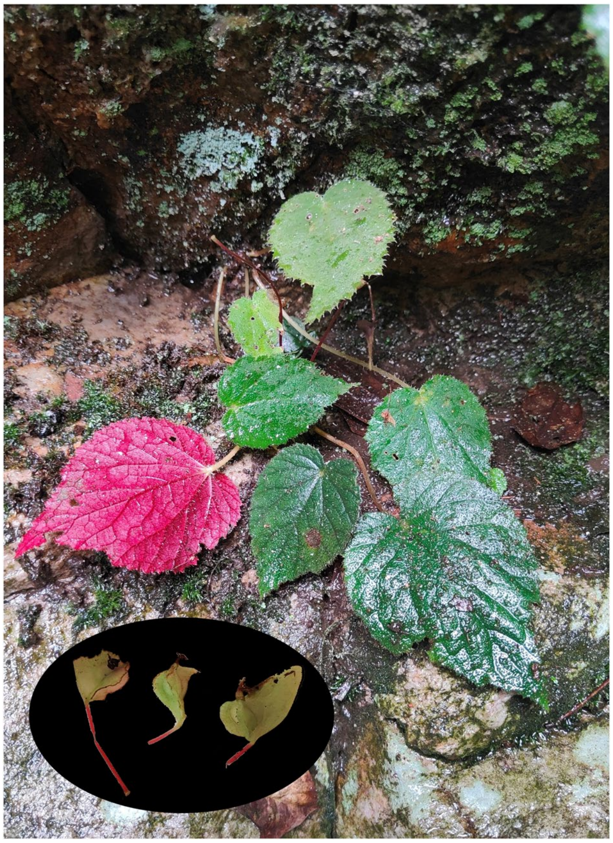

Fig. 1 Photographs of the plant and fruits.

Data	Number of bases (Gb)	Number of reads (M)	N50 length (bp)	Depth (×)	Sample
ONT	125.68	8.97	23,426	284	Leaf
NGS	143.04	953.59	150	325	Leaf
Hi-C	148.93	992.86	150	338	Leaf
RNA-seq	34.48	229.88	150	\	Leaf
RNA-seq	37.24	248.30	150	\	Fruit

Table 1. Summary of the sequencing data.

38.22 Mb. Annotation of repeat elements revealed that 64.63% (274.62 Mb) of the genome comprises repeat elements, with long terminal repeats (LTRs) accounting for 53.85% (146.60 Mb). Our analyses predicted a total of 25,563 protein-coding genes and 274 tRNAs. The high-quality genome of *B. fimbristipula* will advance our understanding of the evolutionary relationships within the genus *Begonia* and contribute to the conservation of this economically valuable species.

Methods

Sample collection and sequencing. Samples of *B. fimbristipula* were collected from Dinghu Mountain in Zhaoqing city, Guangdong Province, China (23°10′48″ N, 112°31′53″ E). Tissues, including fresh and young leaves and fruits, were immediately frozen in liquid nitrogen after collection and subsequently stored in refrigerator at −80 °C for DNA and RNA extraction. A voucher specimen (ID: gexj230012) has been deposited in the herbarium of the South China Botanical Garden, Chinese Academy of Sciences (IBSC).

Genomic DNA extraction and sequencing were performed according to the protocols described in our previous study[7]. Specifically, total DNA was extracted using Grandomics Genomic DNA Kit (GrandOmics Biosciences, Wuhan, China). DNA degradation was assessed via a 0.75% gel electrophoresis experiment, while DNA purity and concentration were evaluated using a NanoDrop One UV–Vis spectrophotometer and a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), respectively. For NGS sequencing, a short-read library (2 × 150 bp) with an insert size of 200–300 bp was prepared using the TruSeq Nano DNA HT Sample Preparation Kit, and sequencing was conducted on the Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA), generating 143.04 Gb (~325×) of raw data (Table 1).

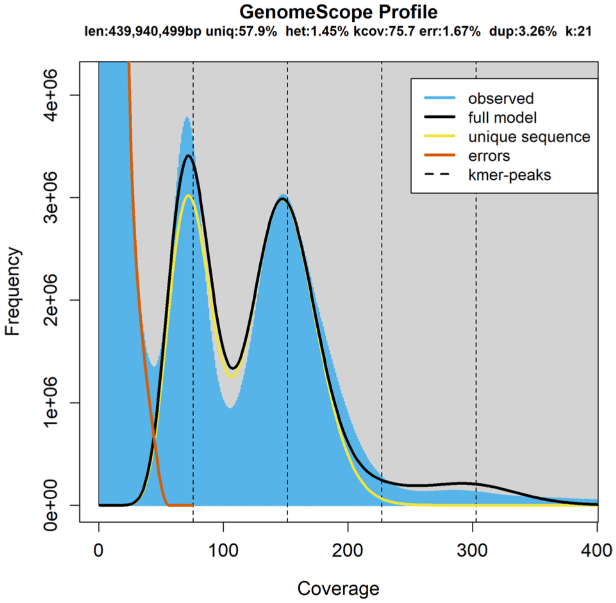

Fig. 2 Genome survey of *Begonia fimbristipula* based on 21-mer analysis.

For ONT long-read sequencing, the Nanopore library was prepared with the LSK109 Ligation Sequencing Kit, following the manufacturer's instructions, and sequenced with a Nanopore PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). This process yielded approximately 125.68 Gb (~284×) of ONT data with a mean read length of 14.84 kb and a read N50 length of 23.43 kb (Table 1). For Hi-C sequencing, the library was constructed from cross-linked DNA after digestion, biotinylation, ligation, enrichment, shearing, blunt-end repair, and additional steps. Sequencing of the final Hi-C library with pair-end read lengths of 150 bp was conducted on the Illumina HiSeq X Ten platform, which produced a total of 148.93 Gb (~338×) of Hi-C data (Table 1). For transcriptome sequencing, a pair-end RNA library (2 × 150 bp) was prepared according to the TruSeq RNA Library Preparation Kit instructions and sequenced on the Illumina HiSeq X Ten platform, yielding 34.48 Gb and 37.24 Gb of RNA-seq data for the leaf and fruit samples, respectively (Table 1). All sequencing was conducted at GrandOmics Co., Ltd (Wuhan, China).

Adapters in ONT data were trimmed using Porechop v0.2.4[8]. For NGS, Hi-C, and RNA-seq data, adapters and low-quality reads were removed using fastp v0.23.3[9].

Genome size estimation. The k-mer frequency distribution was calculated using Jellyfish v2.3.0 (-m 21)[10] with the NGS data. The resulting file was then utilized to predict the genome features using GenomeScope v1.0[11] with k-mer and read length set at 21 and 150, respectively. Based on this analysis, the genome of *B. fimbristipula* was estimated to be 439.94 Mb, with a heterozygosity rate of 1.45% (Fig. 2).

Chromosome-level genome assembly. The genome was assembled using NextDenovo v2.5.1[12], which has been shown to outperform other assemblers when applied to high repetitive and heterozygous genomes with ONT data[13]. The assembled haploid draft genome was 756.92 Mb, consisting of 980 contigs with a contig N50 length of 6.02 Mb. ONT reads were subsequently aligned to the draft genome using minimap2 v2.24-r1122[14]. The aligned bam file was processed with Purge Haplotigs v1.1.2 (-l 10 -m 110 -h 300)[15] to remove haplotypic duplications. After purging, 197 contigs were retained, totaling 460.91 Mb. The purged genome assembly was polished for two rounds with Racon v1.5.0[16] using ONT data, followed by two rounds of polishing with Polypolish v0.5.0[17] using NGS data. Hi-C reads were used to scaffold the polished genome assembly with Juicer v1.6[18] and 3d-dna v180922[19] (using the '-r 0' option), with manual adjustments made in Juicebox v1.11.08[20]. Gaps in the genome assembly were processed using TGS-GapCloser v1.1.1[21] with ONT long reads. Initially, 63 gaps were present; after gap-filling, only three remained on chromosomes 6, 7, and 8. The gap-filled genome was further polished in two additional rounds with Racon and Polypolish, as previously described.

The final genome assembly was 462.11 Mb in length, consisting of 67 scaffolds with a scaffold N50 length of 38.22 Mb. A total of 91.96% (424.94 Mb) of the sequences were anchored to 11 pseudochromosomes. The lengths of individual chromosomes ranged from 23.65 Mb (chr3) to 74.55 Mb (chr8) (Fig. 3a). The circular plot of chromosomes and Hi-C interaction heatmap were visualized by circos v0.69-9[22] and HiCExplorer v3[23], respectively. Tandem repeats were identified using Tandem Repeats Finder v4.09[24] in quarTeT v1.2.2[25]. The counts of tandem repeats varied from 11,339 on chr1 to 74,946 on chr8. GC content of the genome was calculated using bedtools v2.30.0[26], revealing that individual chromosomes had GC contents ranging from 37.39% (chr7) to 38.63% (chr3), with an overall mean of 38.00%.

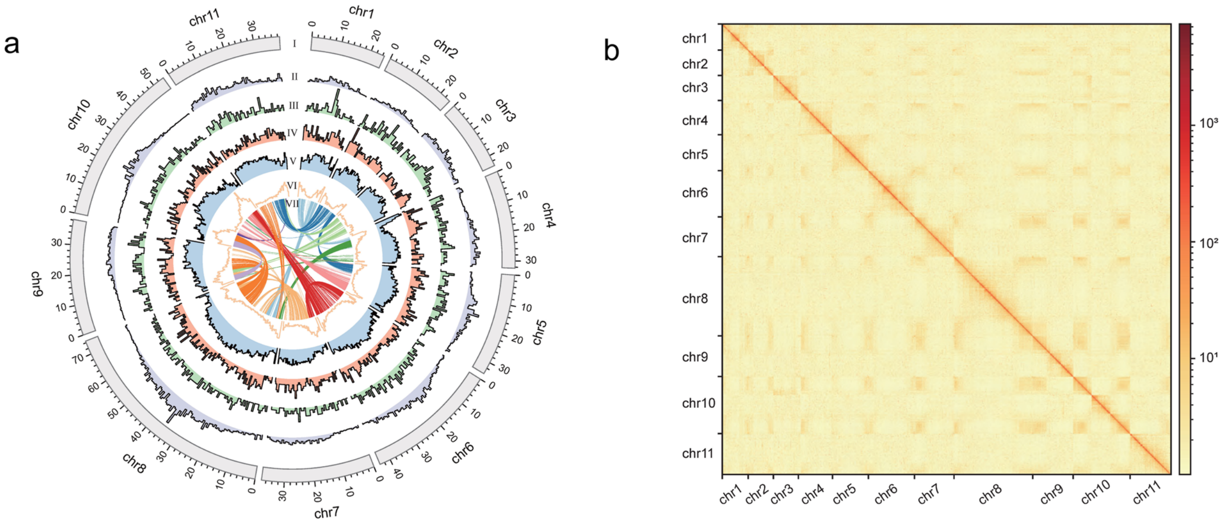

Fig. 3 Chromosome features of *Begonia fimbristipula*. (**a**) The tracks from outer to inner (I–VII) represent the chromosome, tandem repeat density, *Gypsy* density, *Copia* density, GC content, gene density, and sequence synteny within the genome, respectively (window size = 700 kb). (**b**) Hi-C interaction heatmap (bin size = 10 kb).

Class	Count	Masked (bp)	Masked (%)
LTR/*Copia*	84,247	91,350,832	21.50%
LTR/*Gypsy*	131,804	120,797,773	28.43%
LTR/unknown	33,400	16,668,911	3.92%
TIR/CACTA	9,209	2,965,375	0.70%
TIR/Mutator	39,646	13,302,388	3.13%
TIR/PIF_Harbinger	2,509	964,129	0.23%
TIR/Tc1_Mariner	2,398	606,144	0.14%
TIR/hAT	18,297	8,782,369	2.07%
non-LTR/LINE_element	173	61,602	0.01%
non-LTR/unknown	155	30,088	0.01%
non-TIR/helitron	14,055	5,505,831	1.30%
repeat_region	62,492	13,585,381	3.20%
Total	398,385	274,620,823	64.63%

Table 2. Summary of repeat classes identified by EDTA.

Repeat and gene annotation. The Extensive *de novo* TE Annotator (EDTA) v2.1.0[27] was employed to identify transposable elements. A total of 398,385 repetitive sequences were identified, representing 64.63% (274.62 Mb) of the genome (Table 2). Among these repeats, LTRs were the most prevalent, constituting 53.85% (146.60 Mb) of the genome (Table 2). *Gypsy* elements (28.43%) were the predominant LTRs, followed by *Copia* elements (21.50%). The total lengths of terminal inverted repeats (TIRs) and non-LTR elements were 5.32 Mb and 1.87 Mb, respectively. These elements accounted for 6.27% and 1.32% of the genome (Table 2).

The transposable element (TE) library generated by EDTA served as input for RepeatMasker v4.1.2[28] to produce a soft-masked genome. Gene prediction and functional annotation for the soft-masked genome were performed using *de novo*, homology protein-based, and transcriptome-based methods via the funannotate pipeline v1.8.15[29]. RNA-seq data were utilized to train the gene prediction models with the 'funannotate train' function. Subsequently, Augustus v3.5.0[30], GeneMark-ET v4.72[31], GlimmerHMM v3.0.1[32], and SNAP v2013-02-16[33] were employed for protein-coding gene prediction via the 'funannotate predict' function. At this stage, the protein-coding sequences of *B. loranthoides*[34], *B. masoniana*[35], *B. darthvaderiana*[36], and *B. peltatifolia*[37] were obtained from China National GeneBank DataBase as protein evidence. tRNAs were annotated using tRNAscan-SE v2.0.11[38]. Subsequently, the gene model predictions were refined and untranslated regions (UTRs) were incorporated using the 'funannotate update' feature. For functional annotation, the predicted genes were queried against public databases, including pfam v32.0, gene2product v1.45, interpro v76.0, dbCAN v8.0, busco_outgroups v1.0, merops v12.0, mibig v1.4, go v2023-05-10, repeats v1.0, unipot v2023_02, and eggNOG v5.0, using the InterProScan v5.62–94.0[39] and EggNOG-mapper v2.1.11[40] pipelines. The functional annotations were further processed using the 'funannotate annotate' feature. In total, 25,563 genes encoding 27,671 proteins were predicted, with an average gene length of 3,281 bp. Furthermore, 274 tRNAs were annotated. Among the protein-coding genes, 24,871 (97.29%) were functionally annotated by the eggNOG database, while 22,443 (87.79%) of the genes were identified by InterProScan database. Synteny blocks were identified using jcvi v1.3.8[41].

Data Records

The raw data, including ONT long reads, Illumina short reads, Hi-C reads, and RNA short reads, have been deposited in the Genome Sequence Archive in the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) with the accession number of CRA019543 under BioProject PRJCA031018[42]. The final genome assembly, annotation, and protein-coding sequences are accessible via Figshare[43]. Genome assembly has been submitted to the National Center for Biotechnology Information (NCBI) with the accession number of JBIQHB000000000 under BioProject PRJNA1173897[44].

Technical Validation

The completeness of the genome assembly was evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.3.2[45] with the embryophyta_odb10.2020-09-10 database. Of the core 1,614 conserved plant genes evaluated, the complete BUSCOs for *B. fimbristipula* were 90.3%, with 87.0% complete and single-copy BUSCOs, 3.3% complete and duplicated BUSCOs, 1.2% fragmented BUSCOs, and 8.5% missing BUSCOs. Additionally, the genome quality was also evaluated by calculating the LTR assembly index (LAI) using the LAI program[46], yielding a LAI value of 17.73. The base accuracy was assessed by Merqury v1.3[47] based on the NGS data, which demonstrated a k-mer-based QV of 24.61 and k-mer completeness of 61.41%. The Hi-C heatmap revealed that the 11 pseudochromosomes of *B. fimbristipula* exhibited strong interactive signals along the diagonals (Fig. 3b). ONT reads and RNA-seq reads were aligned to the final genome assembly using minimap2 v2.24-r1122[14] and and HISAT2[48], respectively. The mapping rate of ONT and RNA-seq reads were 89.30% and 87.16%, which were calculated using the 'stats' function in bamtools v2.5.1[49]. Moreover, the BUSCO completeness for genome annotation was assessed with BUSCO v5.3.2, yielding a value of 84.6%. Overall, these metrics indicate that the genome assembly of *B. fimbristipula* is of high quality and well-annotated.

Code availability

All software and pipelines utilized in this study were performed following the guidelines of the published tools. The parameters and version numbers of software and databases are detailed in the Methods section. Any elements not specified in the Methods were executed using default parameters. No custom scripts were employed.

Received: 8 November 2024; Accepted: 6 March 2025;
Published online: 12 March 2025

References

1. Moonlight, P. W. *et al*. Dividing and conquering the fastest–growing genus: Towards a natural sectional classification of the mega–diverse genus *Begonia* (Begoniaceae). *Taxon* **67**, 267–323, https://doi.org/10.12705/672.3 (2018).
2. Kubitzki, K. *The Families and Genera of Vascular Plants Vol. X, Flowering Plants - Eudicots - Sapindales, Cucurbitales, Myrtaceae*. (Springer, 2011).
3. Wu, Z.-Y., Raven, P. H. & Hong, D.-Y. *Flora of China*. (Science Press, 1999).
4. Radbouchoom, S., Phutthai, T. & Schneider, H. *Begonia fimbristipula* subsp. *siamensis* (sect. *Diploclinium*, Begoniaceae), a new taxon of the megadiverse genus endemic to Thailand. *PhytoKeys* **218**, 1–10, https://doi.org/10.3897/phytokeys.218.85699 (2023).
5. Yang, X. *et al*. Inhibitory activity of essential oil from *Begonia fimbristipula* against *Streptococcus iniae* in tilapia. *Jiangxi Fishery Science and Technology* **4**, 11–14 (2018).
6. Li, L. *et al*. Genomes shed light on the evolution of *Begonia*, a mega-diverse genus. *New Phytol.* **234**, 295–310, https://doi.org/10.1111/nph.17949 (2022).
7. Xiao, T.-W. *et al*. Chromosome-level genome assemblies of *Musa ornata* and *Musa velutina* provide insights into pericarp dehiscence and anthocyanin biosynthesis in banana. *Hortic. Res.* **11**, uhae079, https://doi.org/10.1093/hr/uhae079 (2024).
8. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132, https://doi.org/10.1099/mgen.0.000132 (2017).
9. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
10. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
11. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
12. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107, https://doi.org/10.1186/s13059-024-03252-4 (2024).
13. Sun, J., Li, R., Chen, C., Sigwart, J. D. & Kocot, K. M. Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Philos. Trans. R. Soc. B, Biol. Sci.* **376**, 20200160, https://doi.org/10.1098/rstb.2020.0160 (2021).
14. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).
15. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 460, https://doi.org/10.1186/s12859-018-2485-7 (2018).
16. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746, https://doi.org/10.1101/gr.214270.116 (2017).
17. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comp. Biol.* **18**, e1009802, https://doi.org/10.1371/journal.pcbi.1009802 (2022).

18. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
19. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
20. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
21. Xu, M. *et al*. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, giaa094, https://doi.org/10.1093/gigascience/giaa094 (2020).
22. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645, https://doi.org/10.1101/gr.092759.109 (2009).
23. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184, https://doi.org/10.1093/nar/gkaa220 (2020).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
25. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
26. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, https://doi.org/10.1093/bioinformatics/btq033 (2010).
27. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
28. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
29. *Zenodo* https://zenodo.org/records/4054262 (2020).
30. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57, https://doi.org/10.1002/cpbi.57 (2019).
31. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506, https://doi.org/10.1093/nar/gki937 (2005).
32. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
33. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
34. *China National GeneBank DataBase* https://db.cngb.org/search/assembly/CNA0013974 (2021).
35. *China National GeneBank DataBase* https://db.cngb.org/search/assembly/CNA0013975 (2021).
36. *China National GeneBank DataBase* https://db.cngb.org/search/assembly/CNA0013973 (2021).
37. *China National GeneBank DataBase* https://db.cngb.org/search/assembly/CNA0013976 (2021).
38. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
39. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
40. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122, https://doi.org/10.1093/molbev/msx148 (2017).
41. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211, https://doi.org/10.1002/imt2.211 (2024).
42. *National Genomics Data Center, China National Center for Bioinformation* https://ngdc.cncb.ac.cn/gsa/browse/CRA019543 (2024).
43. *Figshare* https://doi.org/10.6084/m9.figshare.27247158 (2024).
44. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBIQHB000000000 (2024).
45. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).
46. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126, https://doi.org/10.1093/nar/gky730 (2018).
47. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
48. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
49. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692, https://doi.org/10.1093/bioinformatics/btr174 (2011).

Acknowledgements

This work was supported by Science & Technology Fundamental Resources Investigation Program (Grant No. 2022FY100500).

Author contributions

H.F.Y. and Z.F.W. designed this study, collected the samples, and revised the manuscript. T.W.X. performed data analyses and wrote the manuscript. All authors approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.-F.W. or H.-F.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.