scientific data

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of *Wurfbainia villosa* (Zingiberaceae)

Yingmin Zhang[1,2,3], Congwei Yang[2,3], Hui Li[2] & Guodong Li[1,2] ✉

Wurfbainia villosa **is a well-known medicinal plant widely distributed in southern China and Southeast Asia. Its dried fruits, known as Fructus Amomi in traditional Chinese medicine, are commonly used for treating gastrointestinal disease. In this study, the *de novo* chromosome-level genome assembly was generated for this diploid plant. The assembled genome was approximately 2,731.09 Mb in length, with a contig N50 of 30.23 Mb. Approximately 99.91% of the assembled sequences were anchored onto 24 pseudochromosomes. A total of 42,713 genes were annotated, along with 88.83% of repetitive elements. These genomic resources provide valuable foundations for future research on active component biosynthesis, functional genomics, and DNA-informed breeding of *W. villosa*.**

Background & Summary

Wurfbainia villosa (Lour.) Škorničk. & A.D.Poulsen (2n = 48) is a perennial herb in the family Zingiberaceae, naturally distributed throughout Southeast Asia, especially in southern China[1,2] (Fig. 1A–C). Its dried fruit, known as Fructus Amomi (Sharen in traditional Chinese medicine), is widely used as both medicine and food in China. Fructus Amomi is one of the famous "Four Major Southern China Medicines" and is widely used clinically for warming the spleen, eliminating dampness and the prevention of miscarriage diseases[3]. Moreover, *W. villosa* is also widely used in food, liquors, and tea as a health product and condiment[4]. Recent studies have demonstrated various pharmacological properties of *W. villosa*, including gastroprotective, antihyperglycemic, antitumor, anti-inflammatory, and antioxidant effects[5,6]. Volatile oil is recognized as the primary bioactive component and serves as a quality-control marker in the Chinese Pharmacopoeia[7].

In recent years, the *W. villosa* industry has experienced substantial growth due to increasing demand for medicinal, edible, and health-related products, reaching an annual production value exceeding 100 billion yuan[8]. However, the industry faces several agricultural challenges. These include difficulties in natural pollination, poor capsule formation, and low seed germination rates, leading to significantly reduced yields[9]. Moreover, limited genetic diversity and extensive cultivation of genetically similar varieties have resulted in germplasm degradation and inconsistent product quality[10]. Therefore, breeding improved varieties of *W. villosa* is urgently required to address these challenges.

In this study, we assembled and annotated a high-quality genome of *W. villosa* using oxford nanopore technology (ONT), next-generation sequencing (NGS), and high-throughput chromosome conformation capture (Hi-C) methods. The assembled genome had a total length of 2,731.09 Mb with a contig N50 (a continuity metric) of 30.23 Mb. Repeat annotation indicated that repetitive elements account for 88.83% (2,426.10 Mb) of the genome, among which long terminal repeats (LTRs) comprised 78.38% (2,140.81 Mb). Additionally, our annotations predicted 42,713 protein-coding genes, 1,434 tRNAs, 219 miRNAs, 1,000 rRNAs, and 6,573 snRNAs. This genome assembly and annotation represent improvements in terms of completeness and continuity compared with previously reported genomes of *W. villosa*[11,12] (Table 1).

Methods

Sample collection and sequencing. Samples of *W. villosa* were collected on 24 November 2023 from Wenshan city, Yunnan Province, China (23°34' N, 104°45' E). A voucher specimen (ID: 532623SR024) was deposited in the herbarium of Yunnan University of Chinese Medicine. Immediately after collection, fresh young leaves were frozen in liquid nitrogen and stored at −80 °C prior to DNA extraction. High-quality genomic DNA was extracted from leaf tissues using the DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA).

[1]Yunnan Key Laboratory of Dai and Yi Medicines, Yunnan University of Chinese Medicine, Kunming, 650500, China. [2]College of Chinese Material Medica, Yunnan University of Chinese Medicine, Kunming, 650500, China. [3]These authors contributed equally: Yingmin Zhang, Congwei Yang. ✉e-mail: gammar116@163.com

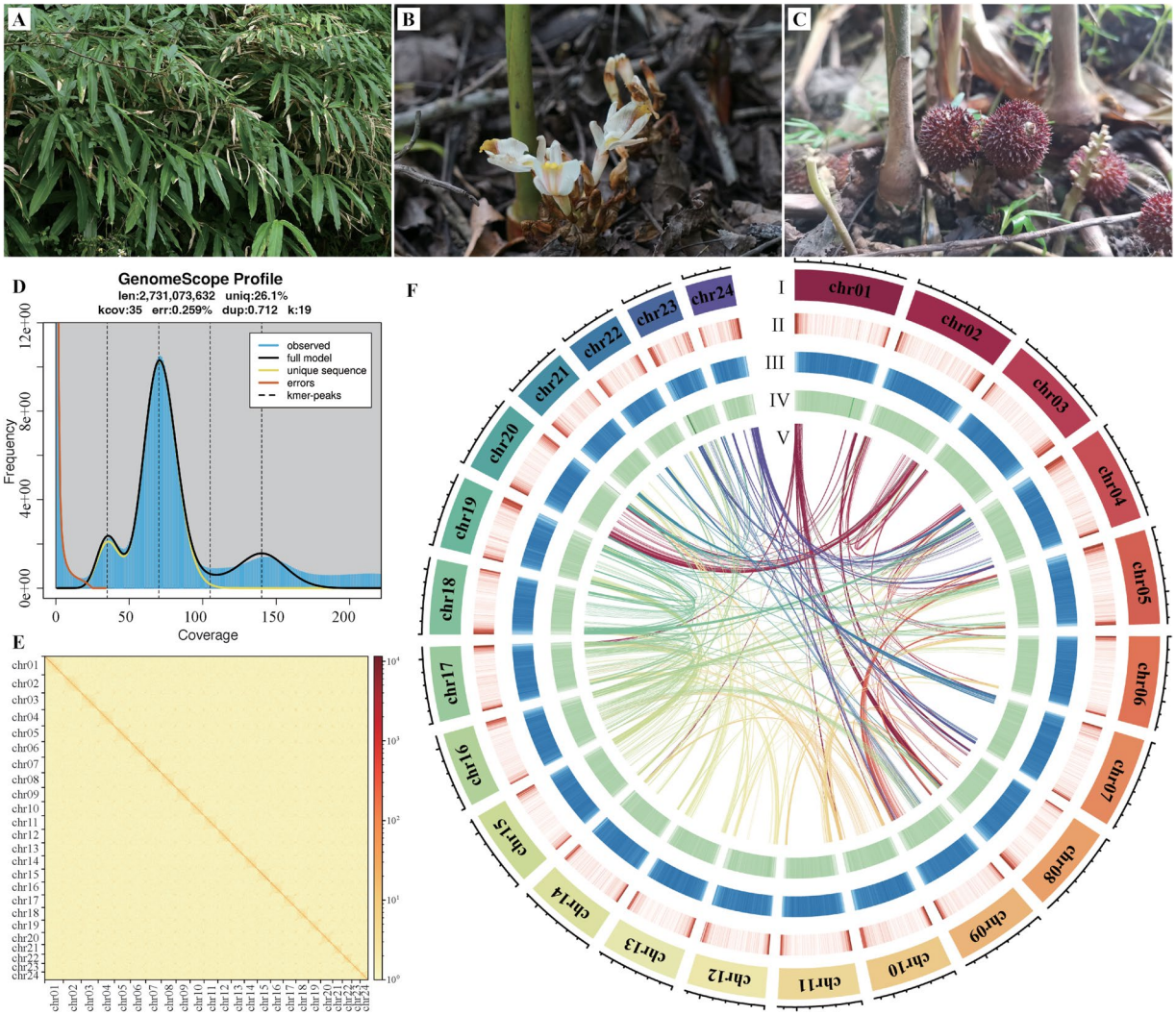

Fig. 1 Morphological characteristics and chromosome features of *Wurfbainia villosa*. (**A**) Plant of *W. villosa*; (**B**) Inflorescence of *W. villosa*; (**C**) Fruit of *W. villosa*; (**D**) Genome survey of *W. villosa* based on 19-mer analysis; (**E**) Hi-C interactive heatmap; (**F**) The tracks from outer to inner (**I**–**V**) represent the chromosome, gene density, repeat sequence density, GC content, and collinearity between the chromosomes, respectively (window size = 10 Mb).

For short-reads sequencing, paired-end libraries with an insert size of 350 bp were prepared using the NextEra DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA) and subsequently sequenced on the DNBSEQ-T7 platform (MGI Tech, Shenzhen, China). Quality control of short reads was performed using fastp v0.21.0[13] with default parameters to remove adapters, contaminants, PCR duplicates, and low-quality reads containing more than 30% low-quality bases. After filtering, a total of 229.60 Gb of clean short reads (84.12 × coverage) were obtained.

The Hi-C library was constructed according to the protocol described by Lieberman-Aiden[14] *et al.* and subsequently sequenced on an Illumina NovaSeq 6000 platform (paired-end, 150 bp reads). Raw Hi-C sequencing data were filtered and quality-controlled using fastp, resulting in 355.99 Gb of clean reads (130.42 × coverage).

For ONT sequencing, equal amounts of genomic DNA extracted from leaf samples were pooled to prepare the sequencing library using the Ligation Sequencing Kit (SQK-LSK109). The library was sequenced on a PromethION P48 sequencer (Oxford Nanopore Technologies, Oxford, UK), generating 294.86 Gb of clean long reads (108.02 × coverage).

Genome survey. The k-mer frequency distribution was calculated from clean short reads using Jellyfish v2.3.0[15] (count -C -m 19; histo -h 10000). Subsequently, genome heterozygosity was assessed using GenomeScope v1.0[16] (-k 19). Based on these analyses, the genome size of *W. villosa* was estimated to be 2,729.62 Mb, with a heterozygosity level of 0.63% (Fig. 1D).

Genome and chromosome assembly. ONT long reads were assembled to generate a draft genome using NextDenovo v2.5.1[17]. The initial genome assembly was subsequently polished twice with ONT data using Racon

	This study	Yang *et al.*[11]	Chen *et al.*[12]
Contig			
Number	164	1,110	2,901
Length (bp)	2,731,073,632	2,799,196,719	2,834,193,068
Minimum length (bp)	173,871	—	810
Maximum length (bp)	120,153,141	97,951,534	56,657,728
Average (bp)	16,652,888	—	976,971.07
N50 (bp)	30,227,502	9,134,317	8,295,749
N90 (bp)	8,501,241	1,251,635	1,378,773
Scaffold			
Number	34	309	—
Length (bp)	2,731,086,632	2,799,276,919	2834.2
Minimum length (bp)	173,871	—	—
Maximum length (bp)	155,807,329	139,364,153	151,371,763
Average length (bp)	80,326,077	—	—
N50 (bp)	116,484,485	109,948,433	112,800,000
N90 (bp)	82,385,971	75,970,413	68,159,171
Complete BUSCO (%)	95.1	97.9	99.8
Annotation			
Number of gene	42,713	42,588	42,473
Complete BUSCO (%)	97.6	95.2	—

Table 1. Comparison of the three *Wurfbainia villosa* genomes.

v1.5.0[18], followed by two additional polishing rounds using short-read data with Polypolish v0.5.0[18,19]. Haplotypic duplications were removed from the aligned bam file using Purge Haplotigs v1.1.2[20] (-l 10 -m 110 -h 300). After haplotig removal, the final haploid genome draft assembly was 2,731.07 Mb in length and comprised 164 contigs, with a contig N50 of 30.23 Mb.

Clean Hi-C reads were mapped to the draft genome using HiCUP v0.8.2[21] (-format sanger -longest 800 -shortest 150 -nofill N) to generate valid read pairs. Subsequently, invalid pairs, unmapped reads, and PCR duplicates were removed. Next, ALLHiC v0.9.8[22] (-e GATC -k 10) was used to cluster contigs into chromosome-level groups, followed by sorting and orientation. Interaction data between contigs were converted into binary files for visualization using 3D-DNA v180419[23] and Juicer v1.6[24]. Visual corrections to the assembly were finalized using JuiceBox v1.11.08[25], guided by observed chromosome interaction intensities. Contigs exhibiting no interactions were categorized as "unassigned," and gaps between contigs were filled with sequences of 100 Ns. Ultimately, 99.91% of assembled sequences were anchored onto 24 pseudo-chromosomes, with chromosome length raging from 66.82 Mb to 155.81 Mb (Fig. 1E, Table 2). The final genome assembly had a total length of 2,731.09 Mb and a scaffold N50 of 116.48 Mb (Table 1, Fig. 1F).

Repeat annotation. Repetitive elements were identified using the Extensive *de novo* TE Annotator (EDTA) v2.0.1[26] (parameters: -sensitive 1 -anno 1). In total, 2,426.10 Mb of repetitive sequences were predicted in the genome of *W. villosa* (Table 3). Among these repetitive sequences, long terminal repeat (LTR) retrotransposons constituted 78.38% (2,140.81 Mb). LTR/Copia elements represented the predominant type, covering 55.19% (1,507.37 Mb) of the genome, while LTR/Gypsy elements represented 23.19% (633.44 Mb).

Gene prediction and annotations. Genes were predicted by integrating three methods: *de novo* prediction, RNA-seq-based prediction, and homology-based approaches prediction. For *de novo* prediction, AUGUSTUS v3.2.3[27] and GlimmerHMM v3.0.4[28] were employed with default parameters. For RNA-seq-based prediction, full-length transcript sequences[29–32] were aligned to the reference genome using Minimap2 v2.17[33] (-ax map-ont -x splice -G 1000000). The resulting alignments were subsequently assembled into transcripts using StringTie v1.3.3[34], and coding sequences were predicted using TransDecoder v2.0 (http://transdecoder.github.io). For homology-based prediction, protein sequences from five reference species were utilized: *Wurfbainia villosa*[11,12], *Zingiber officinale*[35], *Musa acuminata*[36], *Oryza sativa*[37], and *Arabidopsis thaliana*[38]. Finally, we used MAKER v2.31.10[39] to combined *de novo* (AUGUSTUS), RNA-seq (Minimap2/StringTie), and homology-based predictions. In total, 42,713 genes were predicted, with an average mRNA length of 8,796.95 bp, an average coding-sequence length of 1,105.59 bp, an average exon length of 284.56 bp, and an average of 4.97 exons number per gene (Table 4).

Functions of protein-coding genes were annotated by comparing predicted proteins sequences against the NCBI Non-redundant protein (NR) database (ftp://ftp.ncbi.nih.gov/pub/nrdb/), Universal Protein Knowledgebase (UniProt) database[40], and Kyoto Encyclopedia of Genes and Genomes (KEGG) database[41]. Sequence comparisons were conducted using Diamond v2.0.11.149[42] (-evalue 1e-5). Additionally, protein motifs and functional domains were identified using InterProScan v5.52-86.0[43] by screening against multiple databases, including ProDom, PRINTS, Pfam, SMRT, PANTHER, and PROSITE. In total, 39,153 genes, representing 91.67% of the predicted protein-coding genes, were successfully annotated. Among these annotated

Chromosome	Length (bp)
chr01	155,807,329
chr02	152,277,517
chr03	141,532,943
chr04	136,752,181
chr05	134,269,119
chr06	130,719,375
chr07	129,765,143
chr08	125,081,870
chr09	120,153,141
chr10	117,077,882
chr11	116,484,485
chr12	112,146,917
chr13	111,159,626
chr14	110,714,412
chr15	110,262,413
chr16	107,835,141
chr17	106,574,171
chr18	106,393,393
chr19	103,109,231
chr20	102,006,460
chr21	82,385,971
chr22	78,476,267
chr23	70,745,984
chr24	66,820,366

Table 2. Length of pseudo-chromosome based on Hi-C data for *Wurfbainia villosa*.

Class	Superfamily	Length (bp)	Fraction of genome (%)
Class I	LTR/Copia	1,507,366,160	55.19
	LTR/Gypsy	633,439,338	23.19
	LINE	854,231	0.03
Class II	TIR/CACTA	40,781,423	1.49
	TIR/Mutator	63,714,455	2.33
	TIR/PIF-Harbinger	6,012,941	0.22
	TIR/Tc1-Mariner	1,976,824	0.07
	TIR/hAT	18,721,029	0.69
	Helitron	29,919,420	1.10
Unknown		123,318,167	4.52
Total		2,426,103,988	88.83

Table 3. Classification statistics of repeat elements.

genes, 81.54%, 86.14%, 22.26%, 88.00%, 26.94%, and 59.54% of these genes matched annotations in UniProt, NR, KEGG, InterPro, GO, and Pfam databases, respectively.

Transfer RNAs (tRNAs) were identified using tRNAscan-SE v2.0.7[44]. Other non-coding RNAs (ncRNAs), including microRNAs (miRNAs) and small nuclear RNAs (snRNAs), were predicted using Infernal v1.1.2[45] by screening against the Rfam database. In total, 1,000 rRNAs, 6,573 snRNAs, 219 miRNAs, and 1.434 tRNAs were predicted in the *W. villosa* genome.

Data Records

The NGS short reads, ONT long-reads, and Hi-C reads data have been deposited in the NCBI Sequence Read Archive with accession numbers SRR32962983[46], SRR33132822[47], and SRR33103089[48]. The genome assembly has been deposited in DDBJ/ENA/GenBank under the accession number JBNJRT000000000[49]. The chromosomal assembly and dataset of gene annotation have been deposited in the FigShare database[50].

Technical Validation

Assessment of the genome assembly. The integrity of the genome assembly was assessed by performing a sequence identity analysis. Short reads from a small-fragment sequencing library were selected and aligned to the assembled genome using BWA v0.7.17-r1188[51]. The alignment rate was approximately 99.67%, and a genome coverage reached approximately 95.56%, demonstrating high consistency between the sequencing reads and the assembled genome.

Feature	Number
Number of gene	42,713
Average of mRNA length (bp)	8,796.95
Average CDS length of per gene (bp)	1,105.59
Average exon number of per gene	4.97
Average of exon length (bp)	284.56
Average of intron length (bp)	1,857.49
Number of exons	212,227
Number of introns	169,514
the total intron length (bp)	314,871,345

Table 4. Statistical results for the genetic structure of *Wurfbainia villosa*.

Features	Assembly	Annotation
Complete BUSCOs	1,535 (95.1%)	1,575 (97.6%)
Complete and single-copy BUSCOs	1,454 (90.1%)	1,474 (91.3%)
Complete and duplicated BUSCOs	81 (5.0%)	101 (6.3%)
Fragmented BUSCOs	25 (1.5%)	13 (0.8%)
Missing BUSCOs	54 (3.4%)	26 (1.6%)

Table 5. Statistics of BUSCO evaluation of the assembly and annotation.

The completeness of the genome assembly was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.4[52] based on the embryophyta_odb10 dataset. The BUSCO analysis indicated that 95.10% of complete orthologous genes were present in the assembly (Table 5). Additionally, 99.91% of assembled sequences were successfully onto 24 chromosomes, demonstrating a high level of chromosome-level continuity. To further validate chromosome assembly accuracy, Hi-C reads were mapped to the final genome assembly using Juicer v1.6 and visualized using Juicebox v1.11.08. The resulting Hi-C interaction heatmap clearly displayed chromosome clustering, with significantly stronger interactions along diagonal positions than off-diagonal positions, indicating frequent interactions between adjacent sequences (Fig. 1E). This chromosomal clustering pattern aligns with established principles of Hi-C-assisted genome assembly, confirming effective genome anchoring and high assembly accuracy.

Assessment of the gene annotation. Gene annotations were validated using BUSCO analysis, which indicated high annotation quality, with approximately 97.60% of complete BUSCO genes successfully detected (Table 5). Moreover, the number of predicted genes and structural characteristics of the annotated *W. villosa* genome were consistent with those of genomes closely related species, further confirming the reliability of the annotation results.

Code availability

All software and pipelines were executed according to the manual and protocols of the published bioinformatics tools. The version and code/parameters of the software have been detailed and described in Methods. No custom code was used during the compilation of the dataset.

Received: 12 June 2025; Accepted: 19 August 2025;
Published online: 26 August 2025

References

1. Chen, Z. Y. *et al.* Preliminary report of chromosome numbers on Chinese Zingiberaceae. *Guihaia* **2**, 153–157 (1982).
2. Boer, H. *et al.* Convergent morphology in Alpinieae (Zingiberaceae): Recircumscribing *Amomum* as a monophyletic genus. *Taxon* **67**(1), 6–36 (2018).
3. National Pharmacopoeia Commission of China. Pharmacopoeia of the People's Republic of China (China Medical Science Press, 2020).
4. Gong, L. *et al.* DNA barcode reference library construction and genetic diversity and structure analysis of *Wurfbainia villosa* Lour. (Zingiberaceae) populations in Guangdong Province. *PeerJ* **9**, e12325 (2021).
5. Liu, H. F. *et al.* Effects of a polysaccharide extract from *Amomum villosum* Lour. on gastric mucosal injury and its potential underlying mechanism. *Carbohydrate Polymers* **294**, 119822 (2022).
6. Zeng, Y. W. *et al.* Chemical constituents of functional food *Wurfbainia villosa* to combat human diseases. *Current Chinese Science* **2**(1), 57–67 (2022).
7. Chen, L. *et al.* Comparison of volatile compounds in different parts of fresh *Wurfbainia villosa* Lour. from different geographical areas using cryogenic grinding combined HS–SPME–GC–MS. *Chinese Medicine* **15**(1), 97 (2020).
8. Yin, C. Y. *et al.* Analysis of the Existing Standard and Construction of Standard System for Amomi Fructus Industry. *Journal of Chinese Medicinal Materials* **47**(01), 1–6 (2024).
9. Zhao, H. Y. *et al.* Research progress on cultivation of *Wurfbainia villosa* Lour.: original plant of the southern famous medicinal fructus Amomi Villosi. *World Chinese Medicine* **17**(08), 1163–1170 (2022).
10. Xu, J. *et al.* Research on germplasm diversity of *Wurfbainia villosa* Lour in genuine producing area. *PLoS One* **17**(8), e0268246 (2022).

11. Yang, P. *et al*. Chromosome-level genome assembly and functional characterization of terpene synthases provide insights into the volatile terpenoid biosynthesis of *Wurfbainia villosa*. *The Plant Journal* **112**(3), 630–645 (2022).
12. Chen, X. *et al*. Multiomics comparison among populations of three plant sources of Amomi Fructus. *Horticulture Research* **10**(8), uhad128 (2023).
13. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
14. Lieberman-Aiden, E. *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
16. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
17. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
18. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
19. Wick, R. R. & Holt, K. E. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comp. Biol.* **18**, e1009802 (2022).
20. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 460 (2018).
21. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
22. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
23. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
25. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
26. Ou, S. J. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
27. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
28. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
29. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR18576716 (2022).
30. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR18576717 (2022).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR18576729 (2022).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR18576727 (2022).
33. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
34. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
35. Chen, Z. J. *et al*. A genome assembly of ginger (*Zingiber officinale* Roscoe) provides insights into genome evolution and 6-gingerol biosynthesis. *The Plant Journal* **118**(3), 682–695 (2024).
36. D'hont, A. *et al*. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**(7410), 213–217 (2012).
37. *Oryza sativa* Indica Group cultivar 93-11, whole genome shotgun sequencing project. *GenBank GenBank* https://identifiers.org/ncbi/insdc:AAAA00000000 (2014).
38. Swarbreck, D. *et al*. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research* **36**, D1009–D1014 (2007).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* **12**, 1–14 (2011).
40. Apweiler, R. *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
41. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
43. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
44. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
45. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32962983 (2025).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33132822 (2025).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33103089 (2025).
49. *Wurfbainia villosa* isolate Wenshan, whole genome shotgun sequencing project. GenBank https://identifiers.org/ncbi/insdc:JBNJRT000000000 (2025).
50. Zhang, Y. M. Genome annotations of *Wurfbainia villosa*. *Figshare* https://doi.org/10.6084/m9.figshare.28746839 (2025).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82260739), the Yunnan Provincial Science and Technology Department – Applied Basic Research Joint Special Funds of Yunnan University of Traditional Chinese Medicine (202101AZ070001-005), and the Yunnan Revitalization Talent Support Program "Young Talent" Project (YNWR-QNBJ-2020-278).

Author contributions

G.L. designed the study. Y.Z., C.Y. and H.L. performed the experiments and analyzed the data. Y.Z. and C.Y. wrote the paper. G.L. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.L.

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.