scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale genome assembly and annotation of *Garuga floribunda* var. *gamblei* (King ex W. W. Sm.) Kalkman

Rong Chen[1,2,3,4,5], Rui Rao[1,2,3,4,5] & Liang-Liang Yue[1,2,3,4] ✉

Garuga floribunda **var. *gamblei* (King ex W. W. Sm.) Kalkman, a tree species of Burseraceae (Sapindales), is rare in China and prized for its high-quality timber. Despite its economic and ecological significance, genetic architecture remains poorly understood, underscoring the necessity for comprehensive genomic studies. In this study, we present a high-quality, chromosome-level genome assembly of G. *floribunda* var. *gamblei*, constructed using a combination of PacBio HiFi long reads, Illumina short reads, and Hi-C sequencing technologies. The assembled genome spans approximately 448.57 Mb with a scaffold N50 of 30.54 Mb, and 13 pseudochromosomes were successfully constructed. BUSCO analysis indicated a high level of completeness, with 97.7% of conserved genes detected. Repeat annotation identified 160.78 Mb of repetitive sequences, accounting for 35.84% of the genome. The genome annotation predicted 19,620 protein-coding genes and 14,552 non-coding RNAs. This high-quality genomic resource lays a solid foundation for future comparative genomic studies and provides valuable insights into the evolutionary history and genetic diversity of Burseraceae and related lineages.**

Background & Summary

Garuga floribunda var. *gamblei* (King ex W. W. Sm.) Kalkman, first published in 1953[1], is a tree species mainly distributed across the lowland tropical area of Southern Asia, such as Myanmar, east India and southern China[1–4]. The species is currently threatened by the loss of tropical forestry within its native range. Particularly in China, most of the collection locations from herbarium specimens are lost to land use changes, such as agricultural, plantation forestry, and urban development[5], making it touching the scope of Species with Extremely Small Population (SESP)[6,7]. Despite its rarity, *G. floribunda* var. *gamblei* has historically been utilized as a source of "Red Wood" in China and other East Asian countries. Its bright yellow flowers make it a valuable ornamental tree for horticulture and urban landscaping (Fig. 1a,b). We were once captivated by its striking bright yellow canopy standing prominently on a steep karst slope along the highway, an image that sparked our interest in the conservation biology of this species. To date, literature and genetic information concerning the genome, phylogeny, and population genetics of this species remain extremely limited, with only a preliminary evolutionary analysis involving three *Garuga* species previously conducted by our team[8]. Consequently, the acquisition of high-quality reference genomes is essential for enhancing the utilization of resources in *G. floribunda* var. *gamblei* and for investigating the phylogeny of Burseraceae plants.

In this study, we assembled and annotated the genome of *G. floribunda* var. *gamblei* using an integrated approach that combined PacBio HiFi long-read sequencing, Illumina short-read sequencing, high-throughput chromosome conformation capture (Hi-C), and RNA-seq data. The assemblies span approximately 448.57 Mb with a scaffold N50 of 30.54 Mb. A total of 95.40% of the assemblies were anchored to 13 pseudo-chromosomes. Repeat annotation revealed that repetitive elements account for 35.84% of the genome. Circular consensus sequencing (CCS) reads were mapped back to the genome, achieving a 93.85% alignment rate across 1,204,045

[1]Yunnan Key Laboratory of Plateau Wetland Conservation, Restoration and Ecological Services, Southwest Forestry University, Kunming, China. [2]National Plateau Wetlands Research Center, Southwest Forestry University, Kunming, China. [3]National Wetland Ecosystem Fixed Research Station of Yunnan Dianchi, Southwest Forestry University, Kunming, China. [4]Dianchi Lake Ecosystem Observation and Research Station of Yunnan Province, Kunming, China. [5]These authors contributed equally: Rong Chen, Rui Rao. ✉e-mail: yueliangliang@swfu.edu.cn

Fig. 1 *G. floribunda* var. *gamblei*: (**a**) Fruit; (**b**) Flower (by L.L.Yue).

reads, with an average depth of 27.7 × and a coverage of 99.94%. Reference-free quality assessment using Merqury produced a QV score of 60.41, corresponding to an error rate of only 9.1×10^{-7}, and indicated a k-mer completeness of 85.2%. This difference between high QV and moderate k-mer completeness reflects the absence of low-frequency or sample-specific k-mers, rather than substantial gaps, confirming that the assembly is highly accurate at the base level and largely complete. Genome completeness was further evaluated using BUSCO (v5.4.7) with the embryophyta_odb10 dataset (1,614 conserved genes), revealing 97.7% complete BUSCOs (96.0% single-copy and 1.7% duplicated), with only 0.6% fragmented and 1.7% missing. Genome annotation identified 19,620 protein-coding genes and 14,552 non-coding RNAs. This high-quality genome provides a valuable reference for the genus *Garuga* and the Burseraceae family, offering important resources for comparative genomics and advancing our understanding of evolutionary relationships and functional genomics among closely related species.

Methods

Sample collection. Fresh leaves of *G.floribunda* var. *gamblei* were collected from Wild Elephant Valley, Xishuangbanna, Yunnan Province, China (22°11′0.01″N, 100°51′39.27″E). Young leaves, fruits, and floral buds were collected from the same individual for subsequent transcriptome sequencing. All samples were immediately frozen in liquid nitrogen after collection and stored at −80 °C until nucleic acid extraction. DNA and RNA extraction, as well as sequencing, were carried out by Shanghai Personal Biotechnology Co., Ltd.

DNA/RNA extraction and genome sequencing. Genomic DNA was extracted using an improved CTAB method[9]. For PacBio HiFi sequencing, high-quality DNA libraries were prepared using the SMRTbell® Prep Kit 2.0 (Pacific Biosciences) following the manufacturer's instructions and sequenced on the PacBio Sequel II platform. For Hi-C sequencing, genomic DNA was purified using the QIAamp DNA Mini Kit (Qiagen, CAT#51306), and Hi-C libraries were constructed and sequenced on the BGI MGISEQ-2000 platform. Transcriptome sequencing was performed by extracting total RNA with TRIzol reagent (Thermo Fisher Scientific, MA, USA) following the manufacturer's protocol. A 150 bp paired-end RNA-seq library was constructed and sequenced on the Illumina NovaSeq 6000 platform. Additionally, short-read genomic libraries were prepared with the MGIEasy Universal DNA Library Prep Set (BGI) and assessed using the Qubit™ dsDNA BR Assay Kit and Qubit® ssDNA Assay Kit (Invitrogen), then sequenced on the BGI MGISEQ-2000 platform to generate 150 bp paired-end reads (PE150). All genomic and Hi-C sequencing data were obtained from the same individual plant. The RNA-seq data were used for gene structure annotation.

Sequencing data and quality assessment. PacBio HiFi sequencing generated approximately 15.50 Gb of high-fidelity long reads, with > 99.39% of bases reaching Q7. The average read length was 16,473 bp, with a maximum read length of 61,149 bp, and a GC content of 35.41%, providing ~34 × genome coverage. Illumina short-read sequencing, after quality filtering using fastp[10], produced 25.18 Gb of high-quality data, with Q20 and Q30 values of 98.79% and 96.25%, respectively. The average read length was 149 bp, and the GC content was 34.35%. These short reads were used for error correction and polishing of the genome assembly. Hi-C sequencing generated approximately 52.18 Gb of raw data (349 million paired-end reads). After filtering, 51.13 Gb of clean reads (342 million read pairs) were retained, with Q20 and Q30 scores of 98.85% and 96.46%, respectively, an average read length of 149 bp, and a GC content of 37.5%, contributing ~120 × coverage for chromosome-scale scaffolding. RNA-seq libraries were constructed from total RNA extracted with TRIzol and sequenced on the Illumina NovaSeq 6000 platform, generating 11 Gb of clean data (150 bp paired-end reads). These RNA-seq reads were used for gene structure annotation and transcript-supported functional characterization (Table 1).

Genome survey. The congeneric species *Garuga pinnata* possesses a diploid chromosome number of 2n = 26[11]. *G. floribunda* var. *gamblei* is likewise diploid. In this study, by employing Jellyfish v2.3.0[12] and GenomeScope2[13] with 21-kmer to estimate the genome size, heterozygosity, and repeat content based on Illumina

Library types	Platform	Bases (Gb)	Mean length(bp)	Sequence coverage(X)
PacBio	PacBio Sequel IIe	15.5	16473	30
Illumina	Illumina NovaSeq 6000	25.18	149	50
Hi-C	Illumina NovaSeq 6000	52.18	150	120
RNA-seq	Illumina NovaSeq 6000	11	150	—

Table 1. Data for Genome Sequencing.

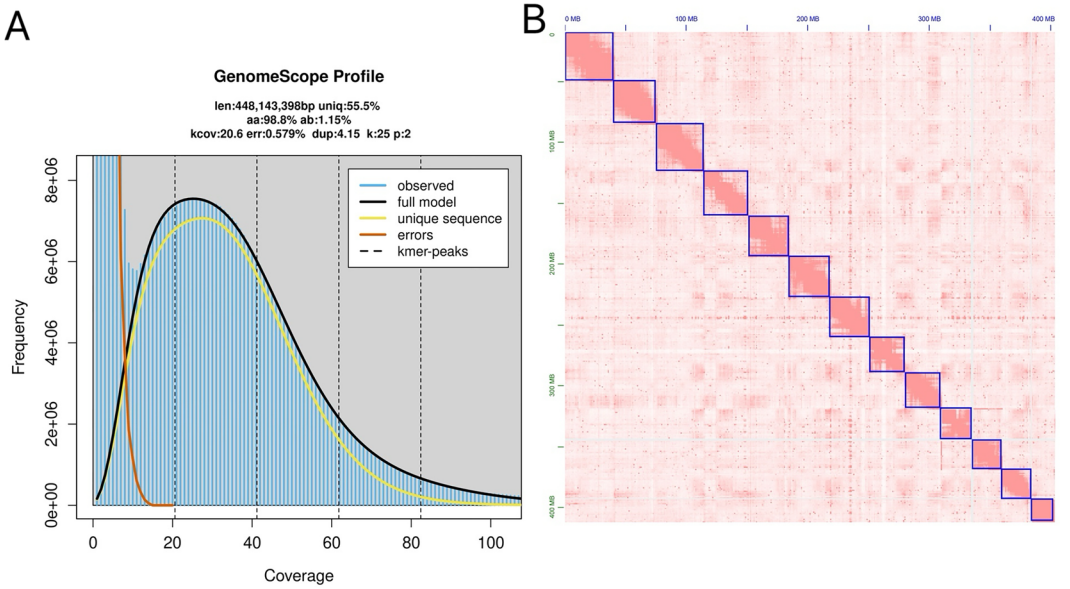

Fig. 2 Chromosome-Level Genome Assembly: (**A**) K-mer analysis of *G. floribunda* var. *gamblei* genome using GenomeScope profle; (**B**) Pore-C interaction Heatmap illustrating chromosomal interactions in the *G. floribunda* var. *gamblei* genome.

short reads, the genome of *G. floribunda* var. *gamblei* was estimated to be 435–448 Mb in size, with a repeat content of 44.5% and a high level of homozygosity, as 98.8% of the genome consists of homozygous regions (Fig. 2A).

Genome assembly. PacBio HiFi reads were assembled using Hifiasm v0.19.5[14] with parameter (−l = 2, -n = 4). Detailed assembly statistics are summarized in Table 2, and the resulting assembly graph was converted into contig sequences in FASTA format using Gfastats[15], while the redundant sequences were removed using purge_dups v1.2.5[16] with default parameters. The draft genome of *G. floribunda* var. *gamblei* was 448.57 Mb in size, with a contig N50 of 34 Mb (Table 3). Using Hi-C data generated by sequencing, the assembled contigs were anchored to the near-chromosome level with ALLHIC v0.9.8 software[17] (enz = DpnII, CLUSTER = n), including chromosome clustering, orientation, and sorting. Manual inspection and adjustment were performed using Juicebox v1.11.08 software[18] (pre -n -q 0 or 1), primarily focusing on refining chromosome segment boundaries and correcting assembly errors. Finally, the genome was obtained at the chromosome level, the size was 448.57 Mb, with 95.40% of sequences anchored to 13 pseudo-chromosomes and a scaffold N50 size of 30.54 Mb (Fig. 2B).

Genome annotation. Genome annotation mainly includes repeat sequence annotation, gene structure annotation, gene function annotation, and non-coding RNA annotation. We employed a combination of de novo and homology-based approaches to detect repeat elements within the genome of *G. floribunda* var. *gamblei*. For de novo prediction, LTR_FINDER[19] was used to identify repeat elements, followed by the construction of a non-redundant repeat library using RepeatMasker[20] and its associated script Repeat-ProteinMask, with default parameters, to extract repeat regions. For the homology-based approach, TRF was used to detect tandem repeat, RepeatMasker[20] was employed to search for repeat elements against the RepBase database[21,22]. A total of 324,237 repeat sequences were identified, with a cumulative length of 160,788,694 bp, accounting for 35.84% of the genome (Table 4).

The protein-coding genes were annotated by incorporating transcriptional evidence, homology-based, and *ab initio* methods. For transcriptional evidence, fastp[10] was used to trim the data separately for different tissues, and the clean data were aligned to the assembled genome of *G. floribunda* var. *gamblei* by using HISAT2 v2.2.1[23]. The transcripts were assembled using Stringtie v2.2.1[24] following default parameters, and the transcripts from all samples were merged and subjected to protein-coding sequence prediction and quality filtering via TransDecoder v.5.1.0 (https://github.com/TransDecoder/TransDecoder/wiki), which is available in PASA

Assembly characteristics	Values
Total genome size (bp)	448,571,095
Number of contigs	375
Contig N50 (bp)	31,853,271
Contig N90 (bp)	6,550,000
Number of scaffolds	168
Scaffold N50 (bp)	30,542,712
Scaffold N90 (bp)	6,109,652
Number of pseudochromosomes	13
Chromosome length (bp)	427,838,147
Anchored rate of bases to the pseudochromosomes (%)	95.40%
GC content (%)	33.58
Number of annotated genes	19,620
Genome BUSCO	C:97.7%[S:96.0%,D:1.7%],F:0.6%,M:1.7%,n:1614

Table 2. Genome assembly statistics.

Pseudo-chromosome ID	Length (bp)	GC content(%)
Chr1	40,380,722	32.86%
Chr2	40,371,135	33.06%
Chr3	38,559,720	33.79%
Chr4	36,800,044	33.46%
Chr5	36,388,411	34.33%
Chr6	34,122,100	33.56%
Chr7	33,858,100	34.11%
Chr8	32,470,360	33.67%
Chr9	31,472,155	34.02%
Chr10	29,810,812	33.05%
Chr11	29,104,688	33.79%
Chr12	24,426,290	33.73%
Chr13	20,073,610	32.93%
Unplaced	20,732,948	38.49%

Table 3. Summary of the structure of 13 pseudochromosomes.

Type		Count	bpMasked	%masked
DNA		42,721	20,849,984	4.65%
LINE		6,088	2,607,192	0.58%
LTR	Gypsy	47,139	26,597,548	5.93%
	Copia	61,138	51,518,226	11.48%
	Others	25,959	9,850,302	2.20%
	Total LTR	1,34,236	87,966,076	19.61%
SINE		436	270,998	0.06%
Simple_repeat		2,325	117,912	0.03%
Satellite		0	0	0.00%
Unknown		1,44,589	45,948,938	10.24%
Total		3,58,934	160,788,694	35.84%

Table 4. Repetitive elements in the genome assembly.

v2.4.1[25]. Only complete transcripts were retained for subsequent analysis. Sequences of homologous proteins were downloaded from Ensembl (https://useast.ensembl.org/index.html) and NCBI (including *Boswellia sacra, Bursera cuneata, Bursera palmeri, Bursera bipinnata, Mangifera indica, Pistacia atlantica, Pistacia vera, Arabidopsis thaliana*[26]). Protein sequences were aligned to the genome using TblastN v2.2.26[27], and then the matching proteins were aligned to the homologous genome sequences for accurate spliced alignments with GeneWise[28] software which was used to predict gene structure contained in each protein region. For gene predication based on Ab initio, Augustus v3.5[29] and SNAP v2013.11.29[30] were used in our automated gene prediction pipeline. Transcriptome reads assemblies were generated with Trinity v2.8.5[31] for the genome annotation. To

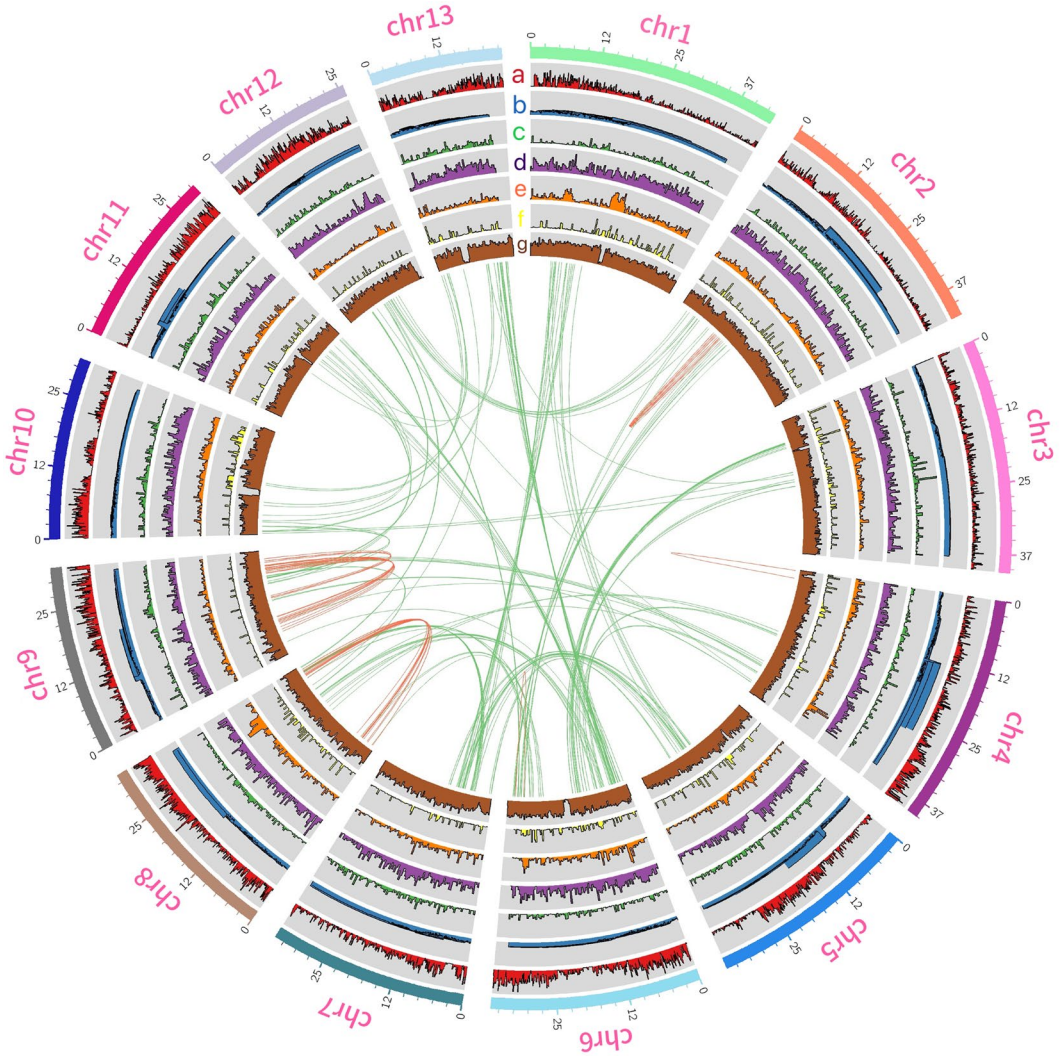

Fig. 3 The circos diagram of *G. floribunda* var. *gamblei* genome. Circles (a–g) represent 13 pseudochromosomes of the *G. floribunda* var. *gamblei* pseudochromosomes: (a) Gene density, (b) GC content, (c) Copia LTR element density, (d) Gypsy LTR element density, (e) EnSpm element density, (f) Tandem repeat density, and (g) Intraspecies collinearity links between syntenic genes. Red ribbons represent intra-chromosomal synteny, while green ribbons indicate inter-chromosomal synteny.

optimize the genome annotation, the RNA-Seq reads from different tissues which were aligned to genome fasta using Hisat v2.2.1[32] with default parameters to identify exons region and splice positions. The alignment results were then used as input for Stringtie v2.2.1[33] with default parameters for genome-based transcript assembly. All predicted gene structures were integrated into a nonredundant gene set by using EVidenceModeler (EVM) v1.1.1[25]. The weight values for high-quality RNA-seq transcripts, homologous proteins, and *ab initio* prediction were set to 10, 7, and 3. The EVM-predicted genes were further corrected by using PASA v2.4.1[25] to identify the untranslated and alternative splicing regions. A total of 19,620 protein-coding genes were predicted (Fig. 3).

Functional annotation of the protein-coding genes was performed based on sequence similarity and the identification of conserved domains. For sequence similarity, the protein-coding genes were matched against the Universal Protein Knowledgebase database[34], the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[35], and the eggNOG database[36] using diamond v2.0.11[37] with a specified E-value cut-off of 1e-5. For the identification of conserved domains, InterProScan v5.52[38] was employed to detect and classify domains and motifs by referring to the Pfam[39], SMART[40], PANTHER[41], PRINTS[42], and ProDom[43] databases. With both approaches, a total of 19,038 protein-coding genes were functionally annotated (Fig. 4).

The non-coding RNAs (ncRNAs) were predicted using a combination of bioinformatic tools. Ribosomal RNAs (rRNAs) were identified using Barrnap v0.9[44], while transfer RNAs (tRNAs) were predicted with tRNAscan-SE v2.0.12[45]. MicroRNAs (miRNAs) and small nuclear RNAs (snRNAs) were detected using the *cmscan* module in Infernal v1.1.4[46], based on the Rfam v14.9[47] database. All predictions were performed using default parameters. In total, 14,552 non-coding RNAs were annotated, including 3,843 rRNAs (partial or complete 5.8S, 18S, and 28S), 2,506 tRNAs, 162 miRNAs, and 3,800 snRNAs (Table 5). It is worth noting that the

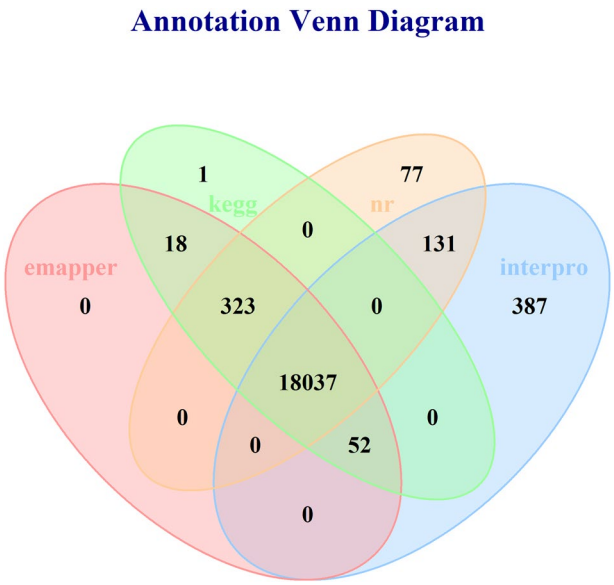

Fig. 4 Functional annotation summary.

Type		Copy number	Average length (bp)	Total length (bp)
miRNA		162	124.41	20,154
tRNA		2,506	76	120,064
rRNA	18S	1,322	1,424.41	1,883,064
	28S	1,300	2,859.78	3,717,720
	5.8S	1,052	149.25	157,010
	5S	1,931	109.98	212,374
	total rRNA	3,843	1,369.67	5,263,628
snRNA	CD-box	3,639	106.37	387,098
	HACA-box	53	127.58	6,762
	splicing	108	149.85	16,184
	total snRNA	3,800	107.91	410,044
other ncRNAs		4,241	84.09	811,302

Table 5. Statistics of ncRNA annotation.

high number of rRNAs and snRNAs may include fragmented or redundant annotations, which could be further refined in downstream analysis.

Data Records

The raw sequencing data generated in this study, including Illumina short reads, PacBio HiFi long reads, Hi-C sequencing data, and multi-tissue RNA-seq datasets, have been deposited in the Genome Sequence Archive (GSA) in the National Genomics Data Center (NGDC)[48,49], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under the accession number CRA033719[50], which can be found with the following GSA run IDs: CRR2340812[51], CRR2340813[52], CRR2340814[53], and CRR2340815[54]. The chromosome-scale genome assembly has been deposited in an INSDC repository (DDBJ/ENA/GenBank) under the accession number JBUERZ000000000[55] and Genome Warehouse (GWH) of NGDC under accession number GWHHCKV00000000.1[56], which is publicly accessible online. Additionally, the genome annotation data have been archived in the Figshare repository[57].

Technical Validation

To assess the assembly accuracy of *G. floribunda* var. *gamblei*, a comprehensive evaluation was conducted across multiple dimensions. HiFi reads were remapped to the assembled genome using Minimap2 v2.26[58], achieving a coverage of 99.94%. Merqury v1.3[59] was used for k-mer[60] quality evaluation. The consensus base-level accuracy reached a QV of 60.4 (corresponding to an estimated error rate of 9.1×10^{-7}), and the k-mer completeness was 85.2%. The slightly lower k-mer completeness reflects low-frequency or sample-specific k-mers (e.g., from heterozygosity, repetitive regions, or sequencing noise) that are not expected to be present in a cleaned nuclear assembly, while most high-confidence k-mers from the raw reads were retained. Long terminal repeat (LTR) retrotransposons accounted for 33.3% of the genome, with Gypsy and Copia elements comprising 16.6%

Category	Proteome	Protein Set
Total BUSCOs searched (n)	1,614	1,614
Complete BUSCOs (C)	82.0% (1,324)	88.8% (1,433)
Single-copy (S)	79.7% (1,287)	61.0% (984)
Duplicated (D)	2.3% (37)	27.8% (449)
Fragmented BUSCOs (F)	9.5% (153)	3.9% (63)
Missing BUSCOs (M)	8.5% (137)	7.3% (118)

Table 6. BUSCO assessment of annotated proteomes.

and 13.2%, respectively. The widespread distribution of LTR elements across scaffolds indicated well-preserved repeat structures and high assembly integrity. BUSCO v5.4.7[61] analysis against the embryophyta_odb10 dataset identified 97.7% complete genes (96.0% single-copy, 1.7% duplicated), confirming a near-complete gene space.

To assess annotation quality, BUSCO was also applied to two predicted protein sets. The EVM-integrated set showed 88.8% completeness (61.0% single-copy, 27.8% duplicated, 7.3% missing), reflecting high annotation quality. A non-redundant set, retaining only the longest transcript per gene, achieved 82.0% completeness, with 79.7% as single-copy genes (Table 6). Although slightly lower in overall BUSCO recovery compared to the full set, the non-redundant protein dataset is more suitable for downstream analyses such as GO/KEGG functional annotation and comparative genomics. Therefore, the full annotated protein set was used for evaluating annotation quality, while the non-redundant dataset was applied in subsequent functional and evolutionary analyses.

Data availability

All raw sequencing data generated in this study—including Illumina short reads, PacBio HiFi long reads, Hi-C sequencing data, and multi-tissue RNA-seq datasets—have been deposited in the Genome Sequence Archive (GSA) of the National Genomics Data Center (NGDC) under the accession number CRA033719, and are publicly available. The corresponding GSA run IDs are CRR2340812, CRR2340813, CRR2340814, and CRR2340815.

The chromosome-scale genome assembly has been deposited in an INSDC repository (DDBJ/ENA/GenBank) under the accession number JBUERZ000000000 and Genome Warehouse (GWH) of NGDC under accession number GWHHCKV00000000.1.

The genome annotation dataset has been deposited in Figshare and is publicly available at: https://doi.org/10.6084/m9.figshare.29442713.

Code availability

No in-house code or scripts were used in this study. Commands and pipelines used for data processing were executed using their corresponding default parameters.

Received: 10 July 2025; Accepted: 10 February 2026;
Published online: 24 February 2026

References

1. Barooha C. & Ahmed, I. Plant Diversity of Assam: A Checklist of Angiosperm & Gymnosperms. (2014).
2. Saikia1a, P. & Choudhury, B. I. & Khan, M. L. Floristic composition and plant utilization pattern in homegardens of Upper Assam, India. *Tropical Ecology* **53**, 105-118, https://doi.org/10.5772/31871 (2012).
3. Mao, A. A. & Dash, S. S. Flowering plants of India an annotated checklist: dicotyledons. (2020).
4. Zhang, L.-B. & Gilbert, M. G. Comparison of classifications of vascular plants of China. *Taxon* **64**, 17–26, https://doi.org/10.12705/641.23 (2015).
5. Volis, S. How to conserve threatened Chinese plant species with extremely small populations? *Plant Diversity* **38**, 45–52, https://doi.org/10.1016/j.pld.2016.05.003 (2016).
6. Ren, H. *et al*. Wild plant species with extremely small populations require conservation and reintroduction in China. *Ambio* **41**, 913–917, https://doi.org/10.1007/s13280-012-0284-3 (2012).
7. Ma, Y. *et al*. Conserving plant species with extremely small populations (PSESP) in China. *Biodiversity and Conservation* **22**, 803–809, https://doi.org/10.1007/s10531-013-0434-3 (2013).
8. Zhu, D. *et al*. Genome Survey Indicated Complex Evolutionary History of Garuga Roxb. Species. *BMC Genomics* **25**, e993, https://doi.org/10.1186/s12864-024-10917-8 (2024).
9. Aboul-Maaty, N. A.-F. & Oraby, H. A.-S. Extraction of high-quality genomic DNA from different plant orders applying a modified CTAB-based method. *Bulletin of the National Research Centre* **43**, 1–10, https://doi.org/10.1186/s42269-019-0066-1 (2019).
10. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107, https://doi.org/10.1002/imt2.107 (2023).
11. Ghosh, R. Karyotype study of somatic chromosomes in *Garuga Pinnata* Roxb. As an aid to its discussion in phylogeny. *Caryologia* **19**, 151–155, https://doi.org/10.1080/00087114.1966.10796213 (1966).
12. 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).
13. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
14. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
15. Formenti, G. *et al*. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. *Bioinformatics* **38**, 4214–4216, https://doi.org/10.1093/bioinformatics/btac460 (2022).
16. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).

17. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
18. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
19. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
20. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4.10. 11–14.10. 14, https://doi.org/10.1002/0471250953.bi0410s05 (2004).
21. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
22. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6**, 1–6, https://doi.org/10.1186/s13100-015-0041-9 (2015).
23. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
24. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome biology* **20**, 1–13, https://doi.org/10.1186/s13059-019-1910-1 (2019).
25. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
26. Lamesch, P. *et al*. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research* **40**, D1202–D1210, https://doi.org/10.1093/nar/gkr1090 (2012).
27. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research* **32**, W20–W25, https://doi.org/10.1093/nar/gkh435 (2004).
28. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988–995, http://www.genome.org/cgi/doi/10.1101/gr.1865504 (2004).
29. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465–W467, https://doi.org/10.1093/nar/gki458 (2005).
30. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 1–9, https://doi.org/10.1186/1471-2105-5-59 (2004).
31. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
32. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
33. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
34. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic acids research* **46**, 2699–2699, https://doi.org/10.1093/nar/gky092 (2018).
35. Kanehisa, M. Kegg: Kyoto encyclopedia of genes and genomes. *Proc. Natl. Acad. Sci.* **98**, 4569–4574, https://doi.org/10.1093/nar/28.1.27 (2001).
36. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
37. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
38. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
39. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic acids research* **42**, D222–D230, https://doi.org/10.1093/nar/gkt1223 (2014).
40. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic acids research* **28**, 231–234, https://doi.org/10.1093/nar/28.1.231 (2000).
41. Mi, H. *et al*. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research* **33**, D284–D288, https://doi.org/10.1093/nar/gki078 (2005).
42. Attwood, T. K. The PRINTS database: a resource for identification of protein families. *Briefings in bioinformatics* **3**, 252–263, https://doi.org/10.1093/bib/3.3.252 (2002).
43. Bru, C. *et al*. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic acids research* **33**, D212–D215, https://doi.org/10.1093/nar/gki034 (2005).
44. Loman T. A novel method for predicting ribosomal RNA genes in prokaryotic genomes. (2017).
45. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
46. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337, https://doi.org/10.1093/bioinformatics/btp157 (2009).
47. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic acids research* **49**, D192–D200, https://doi.org/10.1093/nar/gkaa1047 (2021).
48. Database Resources of the National Genomics Data Center. China National Center for Bioinformation in 2025. *Nucleic Acids Research* **53**, D30–D44 (2025).
49. The Updated Genome Warehouse. Enhancing Data Value, Security, and Usability to Address Data Expansion. *Genomics, Proteomics & Bioinformatics* **23**, qzaf010 (2025).
50. The GSA Family in 2025. A Broadened Sharing Platform for Multi-Omics and Multimodal Data. *Genomics, Proteomics & Bioinformatics* **23**, qzaf072 (2025).
51. *National Genomics Data Center* https://ngdc.cncb.ac.cn/gsa/browse/CRA033719/CRR2340812 (2025).
52. *National Genomics Data Center* https://ngdc.cncb.ac.cn/gsa/browse/CRA033719/CRR2340813 (2025).
53. *National Genomics Data Center* https://ngdc.cncb.ac.cn/gsa/browse/CRA033719/CRR2340814 (2025).
54. *National Genomics Data Center* https://ngdc.cncb.ac.cn/gsa/browse/CRA033719/CRR2340815 (2025).
55. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBUERZ000000000 (2025).
56. *NGDC/CNCB* https://ngdc.cncb.ac.cn/gwh/Assembly/104904/show (2025).
57. Chen, R. Chromosome-scale genome assembly and annotation of *Garuga floribunda* var. *gamblei Kalkman*. *Figshare*. https://doi.org/10.6084/m9.figshare.29442713 (2025).
58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
59. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 1–27, https://doi.org/10.1186/s13059-020-02134-9 (2020).
60. Solis-Reyes, S., Avino, M., Poon, A. & Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PloS one* **13**, e0206409, https://doi.org/10.1371/journal.pone.0206409 (2018).

61. 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. *Molecular biology and evolution* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).

Acknowledgements
We sincerely thank Chunmin Mao for his assistance with field sample collection. This work was supported by the National Natural Science Foundation of China (grant No. 42161015) and the Key Project of Basic Research of Yunnan Province, China (grant No. 202301AS070001).

Author contributions
Rong Chen led the genome assembly and annotation workflow, including raw data processing, genome polishing, repeat annotation, and structural gene annotation. She also prepared the initial draft of the manuscript. Rui Rao contributed substantially to the project, including data interpretation, figure refinement, and major manuscript editing and language polishing. Liang-Liang Yue, the corresponding author, conceived and supervised the project, provided strategic guidance, coordinated resources, and secured funding support. He critically reviewed and revised the manuscript for important intellectual content. All authors read and approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.-L.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.