



OPEN

DATA DESCRIPTOR

A chromosome-scale assembly of *Ormosia boluoensis* (Fabaceae)

Jin-Pi Zhu^{1,8}, Zheng-Feng Wang^{2,3,4,5,8}✉, Fang Cheng¹, En-Ping Yu^{2,3,4,5,6}, Lin Fu^{2,4,5,7}, Yan-Na Zeng¹, Yun-Mei Huang¹, Xin-Gui Chen¹, Sheng-Wen Deng¹, Yue Zhang¹, Jie-Ming Deng¹, Hao-Han Lin¹, Wei-Guang Zhu^{2,3,4,5}, Feng-Xia Xu^{2,4,5,7} & Hong-Lin Cao^{2,3,4,5}✉

The genus *Ormosia* is in the Fabaceae family. China is one of its centers of origin, with almost all species being endemic, suggesting the need for genomic studies to understand its evolution, conservation, and use. Therefore, we generated a chromosome-scale genome assembly of *O. boluoensis* using Oxford Nanopore long reads combined with Illumina short reads. The initial genome assembly generated by NextDenovo had a total length of 1,638,935,015 bp and a contig N50 of 16,933,043 bp. The assembly after Hi-C reads scaffolding was 1,565,663,440 bp with the scaffold N50 of 201,058,012 bp, and 1,561,022,207 bp (99.70%) for sequences were anchored to 8 chromosomes. Repeats accounted for 75.06% of the assembly. There were 51,822 predicted genes. The BUSCO evaluation showed 98.2% completeness at genome assessment level. The high-quality assembly and annotation provide valuable insights into the genomic characteristics of *O. boluoensis*, providing resources for its effective conservation and future utilization.

Background & Summary

The genus *Ormosia* (family Fabaceae) contains trees and shrubs and includes 130–150 species^{1,2}. They are warm-adapted lowland plants, with current distributions including southeast Asia, north Australia, and South America^{1–3}. China is one of its diversity centers, and its northernmost distribution extends to 23° north latitude^{4,5}. Currently, there are approximately 37 *Ormosia* species in China, 34 of which are endemic².

Ormosia species have hard, often red seeds used for jewelry and decorations^{1,4}. They can also be used as landscape trees and harvested for timber^{4,6,7}. Their seeds, roots, stems, bark, and leaves contain valuable secondary metabolites, such as alkaloids, flavonoids, isoflavones, terpenes, and lignin^{8–10}. These substances have antioxidant, anticancer, insecticidal, bactericidal, and other pharmacological activities⁸. As legume species, *Ormosia* forms nitrogen-fixing nodules on its root system^{11–13}.

Currently, four *Ormosia* nuclear genomes have been reported, including those of *O. purpureiflora*, *O. emarginata*, *O. semicastrata*¹⁴ and *O. henryi*¹⁵, all of which were assembled into the chromosome scales. Approximately 15 species, including *O. boluoensis* which we study here, have had their chloroplast genomes reported^{16,17}. However, only one mitochondrial genome has been published, which is from *O. boluoensis*¹⁸. The nuclear genome sizes for *Ormosia* are generally around 1.5 Gb as shown in *O. purpureiflora*, *O. emarginata* and *O. semicastrata*, while the size is approximately 2.7 Gb in *O. henryi*. Regardless of genome size, nearly 70% or more of their sequences are detected as repetitive elements. *Ormosia* is in the Genistoid lineage within Fabaceae¹⁹, and its closest genera are *Hovea* and *Poecilanthus*. To date, no genomes have been reported for these two genera.

O. boluoensis is endemic to China, has only been found at two sites in southeast China: Guangdong Xiangtoushan (XTS) National Natural Reserve and Guangdong Nankunshan (NKS) Provincial Natural Reserve.

¹Management Office of Guangdong Xiangtoushan National Natural Reserve, Huizhou, 516166, China. ²Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Guangzhou, 510650, China. ³Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. ⁴Key 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. ⁵South China National Botanical Garden, Guangzhou, 510650, China. ⁶University of Chinese Academy of Sciences, Beijing, 100049, China. ⁷Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. ⁸These authors contributed equally: Jin-Pi Zhu, Zheng-Feng Wang. ✉e-mail: wzf@scib.ac.cn; caohl@scib.ac.cn

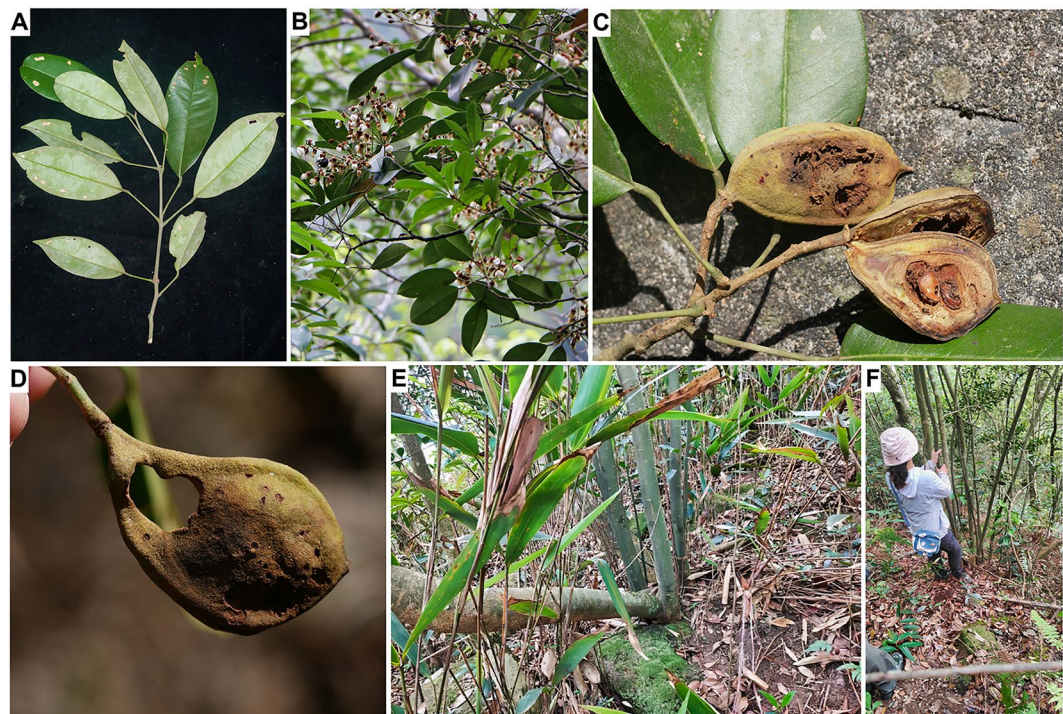


Fig. 1 Picture showing *Ormosia boluoensis*. (A) *O. boluoensis* leaves; (B) *O. boluoensis* flowers; (C) and (D) *O. boluoensis* seeds and fruit showing attack by worms/insects and/or disease; (E) and (F) *O. boluoensis* growth habitat and its clonal growth with root sprouting.

Field investigations by Guo²⁰ and the others in our group indicate that there are less than 1000 individuals in the reserves. Although both the chloroplast and mitochondrial genomes of *O. boluoensis* have been reported, its nuclear genome has not yet been assembled. Therefore, we present the chromosome-scale and fully annotated genome of *O. boluoensis*. Before genome assembly, chromosome number of *O. boluoensis* was observed using microscope and was determined to be $2n = 16$. The assembly and annotation were then performed by using approximately 127.13 Gb Oxford Nanopore Technology (ONT) long reads, 122.07 Gb next-generation sequencing (NGS) short reads, 148.19 Gb high-throughput chromosome conformation capture (Hi-C) reads, and 22.29 Gb RNA-seq reads. The initial *O. boluoensis* genome assembly by Nextdenovo was 1,638,935,015 bp with the contig N50 of 16,933,043 bp. The assembly after Hi-C reads scaffolding by Juicer pipeline, 3d-dna and manually curated was 1,565,663,440 bp with the scaffold N50 of 201,058,012 bp, and 1,561,022,207 bp (99.70%) for sequences were anchored to 8 chromosomes. 75.06% of the assembled sequences were identified as repeats. A total of 51,822 genes were predicted. The BUSCO evaluation for the assembled genome showed 98.2% completeness. The high quality of *O. boluoensis* genome assembly provides valuable resources for its evolutionary and conservation studies.

Methods

Species description. *O. boluoensis* is a shade-tolerant small tree or shrub. It is distributed at altitudes between 640 and 840 m and generally found along streams. It has leathery leaves, which are simple and alternate (Fig. 1A). Its flowers are white (Fig. 1B), bisexual, mostly in panicles and rarely in racemes. It is outcrossing species and pollinated by insects. In 2018 and 2019, individuals with a diameter at breast height (DBH) larger than 5 cm were observed to flower and set seeds, but only a few actually flowered in these years²⁰. In addition, because it suffers from severe pest/disease attacks on fruit/seeds (Fig. 1C,D), its seed yields were very low. No (healthy) seeds have been collected since 2020. Root-derived clonality was frequently observed in the field (Fig. 1E,F), indicating asexual reproduction. It has green bark (Fig. 1E).

Chromosome number observation. The individual used for chromosome number observation in *O. boluoensis* was regenerated from seeds collected from the XTS Reserve. At room temperature, the root tips were pretreated with 0.002 M 8-hydroxyquinoline for 6 h and fixed in 3:1 (v:v) absolute ethanol:glacial acetic acid for 24 h. The fixed root tips were hydrolyzed in 1:1 (v:v) 1 M absolute ethanol:hydrochloric acid for 7 min, washed by water, and stained with carbol fuchsin for 4 min. Meristems were subsequently excised, squashed and observed by microscope. Photographs were collected using an Olympus BX-43 microscope at 100x magnification with an Olympus DP26 camera. The chromosome number observation in *O. boluoensis* indicated $2n = 16$ (Fig. 2A).

Sample collection and sequencing. One *O. boluoensis* individual collected from the XTS Reserve was used for genome assembly. The collection was approved by the Management Office of Guangdong Xiangtoushan National Natural Reserve under the project of “Research on *Manis pentadactyla* and the other Rare and

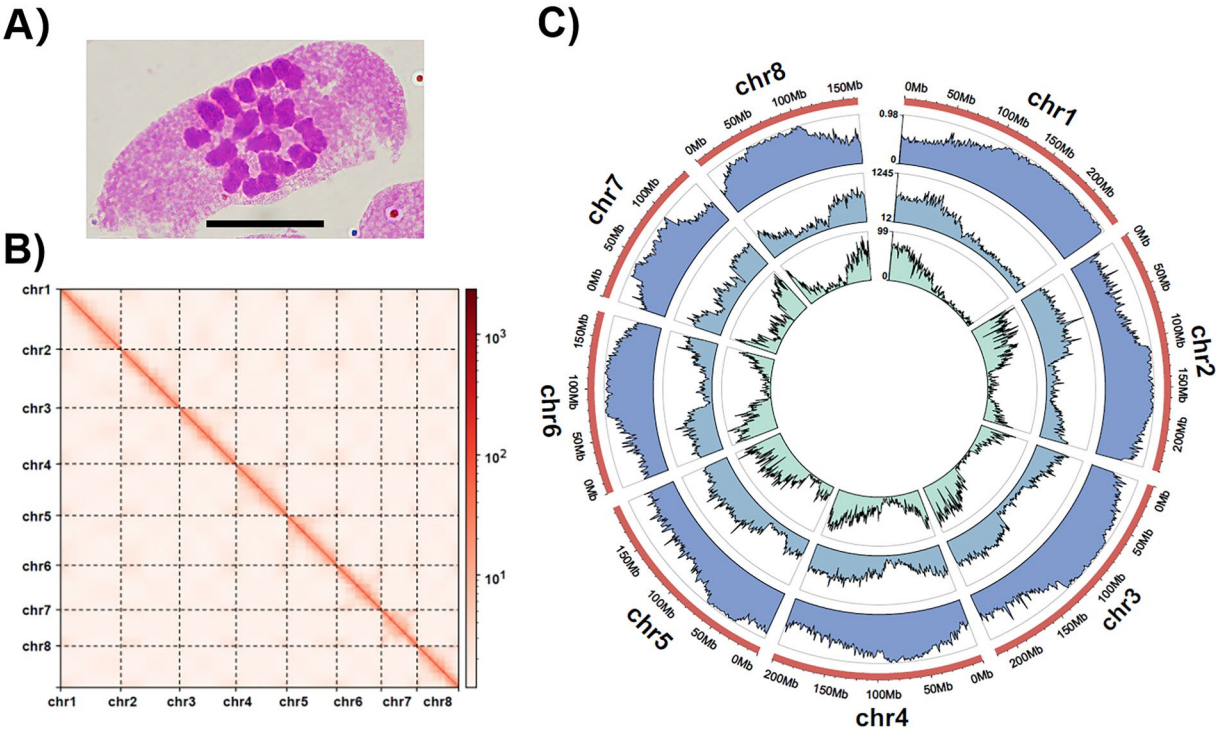


Fig. 2 (A) Chromosome numbers observed in *Ormosia boluoensis* (scale bar: 10 μm); (B) Hi-C interaction heat maps (bin length 100,000 bp) for the *O. boluoensis* genome assembly; (C) Circos plot showing the genome features (chromosome, repeat density in length proportions, repeat density in numbers, gene density). All densities were estimated with a 1-Mbp sliding window.

Data	Number of bases (Gb)	Number of reads	N50 length (bp)	Depth
Long WGS read	127.13	6,942,054	25,315	81×
Short WGS read	122.07	813,888,666	150	78×
Hi-C read	148.19	987,961,096	150	95×
RNA-seq read	22.29	148,623,502	150	/

Table 1. Sequencing data for *Ormosia boluoensis* assembly.

Endangered Animals and Plants in Guangdong Xiangtoushan National Nature Reserve — Genetic Diversity and Seedling Breeding of *O. boluoensis* (2024)”. For genome assembly and annotation, genomic DNA and RNA were isolated from its leaf tissues, and multiple libraries, including long- and short-read whole genome sequencing (WGS), Hi-C and RNA-seq libraries were constructed. Long-read WGS was applied using an Oxford Nanopore Technologies (ONT) PromethION sequencer. Short-read WGS, Hi-C, and RNA sequencing were performed using Illumina HiSeq X Ten platform sequencer with a 150-bp paired-end (insert size 300 bp) sequencing strategy. Detailed libraries construction and sequencing information, including DNA/RNA preparation and library construction, has been reported in our previous studies^{21,22}. Briefly, genomic DNA of *O. boluoensis* was extracted using the cetyltrimethylammonium bromide method. Extracted DNAs were used to build short-read and long-read WGS libraries. For the short-read WGS libraries, the DNAs were randomly fragmented to an average size of 200–400 bp for the final library construction. For the long-read WGS library, the extracted DNAs were size-selected and ligated with adapters to construct sequencing libraries. To perform Hi-C scaffolding, leaf tissues of *O. boluoensis* were immersed into nuclei isolation buffer with 2% formaldehyde for fixation. After fixation, nuclei were sheared into 300–600 bp fragments for the Hi-C library construction. After quality and integrity examination of extracted RNA, the RNA sequencing libraries were generated using RNA library preparing kit. After sequencing, these data were used for genome assembly, annotation, and comparative genomics analysis, with the default parameters in all programs unless otherwise specified.

Finally, the ONT sequencing platform generated approximately 127.13 Gb of WGS reads. The Illumina platform generated approximately 122.07 Gb of short WGS reads, 148.19 Gb of Hi-C reads, and 22.29 Gb of RNA-seq reads (Table 1).

Date pre-processing. After sequencing, the adapters in long WGS reads were trimmed using Porchop v0.2.4²³ with the parameter “--check_reads 100000 --adapter_threshold 80”. Short WGS and Hi-C reads were quality trimmed using Sickle v1.33²⁴ by removing the reads with base quality values less than 30 and lengths

shorter than 80 bp. Short WGS reads were further error corrected using RECKONER v1.1²⁵, and duplicates were removed by FastUniq v1.1²⁶. Using the error-corrected reads of *O. boluoensis*, the genome size of *O. boluoensis* was estimated using Kmergenie v1.7044²⁷ with the parameters “-diploid-k 141”. The genome size of *O. boluoensis* estimated in Kmergenie was 1,467,753,196 bp.

Genome assembly. The *O. boluoensis* genome was assembled by Nextdenovo 2.3.1²⁸ using pre-processed Nanopore long reads with lengths larger than 1000 bp. The specific parameters for our Nanopore reads in Nextdenovo were set using “read_type = ont, input_type = raw” with the others at their default values. After assembly, the genomes were polished by Racon v1.4.21²⁹ and Hapo-G v1.0³⁰ using long and short WGS reads, respectively, and each procedure was performed twice. After polishing, Pseudohaploid³¹ and Purge_Dups v1.2.5³² were used to examine and remove duplications in the assembly.

Scaffhic v1.1³³ was subsequently used to identify misassemblies and break the contigs with trimmed Hi-C reads. The Juicer pipeline 1.6³⁴ and 3d-dna v201008³⁵ were then applied to perform scaffolding with Hi-C reads. The scaffolding results were visualized using Juicebox v2.04.06³⁶, and the errors were manually corrected when necessary. The major corrections included chromosome boundary adjusting, misjoin and inversion error identifications. After scaffolding, TGS-GapCloser v1.0.1³⁷ was used to close the gaps with long WGS reads, and Racon and Hapo-G were used to polish the gap-closed genome. To detect potential chimeric assemblies, we used chimeric-contig-detector v1.0.2³⁸ to check them. The program contains two modes of chimera detection: one is “GC-Content Based Detection” mode, and the other is “Read-Pair Based Detection” mode. The latter mode uses short WGS reads.

The initial assembly in Nextdenovo was 1,638,935,015 bp with the contig N50 of 16,933,043 bp (Table 2). The assembly after Hi-C reads scaffolding was 1,565,663,440 bp with the scaffold N50 of 201,058,012 bp, and 1,561,022,207 bp (99.70%) for sequences assembled into 8 chromosomes (Fig. 2B). The largest chromosome size was 236,629,900 bp, and the shortest was 141,093,973 bp. No chimeras were detected under both “GC-Content Based Detection” and “Read-Pair Based Detection” mode by chimeric-contig-detector.

Repeat and gene prediction. Repeat and gene prediction in *O. boluoensis* followed the procedures previously described by Wang *et al.*^{21,22}. Briefly, repeats were identified using EDTA v2.0.1³⁹ and RED v2.0⁴⁰, and their results were then combined and soft-masked in the assembled genome.

Based on the results of EDTA and RED, 70.35 and 63.04% of genomes, respectively, were identified as repetitive regions. According to EDTA (Table 3), the highest repetitive sequences were long terminal repeats (LTRs), accounting for 61.50% (962,829,486 bp) of the genome size, followed by terminal inverted repeats (TIRs) with 6.74% (105,292,142 bp) of the genome size. For LTRs, the large proportion of repetitive sequences were Gypsy-like, with a sequence size of 555,597,256 bp (35.49%).

Gene prediction was performed using BRAKER3 v3.0.8⁴¹ and Funannotate pipeline v1.8.16⁴². BRAKER3 is a gene annotation pipeline that integrates evidence from transcript reads and homologous proteins. For transcriptome-based gene annotation, RNA-seq reads were aligned to the soft-masked assembled *O. boluoensis* genome using Hisat2 v2.2.1⁴³ and then assembled with StringTie2 v2.2.1⁴⁴. For homology-based gene annotation, protein sequences from 13 reference genomes from Fabaceae (Table 4) were mapped to the assembled genome, generating hints on the gene structure using ProtHint v2.6.0⁴⁵. These results were used to train GeneMark-ETP v1.0⁴⁶ and AUGUSTUS v3.5.0⁴⁷ to predict genes. The predictions were then combined and filtered by TSEBRA v1.1.2.5⁴⁸ to obtain non-redundant and consensus BRAKER3 gene sets, which were further input into Funannotate to generate the final gene annotation¹⁴. After gene prediction, a funannotate pipeline was used for gene functional annotation with different databases, including dbCAN v10.0, EggNOG v5.0.2, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), InterPro v5.62–94.0, MEROPS v12.0, Pfam v35.0, SignalP 5.0b and UniProt v2023_02^{18,19}. The reference proteins used for gene predictions are shown in Table 4.

By combining EDTA and RED results, 1,175,171,650 bp (75.06%) of the assembled genome was annotated and masked as repetitive components. The density of repeat sequences in the genome is shown in Fig. 2C. A total of 51,822 genes encoding 56,242 proteins were predicted. Functional annotation of these protein-coding genes showed that 40,750 genes (72.45%) were annotated to at least one database (Table 5).

Data Records

The assembled genome was submitted to GenBank under the accession number of JAIEWQ000000000⁴⁹. The sequenced reads were deposited to NCBI Sequence Read Archive under the accession number SRR14802411 for the Nanopore reads⁵⁰, SRR14725128 for Illumina WGS reads⁵¹, SRR14800493 for the Illumina Hi-C reads⁵², SRR14710764 for the Illumina RNA-seq reads⁵³. *O. boluoensis* genome assembly and gene annotation files are submitted to figshare <https://doi.org/10.6084/m9.figshare.29109572.v1>⁵⁴.

Technical Validation

To evaluate the completeness of the assembly, Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.5.0⁵⁵ was used with the eudicots_odb10.2020-09-10 database. The database contains 2326 conserved core genes of eudicots. The completeness of the predicted genes with their longest transcripts was also evaluated using BUSCO with the eudicots_odb10.2020-09-10 database.

The completeness of the genome assessed by BUSCO at the genome assessment level with the parameter of “-m geno” was 98.2% complete, including 90.2% complete single-copy genes, 8% complete and duplicated genes, 1.4% fragmented genes, and 0.4% missing genes. Gene prediction completeness, as assessed using BUSCO with the parameter of “-m prot”, indicated a completeness score of 96.7% in *O. boluoensis* (89.9% complete and single-copy, 6.8% complete but duplicated), with 1.5% fragmented and 1.8% missing genes. We discovered that

Contig statistics of initial assembly using Nextdenovo	
The length of sequence (bp)	The order of sequence length
N10 = 43571457	L10 = 3
N20 = 34757770	L20 = 8
N30 = 28126041	L30 = 13
N40 = 22201282	L40 = 20
N50 = 16933043	L50 = 28
N60 = 15213351	L60 = 38
N70 = 11829838	L70 = 51
N80 = 8920529	L80 = 67
N90 = 4468952	L90 = 93
N100 = 24824	L100 = 286
Total length	1638935015 bp
Average length	5730542 bp
Largest length	62261387 bp
Minimum length	24824 bp
Scaffold statistics after Hi-C scaffolding	
N10 = 236629900	L10 = 1
N20 = 230128291	L20 = 2
N30 = 220340375	L30 = 3
N40 = 220340375	L40 = 3
N50 = 201058012	L50 = 4
N60 = 194964083	L60 = 5
N70 = 173489955	L70 = 6
N80 = 173489955	L80 = 6
N90 = 163317618	L90 = 7
N100 = 1000	L100 = 172
Total length	1565663440 bp
Average length	9102694.42 bp
Largest length	236629900 bp
Minimum length	1000 bp
GC content	34.60%
Percentage of total repeat sequences	75.06%
Number of predicted genes	51,822
BUSCO completeness at genome level	98.2%
BUSCO completeness at gene level	96.7%
LAI values	13.23
Quality value (QV) from Merquy	30.01
Mapping ratio of long WGS reads	98.52%
Mapping ratio of short WGS reads	94.99%
Mapping ratio of RNA-seq reads	94.82%

Table 2. Statistics of *Ormosia boluoensis* genome assembly.

the BUSCO completeness score at the gene level was lower than that at the genome level, possibly due to gene prediction incompleteness. Therefore, future improvements to gene prediction are required.

The assembly quality was also examined by AssemblyQC v. 2.1.1⁵⁶ and sequencing read mapping ratios. The former helps estimate the Long Terminal Repeat (LTR) Assembly Index (LAI)⁵⁷, which evaluates assembly contiguity using repetitive sequences, as well as k-mer-based accuracy and completeness by Merquy⁵⁸. While for various types of long WGS, short WGS and RNA-seq reads, they were mapped to the assembly with Minimap2 v2.29⁵⁹, BWA v0.7.19⁶⁰ and Hisat2, respectively. Their mapping ratio were then calculated by Samtools v1.21⁶¹.

The LAI values for the *O. boluoensis* assembly was 13.23, meeting the “Reference genome” standard. Merquy indicated the quality value (QV) score was 30.01. Mapping-based evaluations indicated 98.52%, 94.99% and 94.82% mapping ratio for long WGS, short WGS and RNA-seq reads, respectively.

Data availability

The assembled genome can be downloaded on the website at <https://identifiers.org/ncbi/insdc:JAIEWQ000000000>. The raw Nanopore reads can be downloaded on the website at <https://identifiers.org/ncbi/insdc.sra:SRR14802411>, Illumina WGS reads at <https://identifiers.org/ncbi/insdc.sra:SRR14725128>, Hi-C reads at <https://identifiers.org/ncbi/insdc.sra:SRR14800493>, RNA-seq reads at <https://identifiers.org/ncbi/insdc.sra:SRR14710764>. The website of <https://doi.org/10.6084/m9.figshare.29109572.v1> includes two compressed files. One is assembled genome under the

Repeat class	Repeat family	Number of elements	Total length (bp)	Percentage in the genome
LTR		973174	962829486	61.50%
	Copia	190769	174873107	11.17%
	Gypsy	457370	555597256	35.49%
	unknown	325035	232359123	14.84%
TIR		302860	105292142	6.74%
	CACTA	83831	30165425	1.93%
	Mutator	126278	40047851	2.56%
	PIF_Harbinger	19125	6352392	0.41%
	Tc1_Mariner	14514	4202165	0.27%
	hAT	59112	24524309	1.57%
nonTIR		107273	33274526	2.13%
	helitron	107273	33274526	2.13%
Total		1383307	1101396154	70.35%

Table 3. Repeat content of the Hi-C scaffolded assembly by EDTA.

Species	GenBank accession number or website
<i>Abrus precatorius</i>	GCF_003935025.1
<i>Aeschynomene evenia</i>	http://aeschynomenebase.fr/download
<i>Ammopiptanthus mongolicus</i>	https://ammopiptanthus.github.io/
<i>Cicer arietinum</i>	GCF_000331145.1
<i>Crotalaria pallida</i>	https://figshare.com/articles/dataset/Genome_assembly_of_i_Crotalaria_pallida_i_/24995756
<i>Glycine max</i>	GCF_000004515.6
<i>Lupinus albus</i>	GCA_010261695.1
<i>Lupinus angustifolius</i>	GCF_001865875.1
<i>Ormosia semicastrata</i>	https://doi.org/10.6084/m9.figshare.26826466.v1
<i>Ormosia emarginata</i>	https://doi.org/10.6084/m9.figshare.26826466.v1
<i>Ormosia purpureiflora</i>	https://doi.org/10.6084/m9.figshare.26826466.v1
<i>Senna tora</i>	GCA_014851425.1
<i>Sophora flavescens</i>	https://zenodo.org/records/7750935

Table 4. Protein sequences of the species used for gene prediction.

Annotation database	Number of genes
dbCAN	1,576
EggNOG	39,865
KEGG	19,840
GO	27,897
InterPro	33,240
MEROPS	1,318
Pfam	27,579
SignalP	3,860
UniProt	10,783
Total	40,750

Table 5. Summary of gene functional annotations of the *Ormosia boluoensis* genome using different databases.

file name of “blhd_change_chr_softmask.rar”. In the file, the repetitive sequences are softmasked (showing as lowercase letters) and non-repetitive sequences are shown as uppercase letters. The other is gene prediction and annotation file under the file name of “Ormosia_boluoensis-gene_annotation.rar”. After decompression, six files will be generated. These are 1) all gene file under the name of “Ormosia_boluoensis_all_transcription.gff3” including all predicted genes and their transcripts information, 2) only longest transcripts file under the name of “Ormosia_boluoensis_only_longest_transcription.gff3” including all predicted genes and their longest transcripts information, 3) nucleotide sequences (all isoforms) of predicted genes under the name of “Ormosia_boluoensis_all_transcription-CDS.fa”, 4) translated sequences (all isoforms) of predicted genes under the name of “Ormosia_boluoensis_all_transcription-protein.fa”, 5) nucleotide sequences (only longest transcripts) of predicted genes under the name of “Ormosia_boluoensis_only_longest_transcription-CDS.fa”, 6) translated sequences (only longest transcripts) of predicted genes under the name of “Ormosia_boluoensis_only_longest_transcription-protein.fa”.

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: 22 May 2025; Accepted: 4 September 2025;

Published online: 20 October 2025

References

1. Torke, B. M. *et al.* A dated molecular phylogeny and biogeographical analysis reveals the evolutionary history of the trans-pacifically disjunct tropical tree genus *Ormosia* (Fabaceae). *Mol. Phylogenet. Evol.* **166**, 107329, <https://doi.org/10.1016/j.ympev.2021.107329> (2022).
2. Niu, M. *et al.* Two new synonyms of *Ormosia semicastrata* (Fabaceae, Papilionoideae, Ormosieae). *Phytotaxa* **613**(2), 140–152, <https://doi.org/10.11646/phytotaxa.613.2.3> (2023).
3. Wang, Z., Shi, G., Sun, B. & Yin, S. A new species of *Ormosia* (Leguminosae) from the middle Miocene of Fujian, Southeast China and its biogeography. *Rev. Palaeobot. Palyno.* **270**, 40–47, <https://doi.org/10.1016/j.revpalbo.2019.07.003> (2019).
4. Zhou, C. *et al.* Genetic structure of an endangered species *Ormosia henryi* in southern China, and implications for conservation. *BMC Plant Biol.* **23**, 220, <https://doi.org/10.1186/s12870-023-04231-w> (2023).
5. Jiang, L. *et al.* Study on the potential suitable area of *Ormosia* in Guangdong Province based on the MaxEnt Model. *Guangxi Sciences* **31**(1), 149–166, <https://doi.org/10.13656/j.cnki.gxkx.20230719.001> (2024).
6. Li, L. *et al.* First report of dieback caused by *Lasiodiplodia pseudotheobromae* on *Ormosia pinnata* in China. *Plant Dis.* **104**, 2551–2555, <https://doi.org/10.1094/PDIS-03-20-0647-RE> (2020).
7. Wei, L., Wang, G. & Xie, C. Predicting suitable habitat for the endangered tree *Ormosia microphylla* in China. *Sci. Rep.* **14**, 10330, <https://doi.org/10.1038/s41598-024-61200-5> (2024).
8. Zhang, L.-J. *et al.* A review on chemical constituents and pharmacological activities of *Ormosia*. *Chin. Tradit. Herbal Drugs*. **52**(14), 4433–4442, <https://doi.org/10.7501/j.issn.0253-2670.2021.14.035> (2021).
9. Zhou, Q.-Q. *et al.* Hosimosines A–E, structurally diverse cytisine derivatives from the seeds of *Ormosia hosiei* Hemsl. *et* Wils. *Fitoterapia* **170**, 105661, <https://doi.org/10.1016/j.fitote.2023.105661> (2023).
10. Zhou, W. *et al.* A new lignan from leaves of *Ormosia xylocarpa*. *Rec. Nat. Prod.* **17**(1), 189–194, <https://doi.org/10.25135/rnp.338.2203.2386> (2023).
11. Domínguez-Núñez, J. A. & Berrocal-Lobo, M. in *Biofertilizers* (eds Rakshit, A., Meena, V. S., Parihar, M., Singh, H. B. & Singh, A. K.) Chapter 20 - Application of microorganisms in forest plant. (Woodhead Publishing, Cambridge, United Kingdom, pp. 265–287, <https://doi.org/10.1016/B978-0-12-821667-5.00026-9>. 2021).
12. Ye, N. *et al.* Different leaf nutrient use strategies of nitrogen-fixing and non-nitrogen-fixing leguminous trees in South China. *J. Trop. Subtrop. Bot.* **31**(3), 334–340, <https://doi.org/10.11926/jtsb.4583> (2023).
13. Zhu, T. *et al.* Screening and identification of plant growth-promoting rhizobacteria (PGPR) in the rhizosphere of *Ormosia henryi* and their growth-promoting characteristics. *J. Cent. South Univ. Forest. & Tech.* **43**(1), 43–49, <https://doi.org/10.14067/j.cnki.1673-923x.2023.01.004> (2023).
14. Wang, Z.-F. *et al.* Chromosome-scale assemblies of three *Ormosia* species: repetitive sequences distribution and structural rearrangement. *GigaScience* **14**, gfa047, <https://doi.org/10.1093/gigascience/gfa047> (2025).
15. Zhou, C. *et al.* Chromosome-level genome assembly of the endangered tree species *Ormosia henryi* Prain. *Sci. Data* **12**, 1065, <https://doi.org/10.1038/s41597-025-05402-0> (2025).
16. Tang, J., Zou, R., Wei, X. & Li, D. Complete chloroplast genome sequences of five *Ormosia* species: Molecular structure, comparative analysis, and phylogenetic analysis. *Horticulturae* **9**(7), 796, <https://doi.org/10.3390/horticulturae9070796> (2023).
17. Wang, Z.-F. *et al.* The complete chloroplast genome of *Ormosia purpureiflora* (Fabaceae). *Mitochondrial DNA B* **6**(12), 3327–3328, <https://doi.org/10.1080/23802359.2021.1994901> (2021).
18. Wang, Z.-F., Zhang, Y., Zhong, X.-J., Kang, N. & Cao, H.-L. The complete mitochondrial genome of *Ormosia boluoensis*. *Mitochondrial DNA B* **6**(8), 2109–2111, <https://doi.org/10.1080/23802359.2021.1920503> (2021).
19. Zhao, Y. *et al.* Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol. Plant* **14**(5), 748–773, <https://doi.org/10.1016/j.molp.2021.02.006> (2021).
20. Guo, Y. *Preliminary study on endangered mechanism of Ormosia boluoensis with extremely small population* (Master's thesis. University of Chinese Academy of Sciences, 2020).
21. Wang, Z.-F., Rouard, M., Droc, G., Heslop-Harrison, P. & Ge, X.-J. Genome assembly of *Musa beccarii* shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes. *GigaScience* **12**, giad005, <https://doi.org/10.1093/gigascience/giad005> (2023).
22. Wang, Z.-F. *et al.* Chromosome-level genome assembly and demographic history of *Euryodendron excelsum* in monotypic genus endemic to China. *DNA Res.* **31**(1), dsad028, <https://doi.org/10.1093/dnares/dsad028> (2024).
23. Porchop v0.2.4, <https://github.com/rwrick/Porchop> (Accessed 8 January 2019).
24. Joshi, N. A. & Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). 2011, <https://github.com/najoshi/sickle> (Accessed 28 August 2020).
25. Dlugosz, M. & Deorowicz, S. RECKONER: read error corrector based on KMC. *Bioinformatics* **33**, 1086–1089, <https://doi.org/10.1093/bioinformatics/btw746> (2017).
26. Xu, H. *et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* **7**(12), e52249, <https://doi.org/10.1371/journal.pone.0052249> (2012).
27. Chikhi, R. & Medvedev, P. Informed and automated *k*-mer size selection for genome assembly. *Bioinformatics* **30**, 31–37, <https://doi.org/10.1093/bioinformatics/btt310> (2014).
28. 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).
29. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**(5), 737–746, <https://doi.org/10.1101/gr.214270.116> (2017).
30. Aury, J. M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom. Bioinform.* **3**(2), lqab034, <https://doi.org/10.1093/nargab/lqab034> (2021).
31. Pseudohaploid, <https://github.com/schatzlab/pseudohaploid> (Accessed 28 August 2020).
32. Guan, D. F. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, <https://doi.org/10.1093/bioinformatics/btaa025> (2020).
33. <https://github.com/wtsi-hpag/scaffHiC> (2022). Scaffhic v1.1 Accessed 7 December.
34. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**(1), 95–98, <https://doi.org/10.1016/j.cels.2016.07.002> (2016).

35. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333), 92–95, <https://doi.org/10.1126/science.aal3327> (2017).
36. Robinson, J. T. *et al.* Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst.* **6**(2), 256–258.e1, <https://doi.org/10.1016/j.cels.2018.01.001> (2018).
37. 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**(9), gaaa094, <https://doi.org/10.1093/gigascience/gaaa094> (2020).
38. Chimeric-contig-detector v1.0.2, <https://github.com/megjohnson1999/chimeric-contig-detector> (Accessed 27 August 2025).
39. 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).
40. Girgis, H. Z. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinform* **16**(1), 227, <https://doi.org/10.1186/s12859-015-0654-5> (2015).
41. Gabriel, L., *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**(5), 769–777, <https://doi.org/10.1101/gr.278090.123> (2024).
42. Funannotate v1.8.16, <https://github.com/nextgenusfs/funannotate> (Accessed 12 March 2023).
43. 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).
44. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278, <https://doi.org/10.1186/s13059-019-1910-1> (2019).
45. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform.* **2**(2), lqaa026 (2020).
46. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**(5), 757–768, <https://doi.org/10.1101/gr.278373.123> (2024).
47. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**(suppl_2), W435–W439, <https://doi.org/10.1093/nar/gkl200> (2006).
48. Gabriel, L., Bruna, T., Hoff, K. J., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 1–12, <https://doi.org/10.1186/s12859-021-04482-0> (2021).
49. Zhu, J.-P. *et al.* NCBI GenBank <https://identifiers.org/ncbi/insdc:JAIEWQ000000000> (2025).
50. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR14802411> (2025).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR14725128> (2025).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR14800493> (2025).
53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR14710764> (2025).
54. Zhu, J.-P. *et al.* A chromosome-scale assembly of *Ormosia boluoensis* (Fabaceae). *Figshare* <https://doi.org/10.6084/m9.figshare.29109572.v1> (2025).
55. Seppy, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245, <https://doi.org/10.1093/bioinformatics/btv351> (2019).
56. Rashid, U. *et al.* AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. *Bioinformatics* **40**(8), btae477, <https://doi.org/10.1093/bioinformatics/btae477> (2024).
57. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**(21), e126, <https://doi.org/10.1093/nar/gky730> (2018).
58. 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).
59. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324> (2009).
61. Li, H. *et al.* & 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).

Acknowledgements

The study is supported by Guangdong Science and Technology Plan Project (2023A1111110001); Project of Research on *Manis pentadactyla* and the other Rare and Endangered Animals and Plants in Guangdong Xiangtoushan National Nature Reserve — Genetic Diversity and Seedling Breeding of *O. boluoensis*. Guangdong Provincial Forestry Bureau Project — Planning of the Provincial Plant *Ex Situ* Protection System and National Key Protected Plant *Ex Situ* Protection and Propagation; The National Natural Science Foundation of China (No. 32370406, 31970188).

Author contributions

J.-P.Z., Z.-F.W., F.C. and H.-L.C. conceived and supervised the study. J.-P.Z., Z.-F.W., F.C., L.F. and H.-L.C. designed the methodology; J.-P.Z., Z.-F.W., F.C., L.F. and H.-L.C. performed data analysis and validation. J.-P.Z., Z.-F.W., F.C., L.F. and H.-L.C.; Z.-F.W., F.C., E.-P.Y., Y.-N.Z., Y.-M.H., X.-G.C., S.-W.D., Y.Z., J.-M.D., H.-H.L., W.-G.Z., F.-X.X. collected the samples; J.-P.Z., Z.-F.W., F.C., L.F. and H.-L.C. prepared original draft. J.-P.Z., Z.-F.W., F.C. and H.-L.C. obtained fundings. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025