scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the endangered tree species *Ormosia henryi* Prain

Chengchuan Zhou[1], Qiang Wen[1], Dan Zeng[1], Changqing Guo[2], Zhiwen Guo[1], Liting Liu[1] ✉ & Tianlin Ouyang[2] ✉

Ormosia henryi **Prain, belonging to the Fabaceae family, is an endangered tree species with great economic, ornamental, and medicinal potential. The lack of genetic information and high-quality genome assembly hinders in-depth research on this species as well as its conservation and further utilization. Here, we assembled a chromosome-level genome of *O. henryi* using Pac-Bio HiFi, Hi-C and short read sequencing. The assembled genome size is approximately 2.69 Gb, with a scaffold N50 of 354.08 Mb and a complete BUSCO score of 98.20%. About 99.97% of the assembled sequences could be anchored to eight pseudo-chromosomes, and 69.09% of the assembled genome were repetitive sequences. In total, we predicted 42,260 protein-coding genes, 88.42% of which could be functionally annotated. This genome assembly provides a valuable resource for various applications, including future conservation and utilization, evolutionary studies, and elucidation of pathways involved in key biological traits of *O. henryi*.**

Background & Summary

The genus *Ormosia* (Fabaceae, Papilionoideae) comprises about 130 tree species and exhibits a disjunct distribution between the New World- and the Asian and northeastern Australian tropics and subtropics[1–3]. *Ormosia* species are usually characterized by extremely hard and brightly coloured seeds, which are thought to mimic the fleshy nutritious seeds of other species, thereby deceiving birds and other dispersers to ingest them[4,5]. The seeds of *Ormosia* species, often used as beads for jewellery, inspired the genus name *Ormosia*, which is derived from the Greek word *hormos*, meaning necklace[6]. In addition, many *Ormosia* species are excellent timber trees, and some species are used as medicinal plants with high economic value[7–9].

Ormosia henryi Prain is an evergreen broad-leaved tree species that grows mainly in southern China. The wood of *O. henryi* is hard, durable and beautifully grained, and has been used for centuries to make high quality furniture and handicrafts[10]. *O. henryi* also makes an excellent ornamental plant. In addition, the roots, branches and leaves of *O. henryi* have been used as Chinese folk medicine. It is believed that *O. henryi* can stimulate blood circulation and relieve blood stasis[11]. In southern provinces of China such as Jiangxi and Hunan, local people use *O. henryi* for the treatment of injuries from falling and rheumatic arthritis. Recently, *O. henryi* has been found to be rich in isoflavonoids with significant antidepressant effects[12,13]. However, due to over-exploitation, habitat destruction and difficulties in natural generation, natural populations of *O. henryi* have declined dramatically[14]. Field investigation found that the extant natural populations of *O. henryi* are usually very small, with only a few adult individuals. Therefore, *O. henryi* has been listed as a second-class National Key Protected Wild Plant in China and was evaluated as Vulnerable (VU) by International Union for Conservation of Nature (IUCN).

Despite its economic and ecological importance, relatively little is known about the genetics and genomics of *O. henryi*. A few genes and the chloroplast genome of *O. henryi* have been sequenced to infer the phylogeny of *Ormosia*[8,15]. A recent study based on metabolome and transcriptome has revealed the accumulation of isoflavonoids in five different organs of *O. henryi*, as well as identifying the potential transcriptional regulators[11]. This study also provides a basis for the genome annotation of *O. henryi*. Our preliminary investigation of the genetic structure of ten *O. henryi* populations in southern China based on genotyping by sequencing (GBS) method indicates that the endangered status of *O. henryi* has been underestimated and that immediate conservation

[1]Jiangxi Academy of Forestry, Nanchang, 330013, Jiangxi, P.R. China. [2]Jiangxi Provincial Forestry Science and Technology Experiment Center, Ganzhou, 341699, Jiangxi, P.R. China. ✉e-mail: more_trees@163.com; 13307079666@163.com

Statistical feature	Corresponding value
Assembled genome size (bp)	2,691,380,774
Number of contigs	150
Number of scaffolds	19
Contig N50 (bp)	37,594,281
Scaffold N50 (bp)	354,082,533
Number of chromosomes	8
Genome sequences anchored to chromosomes (bp)	2,690,505,642
Anchoring rate	99.97%
GC content	35.68%
BUSCO complete genes(C)	1585 (98.20%)
BUSCO single copy genes(S)	1472 (91.20%)
BUSCO duplicated genes(D)	113 (7.00%)
BUSCO fragmented genes(F)	18 (1.12%)
BUSCO missing genes (M)	11 (0.68%)

Table 1. Summary of *Ormosia henryi* genome assembly.

Type	Length (bp)	% in genome
DNA transposon	46,770,846	1.74
LINE	56,463,461	2.10
SINE	103,359	0.00
LTR	1,741,306,952	64.70
Satellite	1,141,065	0.04
Simple repeat	3,501	0.00
Other	32,074	0.00
Unknown	42,771,967	1.59
Total	1,858,759,373	69.06

Table 2. Summary of repeat sequences in *Ormosia henryi* genome.

Statistical feature	Corresponding value
Total gene number	42,260
Total transcript number	55,620
Mean gene length (bp)	5068.82
Functionally annotated gene number	37,366

Table 3. Summary of protein-coding gene annotation in *Ormosia henryi* genome.

measures are needed[14]. Thus, a high-quality genome of *O. henryi* is crucial for further elucidating the genetic mechanism of endangerment and developing scientific conservation strategies.

In this study, we assembled a high-quality chromosome-level genome of *O. henryi* using PacBio HiFi sequencing and Hi-C sequencing technologies. The assembly yielded a 2.69 Gb genome with a scaffold N50 length of 354.08 Mb. Approximately 99.97% of the assembled sequences were anchored to 8 pseudochromosomes by Hi-C guided assembly (Table 1). The assembly's quality was validated through Benchmarking Universal Single-Copy Ortholog (BUSCO) analysis, which revealed 98.20% completeness (Table 1). Of this genome assembly, 69.06% were repetitive sequences, mainly long terminal repeats (LTR), accounting for 64.70% of the genome (Table 2). Meanwhile, 42,260 protein-coding genes were predicted, with 37,366 genes being functionally annotated (Table 3). The chromosome-level genome assembly of *O. henryi* will serve as a valuable genetic resource for improving conservation efforts and facilitating the exploration of its diverse applications. In addition, the genome data can help elucidate the evolutionary history and relationships within the genus *Ormosia*, as well as the genetic basis underlying key traits such as hard wood formation and bioactive components biosynthesis.

Methods

Sample collection and sequencing. For genomic DNA extraction, fresh young leaves of *O. henryi* were collected from a single adult plant cultivated in Jiangxi Academy of Forestry, Nanchang city, Jiangxi Province, China (28.7412°N, 115.8168°E). The inflorescences of the same individual plant were collected for RNA extraction. The harvested materials were rapidly frozen in liquid nitrogen and then preserved at −80 °C until DNA and RNA extraction.

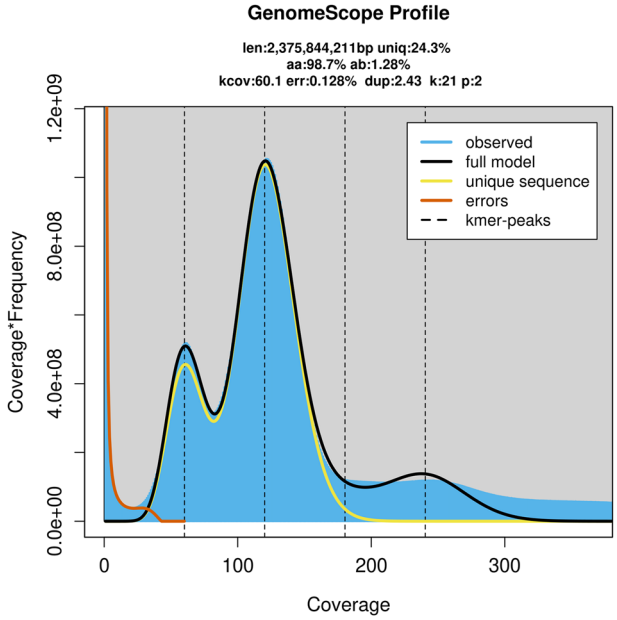

Fig. 1 Genome survey of *Ormosia henryi* based on the kmer distribution analysis using a kmer size of 21.

Genomic DNA was extracted following the modified CTAB method[16]. DNA quality and concentration were examined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA). For genome survey sequencing, the paired-end library was prepared using the DNBSEQ-T7RS High-throughput Sequencing Kit (MGI Tech, China), and the library was sequenced on a DNBSEQ-T7 platform (MGI Tech, China) following the manufacturer's instructions. This yielded about 385.88 Gb of paired-end raw reads, covering about 143.38 × of the *O. henryi* genome (Supplementary Table S1). For PacBio HiFi long-read sequencing, a SMRTbell library was prepared following the PacBio's standard protocol (Pacific Biosciences, USA). The library was then sequenced on a PacBio Revio System in CCS mode, generating 85.41 Gb of HiFi reads with a total depth of 31.74 × (Supplementary Table S1).

To prepare the library for high-throughput chromosome conformation capture (Hi-C) sequencing, fresh leaves were treated with formaldehyde for DNA cross-linking. Subsequently, the Hi-C library was constructed based on the standard protocol and sequenced on a DNBSEQ-T7 platform, generating 137.05 Gb raw read data with a total depth of 50.92 × (Supplementary Table S1).

For genome annotation, a published transcriptome sequencing dataset of five tissues (young leaves, old leaves, stems, barks, roots) from *O. henryi* was used[11] (Supplementary Table S1). Additionally, total RNA was extracted from the collected inflorescences for transcriptome sequencing. The RNA library was prepared using the Ultima Dual-mode mRNA Library Prep Kit (Yeason, China) according to the manufacturer's instructions, and PE150 sequencing was conducted on a DNBSEQ-T7 platform (MGI Tech, China). A total of ~72.22 Gb RNA-seq reads were used in the subsequent annotation analysis (Supplementary Table S1).

Genome assembly. Prior to the genome assembly, a genome survey was conducted to assess the main features of the *O. henryi* genome, including genome size, heterozygosity, and repeat content. The short reads were filtered using Fastp v0.23.2[17] with default parameters. The frequencies of kmer 21 were generated by Jellyfish v2.3.0[18] and used for the genome evaluation by GenomeScope v2.0[19]. The estimated genome size was about 2.38 Gb, with a genome heterozygosity of 1.28% and a repeat content of 75.71% (Fig. 1).

The PacBio HiFi reads were used to perform *de novo* genome assembly using Hifiasm v0.19.6[20] with default parameters. Purge_haplotigs v1.0.4[21] was used to purge the heterozygous and redundancy regions of the draft assembly. Potential contaminants were identified by BLASTN analysis against the NCBI nucleotide database using the BLAST + v2.11.0 software[22]. The bacterial and fungal contaminants were removed from the contigs based on the search results. The initial assembly resulted in a genome size of 2,954,680,083 bp, containing 150 contigs with a N50 length of 37,594,281 bp. Hi-C sequencing data were used to anchor the assembled contigs into pseudochromosomes. The filtered Hi-C data were first mapped to the initial genome assembly with Juicer v1.6[23]. The unique mapped reads were then taken as input for 3D-DNA pipeline v180922[24] with parameters "-r 0". Afterwards, a careful manual inspection and correction of any visual errors in the graph was done by JuiceBox v1.11.08[25]. As a result, eight pseudochromosomes were identified by distinct interaction signals in the Hi-C interaction heatmap (Fig. 2).

We finally obtained a chromosomal-level genome of 2,691,380,774 bp in size with a scaffold N50 length of 354,082,533 bp (Table 1). About 99.97% of the genomic sequences were anchored to the 8 pseudochromosomes, and the chromosomes were numbered according to the order of chromosome length (Fig. 2, Supplementary Table S2). The GC content of the *O. henryi* genome assembly was 35.68% (Table 1). Benchmarking Universal Single-Copy Ortholog (BUSCO) v5.4.3[26] was employed to assess the integrity, purity and completeness of the

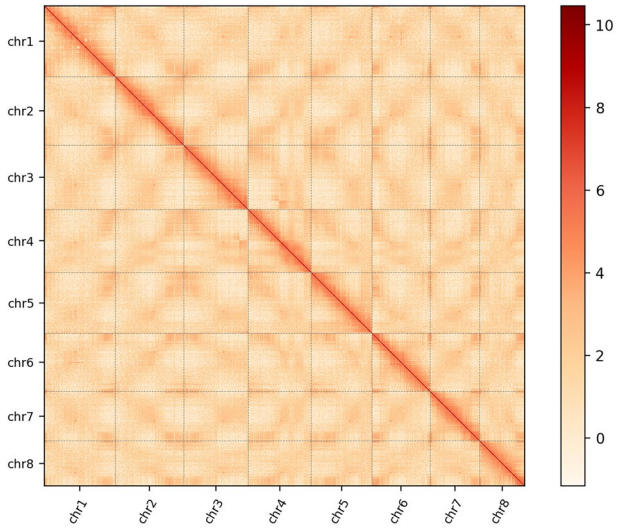

Fig. 2 Hi-C heatmap for the genome assembly of *Ormosia henryi* (bin size = 1 Mb).

genome using the embryophyta_odb10 gene set. A total of 1585 (98.20%) BUSCOs were identified as complete, including 1472 (91.20%) single-copy BUSCOs and 113 (7.00%) duplicated BUSCOs. Additionally, 18 (1.12%) BUSCOs were identified as fragmented, and 11 (0.68%) BUSCOs were identified as missing. (Table 1).

Genome annotation. Repeat elements in the assembled genome were identified by combining *de novo* and homology-based methods. For *de novo*-based identification, RepeatModeler v1.0.11[27] and LTR_FINDER v1.0.7[28] were used to construct the *de novo* repeat libraries following default parameters. RepeatMasker v4.0.9[29] was then applied to detect repeat elements based on these libraries. For homology-based identification, RepeatMasker v4.0.9 was employed against a known repeat database Repbase v23.08[30]. In addition, tandem repeats were annotated using Tandem Repeat Finder (TRF) v4.09 software[31] with default parameters. Finally, we identified a total of 1,858,759,373 bp as repeat sequences, which accounted for 69.06% of the genome assembly (Table 2). Most of these repeats were long terminal repeats (LTRs), which contributed to 64.70% of the genome assembly. The DNA transposons, long interspersed nuclear elements (LINEs), and Unknown repeat elements accounted for 1.74%, 2.10%, and 1.59% of the genome, respectively (Table 2).

Protein-coding genes were predicted by a combination of *ab initio* prediction, homology-based prediction and transcriptome-based prediction. The *ab initio* gene prediction approach was performed by AUGUSTUS v3.3.2[32], GlimmerHMM v3.0.4[33] and Genscan v1.0[34] based on the statistical models of gene structure. For homology-based prediction, we aligned the genome sequence against non-redundant protein sequences from closely related eight species, including *Glycine max*[35], *Medicago truncatula*[36], *Vigna unguiculata*[37], *Lupinus albus*[38], *Lupinus angustifolius*[39], *Pterocarpus macrocarpus*[40], *Pterocarpus santalinus*[40], *Sophora flavescens*[41], using miniprot v0.11[42] with default parameters. For transcriptome-based prediction, RNA-seq reads were mapped to the *O. henryi* genome by HISAT2 v2.1.0[43], and StringTie v1.3.5[44] was used to generate transcripts for reference-guided assembly. The redundant transcripts were removed using CD-HIT v.4.8.1[45]. The open reading frames (ORFs) were predicted by TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/). The gene models predicted by the three methods were integrated using Maker2 v2.31.10[46]. Our final gene set comprised 42,260 protein-coding genes, with an average gene length of 5068.82 bp (Table 3). Gene functional annotation was performed by aligning the predicted protein sequences against public functional gene databases using BLAST + v2.11.0 (e-value $< 10^{-5}$), including NCBI-NR[47], SwissProt[48], TrEMBL[48], InterPro[49], KEGG[50], COG[51], GO[52] and Pfam[53]. As a result, a total of 37,366 genes were functionally annotated, which accounted for 88.42% of the protein-coding genes (Table 3). The genomic features were then visualized by circos v0.69-9[54] (Fig. 3).

Various non-coding RNAs within the *O. henryi* genome were also predicted. The rRNA, miRNA, and snRNA were predicted by INFERNAL v1.1.2[55] based on the Rfam[56] database v14.8. The tRNAs were predicted using tRNAscan-SE v1.3.1[57]. Finally, 2,518 rRNAs, 229 miRNAs, 36,914 snRNAs and 1120 tRNAs were identified in the *O.henryi* genome (Supplementary Table S3).

Data Records

The raw sequencing data in this study, including genome survey short reads, PacBio HiFi reads, Hi-C reads, and RNA-seq short reads, have been deposited to the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa/) in the National Genomics Data Center (NGDC) with the accession number of CRA020446[58], CRA020384[59], CRA020383[60], CRA020387[61] under the BioProject PRJCA032243[62]. The final genome assembly was deposited in the NCBI database under the accession number of GCA_050613405.1[63], and was also available in the Genome Warehouse (GWH) of NGDC under the accession number GWHFHGT00000000.1[64]. The annotation files were deposited in the figshare database for broader accessibility[65].

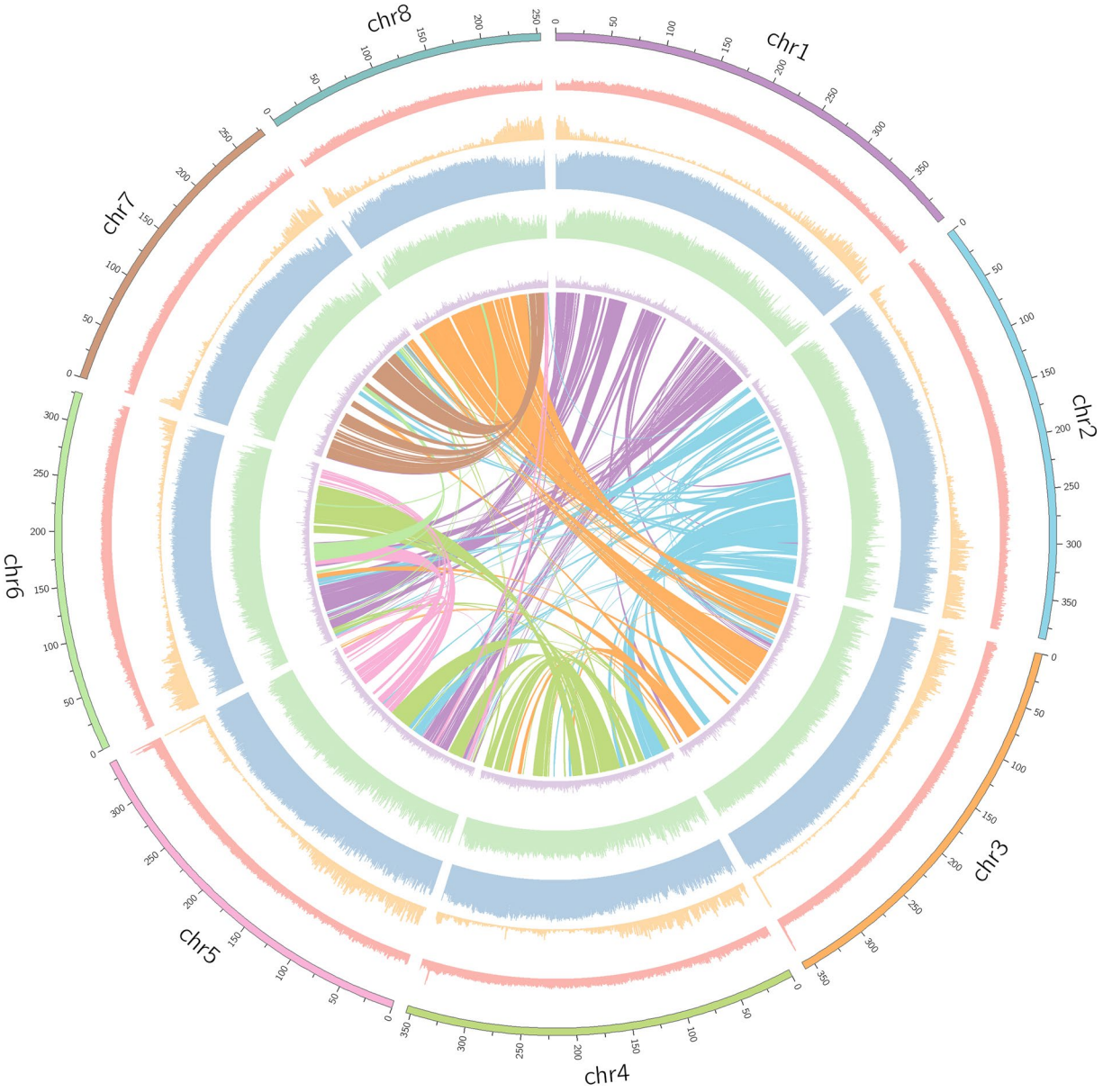

Fig. 3 The genomic features of *Ormosia henryi*. The features are arranged in the order of chromosomes, GC content, gene density, repeat density, LTR/Gypsy density, LTR/Copia density, and syntenic blocks from outside to inside across the 8 pseudochromosomes (window size = 500 kb).

Technical Validation

Genome assembly quality assessment. Two separate methods were employed to assess the quality of the assembled genome. Firstly, we evaluated assembly completeness utilizing BUSCO v5.4.3 against the embryophyta_odb10 reference gene set. The assembly exhibited a BUSCO completeness of 98.20% BUSCO genes, comprising 91.20% single-copy, 7.00% duplicated, 1.12% fragmented, and 0.68% missing. Secondly, the short reads from genome survey sequencing were mapped to the final genome assembly using BWA v0.7.12[66] with default settings, respectively. The mapping rate and coverage percentage of these short reads were 99.83% and 99.94%, respectively.

Gene annotation assessment. The predicted proteins were evaluated by BUSCO v5.4.3 with the embryophyte_odb10 dataset. Among a total of 1614 BUSCOs, 1572 (97.40%) BUSCOs were complete (1466 single-copy BUSCOs and 106 duplicated BUSCOs), 30 (1.86%) BUSCOs were fragmented and 12 (0.74%) BUSCOs were missing, which indicated high-quality annotation of the predicted gene models.

Code availability

No custom codes were used in this study. All bioinformatics tools and software applications were executed in accordance with their respective manuals and protocols. The specific software versions and the parameters used are detailed in the methods section.

Received: 22 November 2024; Accepted: 13 June 2025;
Published online: 23 June 2025

References

1. Sun, H. & Vincent, M. A. *Ormosia* Jackson. in *Flora of China* (eds. Wu, Z. Y., Raven, P. H. & Hong, D. Y.) **Vol. 10** (Science Press and Missouri Botanical Garden Press 2010).
2. Li, X. C. *et al*. *Ormosia* (Fabaceae: Faboideae) from the Miocene of southeastern China support historical expansion of the tropical genus in East Asia. *Hist. Biol.* **33**(12), 3561–3578 (2021).
3. 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**, 796 (2023).
4. Peres, C. A. & van Roosmalen, M. G. M. Avian dispersal of "mimetic seeds" of *Ormosia lignivalvis* by terrestrial granivores: deception or mutualism? *Oikos* **75**(2), 249–258 (1996).
5. Foster, M. S. & Delay, L. S. Dispersal of mimetic seeds of three species of *Ormosia* (Leguminosae). *J. Trop. Ecol.* **14**(4), 389–411 (1998).
6. Rudd, V. E. The American species of *Ormosia* (Leguminosae). *Contr. U.S. Natl. Herb.* **32**(5), 279–384 (1965).
7. Chang, R. H. A study on the Chinese *Ormosia* Jacks. *J. Syst. Evol.* **22**(1), 6–21 (1984).
8. 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 (2022).
9. Liu, P. P. *et al*. Genome Assemblies of Two *Ormosia* Species: Gene Duplication Related to Their Evolutionary Adaptation. *Agronomy* **13**, 1757 (2023).
10. Wu, G., Wei, X., Wang, X. & Wei, Y. Changes and transcriptome regulation of endogenous hormones during somatic embryogenesis in *Ormosia henryi* Prain. *Front Plant Sci.* **14**, 1121259 (2023).
11. Wang, J. *et al*. Integrative analysis of the metabolome and transcriptome reveals the molecular regulatory mechanism of isoflavonoid biosynthesis in *Ormosia henryi* Prain. *Int. J. Biol. Macromol.* **246**, 125601 (2023).
12. Feng, S., Hao, J., Xu, Z., Chen, T. & Qiu, S. X. S. Polyprenylated isoflavanone and isoflavonoids from *Ormosia henryi* and their cytotoxicity and anti-oxidation activity. *Fitoterapia* **83**(1), 161–165 (2012).
13. Lu, Y. *et al*. Phytochemical profile and antidepressant effect of *Ormosia henryi* Prain leaf ethanol extract. *Int. J. Mol. Sci.* **20**(14), 3396 (2019).
14. Zhou, C. *et al*. Genetic structure of an endangered species *Ormosia henryi* in southern China, and implications for conservation. *BMC Plant Biol.* **23**, 220 (2023).
15. Liu, H. *et al*. Genome comparison reveals mutation hotspots in the chloroplast genome and phylogenetic relationships of *Ormosia* species. *BioMed Res. Int*. (1), 7265030 (2019).
16. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Thompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**(5), 2320–2325 (2006).
17. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
18. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
19. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**(14), 2202–2204 (2017).
20. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with Hifiasm. *Nat. Methods* **18**, 170–175 (2021).
21. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**(9), 2896–2898 (2020).
22. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2008).
23. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
24. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes Aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* **356**, 92–95 (2017).
25. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**(1), 99–101 (2016).
26. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
27. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl_1), i351–i358 (2005).
28. Xu, Z. & Wang, H. LTR_Finder: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
29. Tempel, S. Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
31. Benson, G. Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
32. Stanke, M. *et al*. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
33. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
34. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**(1), 78–94 (1997).
35. Valliyodan, B. *et al*. Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.* **100**(5), 1066–1082 (2019).
36. Tang, H. *et al*. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**, 312 (2014).
37. Liang, Q. *et al*. A view of the pan-genome of domesticated Cowpea (*Vigna unguiculata* [L.] Walp.). *Plant Genome* **17**(1), e20319 (2024).
38. Hufnagel, B. *et al*. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nat. Commun.* **11**(1), 492 (2020).
39. Garg, G. *et al*. A pan-genome and chromosome-length reference genome of narrow-leafed lupin (*Lupinus angustifolius*) reveals genomic diversity and insights into key industry and biological traits. *Plant J.* **111**(5), 1252–1266 (2022).
40. Yang, J. *et al*. Chromosome-scale genomes of five Hongmu species in Leguminosae. *Sci. Data* **10**(1), 710 (2023).
41. Qu, Z., Wang, W. & Adelson, D. L. Chromosomal level genome assembly of medicinal plant *Sophora flavescens*. *Sci. Data* **10**(1), 572 (2023).
42. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**(1), btad014 (2023).
43. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods.* **12**(4), 357–360 (2015).

44. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**(3), 290–295 (2015).
45. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(23), 3150–3152 (2012).
46. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
47. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**(D1), D20–D26 (2022).
48. Coudert, E. *et al.* Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**(1), btac793 (2023).
49. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**(D1), D344–D354 (2021).
50. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
51. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**(D1), D274–D281 (2021).
52. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
53. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
54. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**(9), 1639–1645 (2009).
55. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22), 2933–2935 (2013).
56. Griffiths-Jones, S. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2004).
57. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**(5), 955–964 (1997).
58. *Genome Sequence Archive (GSA)* https://ngdc.cncb.ac.cn/gsa/browse/CRA020446 (2024).
59. *Genome Sequence Archive (GSA)* https://ngdc.cncb.ac.cn/gsa/browse/CRA020384 (2024).
60. *Genome Sequence Archive (GSA)* https://ngdc.cncb.ac.cn/gsa/browse/CRA020383 (2024).
61. *Genome Sequence Archive (GSA)* https://ngdc.cncb.ac.cn/gsa/browse/CRA020387 (2024).
62. *NGDC BioProject* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032243 (2024).
63. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_050613405.1 (2025).
64. *NGDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/86360/show (2024).
65. Zhou, C. Annotation files for *Ormosia henryi* genome assembly. *figshare* https://doi.org/10.6084/m9.figshare.27711531 (2024).
66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).

Acknowledgements

This work was supported by the Natural Science Foundation of Jiangxi province (Grant No. 20242BAB20256), the Research Project of Jiangxi Forestry Bureau (Grant No. 202306) and the Special Fund of Fundamental Research and Talents of Jiangxi Academy of Forestry (Grant No. 2023521202, 2024521201).

Author contributions

L.L. and T.O. conceived the project and led the research. C.Z., D.Z., C.G. and Z.G. were involved in sample collection and preparation. C.Z. and Q.W. contributed to the bioinformatics analysis. C.Z. drafted the manuscript. All authors read, revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05402-0.

Correspondence and requests for materials should be addressed to L.L. or T.O.