

# Chromosome-scale assemblies of three *Ormosia* species: repetitive sequences distribution and structural rearrangement

Zheng-Feng Wang<sup>1,2,3,4,\*†</sup>, En-Ping Yu<sup>1,2,3,4,5,†</sup>, Lin Fu<sup>1,3,4,6,†</sup>, Hua-Ge Deng<sup>7</sup>, Wei-Guang Zhu<sup>1,2,3,4</sup>, Feng-Xia Xu<sup>1,3,4,6</sup>, and Hong-Lin Cao<sup>1,2,3,4,\*</sup>

<sup>1</sup>Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

<sup>2</sup>Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

<sup>3</sup>Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

<sup>4</sup>South China National Botanical Garden, Guangzhou 510650, China

<sup>5</sup>University of Chinese Academy of Sciences, Beijing 100049, China

<sup>6</sup>Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China

<sup>7</sup>Management Office of Guangdong Luofushan Provincial Nature Reserve, Huizhou 516133, China

\*Correspondence address. Zheng-Feng Wang, Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Guangzhou 510650, China. E-mail: [wzf@scib.ac.cn](mailto:wzf@scib.ac.cn); Hong-Lin Cao, Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Guangzhou 510650, China. E-mail: [caohl@scib.ac.cn](mailto:caohl@scib.ac.cn)

†Joint first authors: Zheng-Feng Wang, En-Ping Yu, Lin Fu.

## Abstract

**Background:** The genus *Ormosia* belongs to the Fabaceae family; almost all *Ormosia* species are endemic to China, which is considered one of the centers of this genus. Thus, genomic studies on the genus are needed to better understand species evolution and ensure the conservation and utilization of these species. We performed a chromosome-scale assembly of *O. purpureiflora* and updated the chromosome-scale assemblies of *O. emarginata* and *O. semicastrata* for comparative genomics.

**Findings:** The genome assembly sizes of the 3 species ranged from 1.42 to 1.58 Gb, with *O. purpureiflora* being the largest. Repetitive sequences accounted for 74.0–76.3% of the genomes, and the predicted gene counts ranged from 50,517 to 55,061. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis indicated 97.0–98.4% genome completeness, whereas the long terminal repeat (LTR) assembly index values ranged from 13.66 to 17.56, meeting the “reference genome” quality standard. Gene completeness, assessed using BUSCO and OMArk, ranged from 95.1% to 96.3% and from 97.1% to 98.1%, respectively.

Characterizing genome architectures further revealed that inversions were the main structural rearrangements in *Ormosia*. In numbers, density distributions of repetitive elements revealed the types of Helitron and terminal inverted repeat (TIR) elements and the types of Gypsy and unknown LTR retrotransposons (LTR-RTs) concentrated in different regions on the chromosomes, whereas *Copia* LTR-RTs were generally evenly distributed along the chromosomes in *Ormosia*.

Compared with the sister species *Lupinus albus*, *Ormosia* species had lower numbers and percentages of resistance (R) genes and transcription factor genes. Genes related to alkaloid, terpene, and flavonoid biosynthesis were found to be duplicated through tandem or proximal duplications. Notably, some genes associated with growth and defense were absent in *O. purpureiflora*.

By resequencing 153 genotypes (~30 Gb of data per sample) from 6 *O. purpureiflora* (sub)populations, we identified 40,146 single nucleotide polymorphisms. Corresponding to its very small populations, *O. purpureiflora* exhibited low genetic diversity.

**Conclusions:** The *Ormosia* genome assemblies provide valuable resources for studying the evolution, conservation, and potential utility of both *Ormosia* and Fabaceae species.

**Keywords:** comparative genomics, gene families, gene duplication, genetic diversity, outlier loci, population genetics, RNA-seq, repeat-mediated chromosome architectures, SNP calling, structure variation

## Data description

### Context

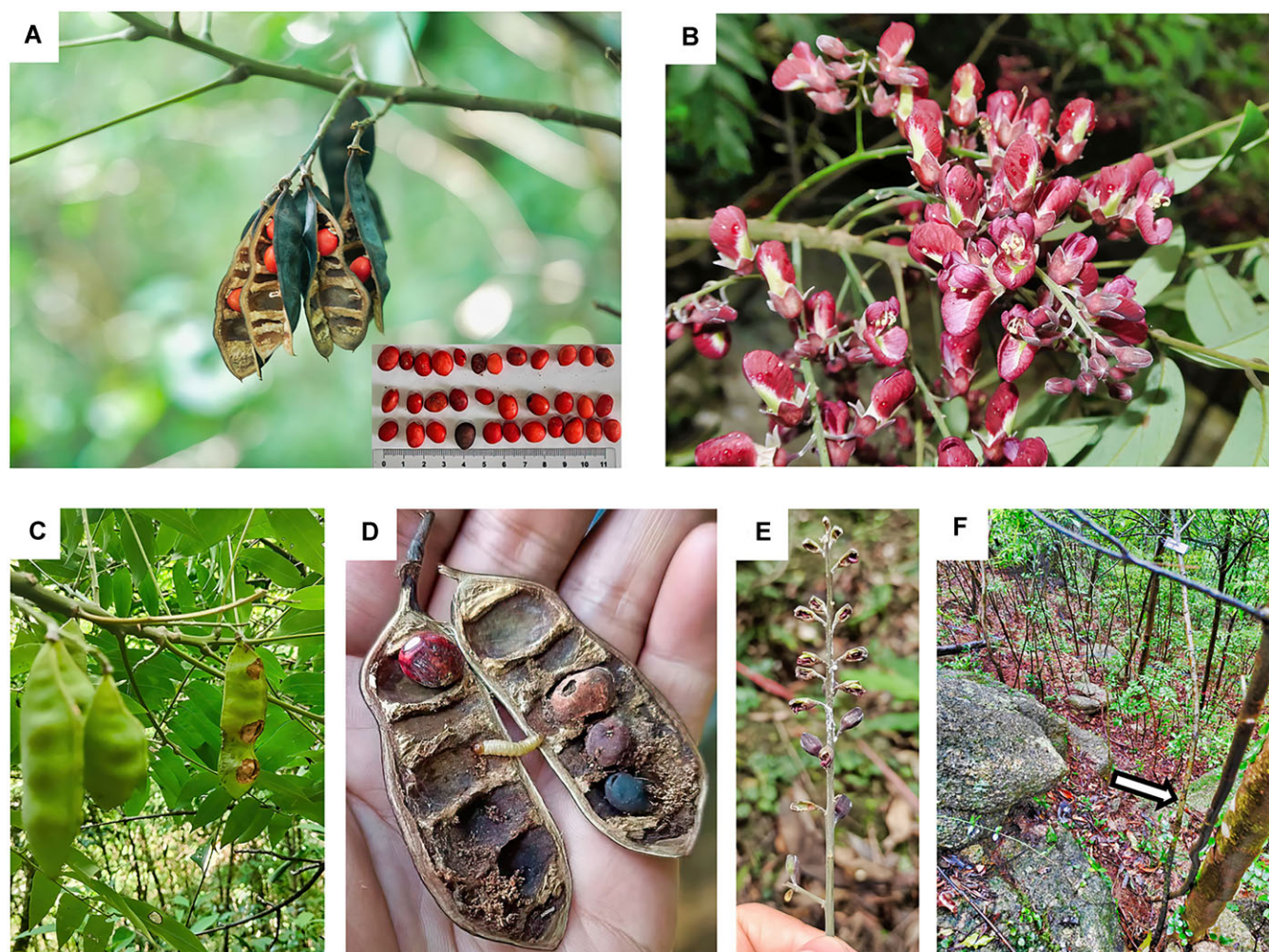
The genus *Ormosia* Jackson, belonging to the Fabaceae family, comprises approximately 130–150 species [1, 2]. These species are trees and shrubs that thrive in warm climates. Fossil records suggest that *Ormosia* species were originally distributed in northern regions of the Northern Hemisphere and migrated southwards during the Paleogene or Neogene period due to climate cooling [3].

Today, their distributions span tropical America, Southeast Asia, and northern Australia [1–3], following a typical Asian–American tropical disjunction pattern [1]. Continental Asia is widely considered the center of origin for the genus.

One of the most distinctive features of *Ormosia* species are their brightly colored seeds, including red, orange, bicolored red/orange, or black (Fig. 1A). These seeds are commonly used in ethnic jewelry and other decorative applications [1]. In addition, certain *Or-*

Received: August 24, 2024. Revised: December 12, 2024. Accepted: March 27, 2025

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1:** Photographs of *Ormosia purpureiflora*. (A) *O. purpureiflora* seeds. Seed sizes measured using a ruler are illustrated in the lower right panel. (B) *O. purpureiflora* flowers. (C) *O. purpureiflora* fruit in the distance showing a diseased state or insect invasion. (D) *O. purpureiflora* seeds showing invasion by worms/insects or a diseased state. (E) *O. purpureiflora* flowers in a diseased state or under insect invasion. (F) *O. purpureiflora* natural habitat; the arrow shows the sampled individual (a small shrub) used for genome assembly.

*mosia* species have high-value timber, and are cultivated as ornamental landscape trees [4, 5]. Extracts from their seeds, roots, stems, bark, and leaves have medicinal applications [6, 7] due to containing bioactive compounds such as alkaloids, flavonoids, isoflavones, terpenes, and lignans [6–8]. Metabolomic and transcriptomic analyses have further revealed that transcription factors play a key role in the regulation of flavonoid and terpenoid biosynthesis in *Ormosia* species [9, 10].

In China, approximately 37 species of *Ormosia* exist, and 34 of them are endemic [2]. *O. purpureiflora* is one such endemic species. Unlike most *Ormosia* species, which typically have white or yellow flowers, *O. purpureiflora* (NCBI:txid2866711) is characterized by its purple flowers (Fig. 1B), as reflected in its species name (*purpureiflora*). This species has been identified at only two locations in southeast China: the Guangdong Longmen Nankunshan (NKS) Provincial Natural Reserve and the Guangdong Luofushan (LFS) Provincial Natural Reserve. Field investigations have revealed approximately 2,000 individuals across these two sites [11]. A detailed survey of five plots (each 20 m × 20 m, 4 in LFS and 1 in NKS) recorded a total of 1,468 individuals. *O. purpureiflora* is a small shrub, with a diameter at breast height not exceeding 5 cm (average, 1.74 cm) and a height of no more than 4.5 m (average, 1.02 m). The species produces bisexual flowers in panicles or racemes and

it is primarily insect pollinated, with bees as the main pollinators. Although it exhibits typical outcrossing, self-pollination is also possible, as confirmed by bagging experiments [11]. Root-derived clonal reproduction has also been observed in the field [11]. The species is subject to severe pest and disease attacks, affecting both flowers and fruit (Fig. 1C–E), resulting in low seed yields in its natural habitat.

Advancements of high-throughput sequencing technologies have enabled the assembly of full genome information in diversified species, facilitating their conservation, restoration, management, and utilization. In this study, we generated a high-quality genome of *O. purpureiflora* by using a combination of long- and short-read whole genome sequencing (WGS), high-throughput chromosome conformation capture (Hi-C) sequencing, and RNA sequencing (RNA-Seq) of different tissues for annotation. In addition, we examined genetic diversity and conducted population genetics analyses for *O. purpureiflora* by resequencing 153 samples collected from 2 sites.

Genomic studies in *Ormosia* have primarily focused on its chloroplasts, with approximately 15 species, including *O. purpureiflora*, having been studied [12, 13]. To date, only 1 mitochondrial genome has been reported, which is from *O. boluoensis* [14]. Moreover, draft nuclear genomes have been reported



for only 2 species, *O. emarginata* (NCBI:txid53908) and *O. semicastrata* (NCBI:txid499992) [15]. According to phylogenetic studies in Fabaceae [16], *Ormosia* belongs to the Genistoid lineage, where it is most closely related to *Hovea* and *Poecilanthus*. However, no genomes have been reported for these two genera.

For comparative genomics, we updated the genome assemblies of previously published *O. emarginata* and *O. semicastrata* genomes [15] by using Hi-C data to generate chromosome-scale assemblies. Compared with *O. purpureiflora*, both *O. emarginata* and *O. semicastrata* are widespread species in southern and southeastern China, with *O. emarginata* extending into Vietnam. Unlike the small shrub *O. purpureiflora*, both *O. emarginata* and *O. semicastrata* are large trees. Phylogenetic analyses conducted by Torke et al. [1] placed *O. emarginata* and *O. semicastrata* in different clades, with *O. emarginata* belonging to the Old World *Ormosia* clade 1 and *O. semicastrata* in the Old World *Ormosia* clade 2. However, *O. purpureiflora* was not included in these phylogenies.

## Methods

### Chromosome number observation

The individual used for chromosome number observation in *O. purpureiflora* was regenerated from seeds collected at LFS. Its root tips were pretreated with 0.002 M 8-hydroxyquinoline for 6 h and then fixed in a 3:1 (v:v) solution of absolute ethanol and glacial acetic acid for 24 h at room temperature. After fixation, the root tips were transferred to 70% ethanol and stored at  $-4^{\circ}\text{C}$  until chromosome counts were performed. For chromosome counting, the fixed root tips were hydrolyzed in a 1:1 (v:v) solution of 1 M absolute ethanol and hydrochloric acid at room temperature for 7 min, rinsed with water, and then stained with carbol fuchsin for 4 min. Meristems were then excised and squashed for microscope observation. Photographs were taken using an Olympus BX-43 microscope (Olympus Corporation, TN, USA) at 100 $\times$  magnification with an Olympus DP26 camera (Olympus Corporation, TN, USA).

### Sample collection and sequencing

An *O. purpureiflora* individual (Fig. 1F) collected from LFS was used for genome assembly. Genomic DNA was isolated from its leaf tissues and multiple libraries were constructed, including long- and short-read WGS and Hi-C libraries. For gene annotation, RNA was extracted from the same individual used for genome assembly, specifically from its leaves, flowers, seeds, and fruit. RNA-seq libraries were then constructed for these tissues. Long-read WGS was performed using the Oxford Nanopore Technologies (ONT) PromethION sequencer (Oxford Nanopore Technologies plc, Oxford, UK). Both long-read and ultra-long-read (50 kb) sequencing libraries were generated on the ONT platform. Short-read WGS, Hi-C, and RNA-seq were conducted using an MGI DNBSEQ-T7 (MGI Tech Co., Ltd., Shenzhen, China) sequencer with a 150 bp paired-end sequencing strategy (insert size, 300 bp). Given the relatively high error rate of ONT reads, the error profile of the ONT data was estimated using SeqFaiLR (Tools To Analyse Long Reads Sequencing Error Profile) [17].

For population genetic studies on *O. purpureiflora*, leaf samples were collected from 153 individuals representing 6 (sub)populations in LFS and NKS (Table 1, Fig. 2A). These individuals were randomly selected to ensure that their distribution covered the entire range of the two sites based on a thorough field investigation [11]. The geographical positions of the sampled individuals were recorded using a handheld GPS. Leaves from each individual were immediately placed into sealed plastic bags

containing silica gel for preservation. Whole-genome resequencing was conducted using an MGI DNBSEQ-T7 sequencer with a PE-150 bp model, generating approximately 30 Gb of data per sample.

Leaf and flower RNA-seq libraries were constructed and sequenced by Annoroad Gene Technology (AGT, Beijing, China), and the remaining libraries were constructed and sequenced by GrandOmics Biosciences (GB, Wuhan, China).

For *O. emarginata* and *O. semicastrata*, Hi-C libraries were constructed using leaf samples from the same individuals used in their primary genome assemblies [15], and sequencing was performed by GrandOmics Biosciences.

Detailed sequencing information, including DNA/RNA preparation and library construction (except for ONT ultra-long WGS sequencing), has been reported in our previous studies [18, 19]. Specifically, ONT long-read WGS sequencing was performed using the protocol described by Wang et al. [18], whereas short-read WGS, Hi-C, and RNA-seq sequencing were conducted using protocols described by Wang et al. [19]. For *O. purpureiflora* ONT ultra-long WGS sequencing, genomic DNA was extracted, and approximately 8–10  $\mu\text{g}$  of DNA fragments longer than 50 kb were selected using the SageHLS HMW library system (Sage Science, Inc., MA, USA). The size-selected DNA was repaired using the NEBNext FFPE DNA Repair Mix (catalog no. M6630, New England Biolabs, MA, USA) in accordance with the manufacturer's instructions. End-repair and dA-tailing were then performed using the NEBNext Ultra II End-Repair/dA-tailing Module (catalog no. E7546, New England Biolabs). Ligation of adaptors was performed by adding Adaptor Mix (SQK-LSK114, Oxford Nanopore Technologies, Oxford, UK). The adaptor-ligated DNA was cleaned and quantified using fluorometry (Qubit 3.0, Thermo Fisher Scientific Inc., MA, USA) before library construction. The final library was sequenced on the Nanopore PromethION platform using an R10.4.1 flowcell (Oxford Nanopore Technologies). Basecalling was performed using Dorado (RRID) v. 0.3.4 [20].

For genome assembly, annotation, and comparative genome analysis, default parameters were used in all programs unless otherwise specified.

### Date preprocessing

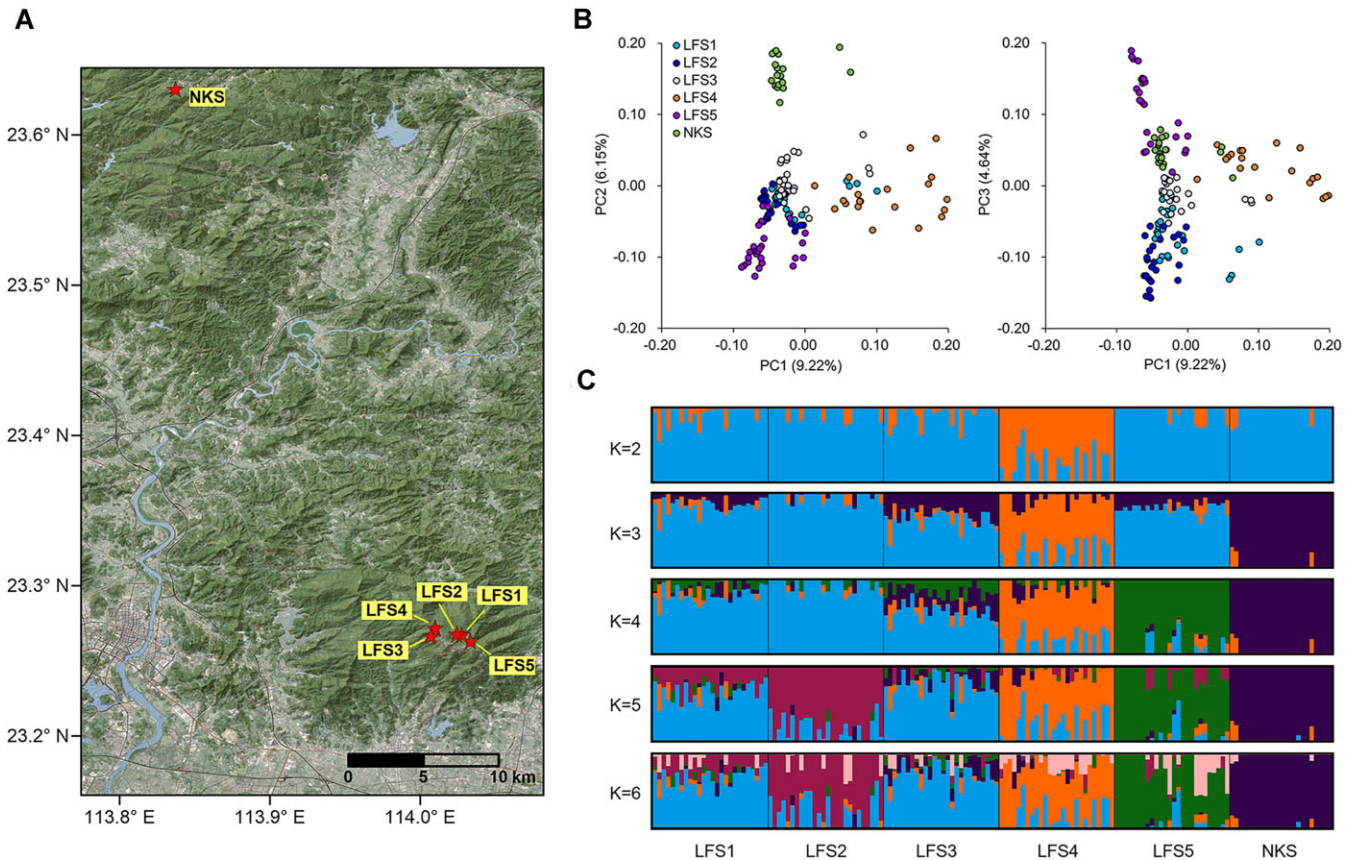
Short WGS reads of *O. purpureiflora* and Hi-C reads from all 3 *Ormosia* species were quality-trimmed using Sickle v. 1.33 (RRID:SCR\_006800) [21]. Reads with base quality values below 30 or lengths shorter than 80 bp were removed. The WGS reads of *O. purpureiflora* were further error corrected using RECKONER v. 1.1 [22]. Based on the error-corrected reads, 21-mer frequencies were generated using Jellyfish 2.3.0 (RRID:SCR\_005491) [23], and the results were analyzed using GenomeScope 2.0 (RRID:SCR\_017014) [24] to estimate the genome size, heterozygosity, and repetitiveness of *O. purpureiflora*. The ploidy level of *O. purpureiflora* was determined using nQuire with the “lrdmodel” function [25]. For ONT (ultra-)long WGS reads of *O. purpureiflora*, adaptors were removed using Porchop 0.2.4 [26]. ONT reads larger than 20 kb were then extracted from the full dataset and defined as the 20 kb ONT read set, which was subsequently used for *O. purpureiflora* genome assembly.

### Genome assembly

Using the 20 kb ONT read set, the *O. purpureiflora* genome was assembled using NextDenovo 2.3.1 [27]. After assembly, Pseudohaploid [28] and Purge\_Dups v. 1.2.6 (RRID:SCR\_021173) [29] were used to identify and remove duplications resulting from heterozygosity. The assembly was then polished sequentially by Racon v. 1.5.0

**Table 1:** Six sampled (sub)populations and their genetic diversities in *Ormosia purpureiflora*

(Sub)population	Sample size	Ho	He	Fis	$\pi$
LFS1	26	0.171	0.148	-0.122	0.148
LFS2	26	0.154	0.130	0.000	0.128
LFS3	26	0.165	0.146	-0.083	0.147
LFS4	26	0.196	0.154	-0.268	0.152
LFS5	26	0.146	0.122	0.058	0.121
NKS	23	0.153	0.122	0.011	0.121



**Figure 2:** Sampled (sub)populations and population genetics of *Ormosia purpureiflora*. (A) Map showing (sub)populations sampled for *Ormosia purpureiflora*. (B) Principal component analysis (PCA) results showing the first 3 PCs (PC1 versus PC2; PC1 versus PC3) for *O. purpureiflora* individuals sampled from different (sub)populations. (C) Admixture results representing data for  $K = 2-6$  clusters.

(RRID:SCR\_017642) [30] (run twice), Hapo-G v. 1.3.2 [31] (run twice), and Polypolish v. 0.5.0 [32]. Depthcharge v. 0.2.0 [33] was applied to correct potential misassemblies, and contigs shorter than 1,000 bp were removed. The corrected assembly was scaffolded using Hi-C reads with Scaffhic v. 1.1 [34], the Juicer pipeline 1.6 (RRID:SCR\_017226) [35], and 3d-dna 201,008 (RRID:SCR\_017227) [36]. Gaps in the scaffolded assembly were closed with TGS-GapCloser v. 1.2.1 (RRID:SCR\_017633) [37]. The gap-closed assembly was polished again using Racon, Hapo-G, and Polypolish. Redundans 0.14a [38] was used to remove redundant sequences unanchored to chromosomes. The assembly was then uploaded to GenBank to check for possible contamination. Sequences identified as bacterial and fungal contaminants were removed. Subsequently, telomeric repeats at each chromosome end were identified (with the parameter of “-motifs TTTAGGG -matchAny”) and recovered using Teloclip v. 0.0.3 [39]. The assembly was then polished by Racon, Hapo-G, and Polypolish to produce a complete genome assem-

bly. To evaluate the assembly completeness, Benchmarking Universal Single-Copy Orthologs (BUSCO) v. 5.5.0 (RRID:SCR\_015008) [40] was applied using the eudicots\_odb10.2020-09-10 database, which contains 2,326 conserved eudicot core genes. Assembly quality was further assessed using AssemblyQC v. 2.1.1 [41] and GAEP v. 1.2.3 [42]. AssemblyQC provided metrics such as the long terminal repeat (LTR) Assembly Index (LAI) [43], which evaluates contiguity based on repetitive sequences, as well as  $k$ -mer-based assembly completeness [44]. GAEP offered mapping-based evaluations, reporting read mapping ratios for various read types (long WGS, short WGS, and RNA-seq reads) and a consensus quality value (QV) for overall mapping accuracy.

For *O. emarginata* and *O. semicastrata*, the primary assemblies [15] were upgraded to chromosome-scale assemblies following the *O. purpureiflora* procedures, starting from Hi-C scaffolding. The final assemblies were evaluated for quality by using the same methods applied to *O. purpureiflora*.

## Repeat sequence and gene prediction

The repeat sequences in 3 *Ormosia* chromosome-scale assemblies were identified using both EDTA v. 2.1.0 (RRID:SCR\_022063) [45] and RED v. 2.0 [46]. The results from both programs for each assembly were combined and used to soft-mask the corresponding assembly with Bedtools v. 2.29.2 (RRID:SCR\_006646) [47] using the commands “merge” and “maskfasta.” An explanation of the merging procedure is shown in [Supplementary Fig. S1](#). The densities of repetitive elements measured by length proportion (percentage of sequence coverage on chromosome) and number per 10<sup>5</sup> or 10<sup>6</sup> bp on the chromosomes were then calculated with Circlize v. 0.4.15 [48] under the parameter of “overlap=FALSE.”

The soft-masked *Ormosia* assemblies were annotated using BRAKER2 v. 2.0 [49] and the Funannotate pipeline v. 1.8.16 [50]. BRAKER2 utilized RNA-seq reads and reference proteins from eight species ([Supplementary Table S1](#)) for transcriptome- and homology-based annotation, except for ab initio-based gene prediction. The results from BRAKER2 were integrated using Funannotate to generate consensus gene sets. Gene prediction in Funannotate followed 3 steps: “train,” “predict,” and “update.” For the “predict” and “update” steps, the parameters “-max\_intronlen 100,000 -busco\_db embryophyta -organism other” were applied. Function annotation of predicted genes in *Ormosia* species was performed using Funannotate with the “annotate” command. The annotation databases included dbCAN v. 10.0 (RRID:SCR\_013208) [51], EggNOG v. 5.0.2 (RRID:SCR\_002456) [52], Gene Ontology (GO, RRID:SCR\_002811) [53, 54], Kyoto Encyclopedia of Genes and Genomes (KEGG, RRID:SCR\_012773) [55], InterPro v. 5.62-94.0 (RRID:SCR\_006695) [56], MEROPS v. 12.0 [57] (RRID:SCR\_007777), Pfam v. 35.0 [58] (RRID:SCR\_004726), SignalP 5.0b (RRID:SCR\_015644) [59], and UniProt v. 2023\_02 (RRID:SCR\_002380) [60].

The completeness of the predicted genes was initially evaluated using BUSCO with the eudicots\_odb10.2020-09-10 database, analyzing the longest transcripts from each *Ormosia* assembly. In addition, prediction quality was assessed using the online tool OMArk v. 0.3.0 [61]. Unlike BUSCO, which focuses solely on conserved single-copy genes, OMArk evaluates completeness based on conserved genes in both single and multiple copies. It also examines the consistency of the predicted genes relative to closely related species (e.g., the proportion of genes in the same lineage) and identifies potential contamination events. Finally, the completeness of the predicted genes was examined against 15,345 representative gene models from 12 Fabaceae species [62]. For this analysis, the gene models of each comparative species ([Supplementary Table S2](#)) were matched to the representative genes by using blastp (RRID:SCR\_004870) v. 2.13.0 [63] with the parameters “-evalue 1e-2 -outfmt 6 -num\_threads 96 -max\_hsp 5 -max\_target\_seqs 5.”

For the comparative genomic analyses, only the longest transcript for each gene across all species was used, unless stated otherwise. Additionally, for genome comparisons, the protein-coding genes of all other species used in our phylogenetic analysis (see below) were functionally annotated following the same procedures applied to the *Ormosia* species.

## Gene family and comparative genomics

Orthologous groups (gene families) in *Ormosia* were identified using OrthoFinder 3.0.0 (RRID:SCR\_017118) [64, 65], with protein-coding gene sequences from 17 other species ([Supplementary Table S2](#)) as inputs. Phylogenetic analysis was subsequently performed using 1,131 single-copy orthologs inferred using OrthoFinder, employing STAG [66] and STRIDE [67], which are inte-

grated within OrthoFinder. The gene family file generated using OrthoFinder was further analyzed to assess gene family expansion or contraction using CAFE v. 5 (RRID:SCR\_018924) [68]. The species tree, along with the divergence time required for CAFE analysis, was constructed using MCMCTree [69], with 12 calibration points from the TimeTree database ([Supplementary Table S3](#)) for calibration. Following the CAFE analysis, GO and KEGG enrichment analyses were performed on the significantly expanded and contracted gene families in *O. purpureiflora* using TBtools v. 2.030 [70].

## Gene duplications, synteny, and structural variation analysis

Ancient whole genome duplication (WGD) events in *Ormosia* and their sister species *Lupinus albus* (see Results in “Gene family” section) were identified using wgd v. 1.1.2 [71]. Gene duplications in *Ormosia* were analyzed using Doubletrouble v. 0.99.1 [72], which classified the duplication origin into categories including WGD, tandem duplications, proximal duplications, transposed duplications, and dispersed duplications [73]. In this analysis *L. albus* was used as an outgroup species. For genes resulting from WGD, tandem, and proximal duplications in *O. purpureiflora*, GO and KEGG enrichment analyses were performed using TBtools.

Syntenic regions within and between *Ormosia* and *L. albus* genome assemblies were identified using MCScanX [74] and visualized using Shinycircos [75] or SynVisio [76]. The parameter of “-s 30” (MATCH\_SIZE) was used for synteny analysis in MCScanX. Structural variations were identified using chromeister v. 1.5.a [77] and plotlr v. 1.1.0 [78].

## Identification of nucleotide-binding leucine-rich repeats and other resistance genes

Nucleotide binding leucine-rich repeats (NLR) genes are the primary plant resistance (R) genes that protect against viruses, bacteria, nematodes, fungi, oomycetes, and insects [79, 80]. These genes typically consist of 3 canonical domains: a variable N-terminal domain, a central nucleotide-binding domain (NB-ARC), and a C-terminal domain composed of leucine-rich repeats (LRRs) [81]. At the N terminus, 3 types have been identified: Toll/interleukin-1 receptor (TIR), coiled-coil (CC), and resistance to powdery mildew8 (RPW8) [82]. The InterPro/Pfam entries associated with these domains include NB-ARC (IPR002182/PF00931), TIR (IPR000157/PF01582/PF13676), CC (IPR038005), RPW8 (IPR008808/PF05659), and LRR (IPR001611/PF00560, IPR013101/PF07723, IPR011713/PF07725, IPR025875/PF12799, IPR026906/PF13306, IPR001611/PF13516/PF13855, PF14580 and IPR032675). In addition to the NLR genes, other R genes were identified based on their InterPro entries, as described by De-la-Cruz et al. [83]. Using gene annotation results from all species (3 *Ormosia* species and the 17 comparative species listed in [Supplementary Table S2](#)), obtained using the “annotate” command from the Funannotate pipeline, the InterPro/Pfam entries of their genes were matched to the corresponding R gene entries. The types and statistics of R genes were subsequently categorized for each species.

Because the above searches were mainly based on the InterPro and Pfam databases, both were generalized domain annotation tools and might provide overlapping or fragmented annotations, leading to inaccurate results in R gene identification. Therefore, NLR genes, in *Ormosia* and the other compared species in our phylogenetic analysis, were also identified by Resistify v. 1.1.5 [84],



which could accurately and extensively identify and classify them by integrating different programs and more filtering steps.

## Transcription factor

Transcription factor (TF) genes in the genomes of *Ormoisa* species and the other species in our phylogenetic analysis were identified by the TF prediction online tool PlantTFDB v. 5.0 [85]. TF genes were also predicted by TransFacPred [86], which combined alignment-free (machine learning method) and alignment-based (BLAST method) methods to achieve high accuracy.

## Single-nucleotide polymorphism calling

Single-nucleotide polymorphisms (SNPs) in 153 *O. purpureiflora* individuals were identified using NGSEP (RRID:SCR\_012827) v. 5.0.0 [87] with the *O. purpureiflora* genome assembled in this study serving as the reference. The mapping results used for this procedure were from BWA v. 0.7.17-r1188 [88]. For NGSEP, the parameters `-h 0.00952 -maxAlnsPerStartPos 2` were used, with all other settings remaining at their defaults. The `-h` parameter specifies the heterozygosity rate, which was derived from the GenomeScope results (see Results in "Genome assembly" section). The raw SNPs called by NGSEP were quality-filtered using VCFtools (RRID) v. 0.1.17 [89], with the parameters of `"-max-missing 0.95 -maf 0.05 -recode -recode-INFO-all -min-meanDP 20 -mac 3 -minQ 30 -non-ref-af 0.001 -max-non-ref-af 0.9999"`. Filtered SNPs were further processed to remove the SNPs deviating from Hardy-Weinberg equilibrium (HWE) and the InDels. Departure from HWE can cause genotyping errors due to the presence of null alleles, sequence duplication, copy number variation, and other sequencing problems related to read depth. The HWE filtering was performed using the script "filter\_hwe\_by\_pop.pl" from the "SNP Filtering Tutorial" [90] with the parameter `"-c 0."`

To accurately infer population genetic diversity and structure, SNPs called by NGSEP were further filtered to remove those in linkage disequilibrium (LD) by using Plink (RRID) v. 1.90p [91–93]. Specifically, SNP loci with an LD association coefficient ( $r^2$ ) greater than 0.2 were excluded. Finally, outlier SNPs (potentially under selection) were identified and removed from the dataset used for genetic diversity and structure inferences. These SNPs were detected using PCAdapt v. 4.3.5 [94, 95] and BayPass v. 2.4 [96].

For PCAdapt, a principal component analysis (PCA) was first performed, and a Scree plot was used to determine the optimal number of PCs for regression with each SNP. Following regression analysis, SNPs with a  $q$  value (adjusted  $P$  value) of  $<0.01$  were considered outliers. For BayPass, the core model with default parameters was applied. This model estimated an  $F_{ST}$ -like XtX statistic while accounting for the variance-covariance structure. To determine significance, a calibrated threshold (99%) was established by simulating pseudo-observed datasets (100,000 SNPs). SNPs falling within the 99.9% quantile of the pseudo-observed XT distribution were considered outliers. Outlier SNPs were identified as those occurring in both PCAdapt and BayPass results.

## Genetic diversity and genetic structure

Genetic diversity parameters, including observed heterozygosity ( $H_o$ ), expected heterozygosity ( $H_e$ ), and inbreeding coefficient ( $F_{is}$ ), were estimated by VCFtools. Nucleotide diversity within populations ( $\pi$ ), nucleotide divergence between populations ( $d_{xy}$ ), and pairwise genetic differentiation ( $F_{st}$ ) were calculated using pixy v. 1.2.7.beta1 [97].

The genetic structure of *O. purpureiflora* was inferred through PCA and ADMIXTURE (RRID:SCR\_001263) [98]. PCA was performed

using SNPRelate v. 1.36.0 [99], and ADMIXTURE was conducted using the AdmixPipe v. 3.2 pipeline [100]. In AdmixPipe, the number of potential genetic groups ( $K$ ) was tested from 1 to 6, with 20 replicates for each  $K$  value. The best  $K$  value was determined based on cross-validation (CV) errors. For the inferred  $K$ , CLUMPAK v. 1.1 [101] was used to estimate the mean membership coefficients for individuals across the 20 replicates.

## Results

### Chromosome number

The ploidy level estimated using nQuire indicated that the *O. purpureiflora* genome is diploid because the diploid model showed a lower delta likelihood than the free model (diploid delta likelihood: 1,609,982.99; triploid delta likelihood: 2,029,931.52; tetraploid delta likelihood: 2,270,455.24) (Supplementary Table S4). Similar estimations for *O. emarginata* and *O. semicastrata* confirmed that these species also have diploid genomes.

The chromosome number of *O. purpureiflora* was determined to be  $2n = 16$  (Supplementary Fig. S2), consistent with the number reported previously in *O. macrocalyx* [102] and *O. arborea* [103].

### Genome sequencing

For *O. purpureiflora*, the ONT sequencing platform generated approximately 181.6 Gb of WGS reads, including 51.3 Gb of ultra-long reads. The short sequencing platform produced approximately 139.3 Gb WGS reads and 146.8 Gb Hi-C reads. RNA-seq data amounted to approximately 20.4 Gb, 21.9 Gb, 23.3 Gb, and 25.3 Gb for leaf, flower, fruit, and seed samples, respectively. For *O. emarginata* and *O. semicastrata*, 148.7 Gb and 123.6 Gb Hi-C reads were generated, respectively.

### Genome assembly

For *O. purpureiflora*, the genome size estimated using GenomeScope was 1,503,292,231 bp, with repetitive sequences accounting for 66.6% of the genome and a heterozygosity rate of 0.952% (Supplementary Fig. S3). The initial genome assembly size was 1,811,176,403 bp, comprising 313 contigs with an N50 of 50,908,349 bp. After redundancy removal, Hi-C scaffolding and gap closing, the final assembly measured 1,584,128,722, with 1,583,483,254 bp (99.96%) anchored to 8 chromosomes (Table 2, Fig. 3A), consistent with chromosome number observation (Supplementary Fig. S2). The longest chromosome was 259,935,025 bp long, and the shortest was 121,398,155 bp.

The initial assemblies for *O. emarginata* and *O. semicastrata* were 1,420,917,605 bp and 1,511,766,959 bp, respectively [15]. GenomeScope estimations using a  $k$ -mer size of 21 revealed repeat contents of 65.5% and 63.4%, and heterozygosity rates of 2.29% and 2.05% for *O. emarginata* and *O. semicastrata*, respectively. Both species displayed higher heterozygosity than *O. purpureiflora*, although the repeat content was similar across the 3 species. After incorporating Hi-C data, the assembly sizes of *O. emarginata* and *O. semicastrata* were refined to 1,420,253,666 and 1,510,687,319 bp, respectively (Table 2). Each assembly achieved 8 chromosome-level scaffolds, accounting for 99.99% and 99.97% of the total in *O. emarginata* and *O. semicastrata*, respectively.

For *O. purpureiflora*, BUSCO evaluation revealed 98.3% complete BUSCOs, with 89.4% of them being single-copy BUSCOs and 8.9% being duplicated BUSCOs. In addition, 0.3% BUSCOs were fragmented, and 1.4% were missing. For *O. emarginata*, the complete BUSCO score was 97.0%, including 89.4% complete and single-copy BUSCOs and 7.6% complete but duplicated BUSCOs,

**Table 2:** Statistics and evaluations of genome assemblies for three *Ormosia* species

Species	<i>O. purpureiflora</i>	<i>O. emarginata</i> <sup>a</sup>	<i>O. semicastrata</i> <sup>a</sup>
Initial assembly statistic (bp)			
N10	122,192,683	81,285,628	89,031,100
N20	120,000,233	63,464,384	79,796,434
N30	75,858,835	43,593,171	73,253,298
N40	61,354,201	37,463,220	56,807,054
N50	50,908,349	28,195,512	48,976,089
N60	45,450,924	25,800,464	45,239,136
N70	36,587,725	20,527,781	31,722,207
N80	15,728,371	13,438,452	22,051,163
N90	3,163,854	7,895,810	12,933,450
N100	34,487	173,104	128,272
Total length	1,811,176,403	1,420,917,605	1,511,766,959
Average length	5,786,506.08	15,787,973.39	23,996,300.94
Largest length	142,757,542	84,853,091	144,833,628
Minimum length	34,487	173,104	128,272
Number of contigs	313	90	63
Assembly after applying Hi-C data (bp)			
chr1	259,935,025	199,918,031	205,218,018
chr2	233,292,245	210,768,611	211,883,283
chr3	229,093,642	183,696,964	200,464,886
chr4	212,222,348	180,298,008	178,099,194
chr5	195,349,128	202,609,791	205,007,630
chr6	187,433,795	149,243,870	185,806,757
chr7	144,758,916	145,867,561	164,432,676
chr8	121,398,155	147,815,325	159,254,978
Unanchored to chromosome	645,468	35,505	519,897
Total length	1,584,128,722	1,420,253,666	1,510,687,319
Assembly quality assessed by AssemblyQC			
LAI	16.08	13.66	17.56
k-mer based assessment			
Completeness	88.36%	78.04%	81.15%
QV	28.83	27.02	28.34
Assembly quality assessed by GAEP			
GC content	35.06%	34.53%	34.63%
Mapping based assessment			
QV	39.74	37.38	38.46
Long WGS reads mapping ratio	97.32% <sup>b</sup>	99.72% <sup>c</sup>	99.59% <sup>c</sup>
Short WGS read mapping ratio <sup>d</sup>	99.76%	98.31%	98.43%
RNA-seq mapping ratio <sup>e</sup>			
Leaf	95.77%	93.52%	95.24%
Flower	91.15%	—	—
Fruit	95.96%	—	—
Seed	92.23%	—	—

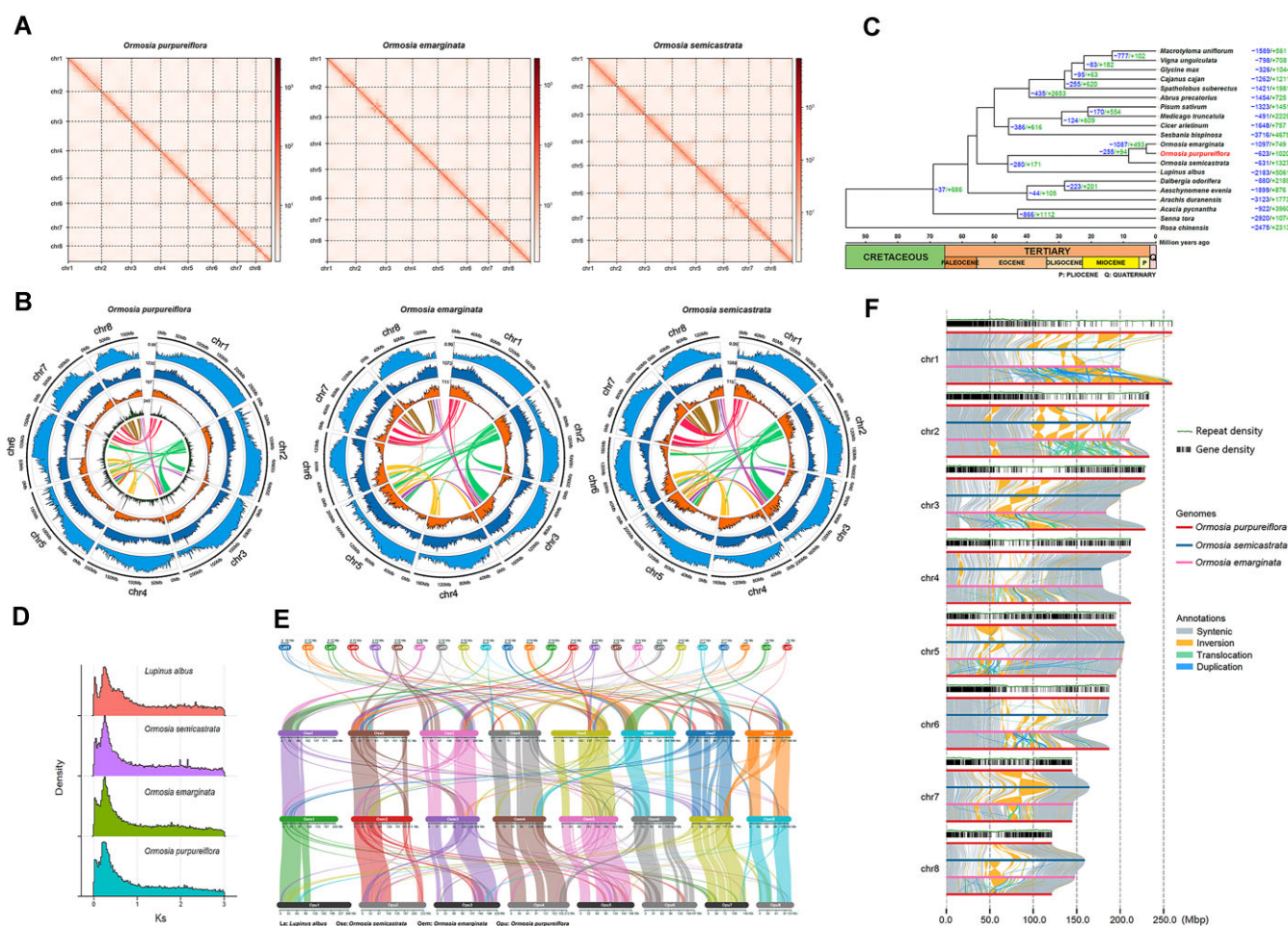
<sup>a</sup>From Liu et al. [15].<sup>b</sup>With reads longer than 20 kb<sup>c</sup>With reads longer than 10 kb.<sup>d</sup>Trimmed and error-corrected.<sup>e</sup>Trimmed using trimmomatic (RRID:SCR\_011848) v. 0.39 [121] with parameter "SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25."

with 0.5% fragmented and 2.5% missing. Similarly, *O. semicastrata* achieved a complete BUSCO score of 98.4%, consisting of 90.4% complete and single-copy BUSCOs and 8.0% complete but duplicated BUSCOs, with fragmented and missing BUSCOs constituting 0.1% and 1.5%, respectively.

The LAI values for the 3 *Ormosia* assemblies were all above 10, ranging from 13.66 to 17.56 (Table 2), meeting the quality standard for a "reference genome." Mapping-based evaluations indicated that all types of reads achieved high mapping ratios, exceeding 91%. However, the k-mer and mapping-based quality value (QV) scores in all 3 *Ormosia* assemblies were below 40, a threshold that corresponds to 99.99% base accuracy and is considered high quality for genome assemblies [42, 44]. In addition, k-mer-based completeness was below 90% across all *Ormosia* assemblies, with *O. purpureiflora* having the highest completeness at 88.36% and *O.*

*emarginata* having the lowest at 78.04%. These assessments suggest that further improvements in the assemblies are warranted.

Assembling genomes with a large size (>1 Gb), high repeat content (>50%), and elevated heterozygosity (>0.5%) presents significant challenges [104], and the species examined in this study exhibited all these features. To address these issues, we used the Nanopore sequencing platform, which generates reads that are longer in length than those produced by the PacBio sequencing platform, particularly in Hi-Fi sequencing mode [105]. For *O. purpureiflora*, we included ultra-long reads (>50 kb) to enhance assembly continuity. In the assembly process, reads longer than 20 kb were used for assembling the *O. purpureiflora* genome, whereas reads longer than 10 kb were used for assembling the genomes of *O. emarginata* and *O. semicastrata*. Programs such as Pseudohaploid and Purge\_Dups were used to remove heterozy-



**Figure 3:** *Ormosia* genomes and comparative genomics. (A) Hi-C interaction heat maps (bin length, 100,000 bp) for the genome assemblies of three *Ormosia* species. (B) Circos plot showing the genome features (chromosome, repeat density in length proportions, repeat density in numbers, gene density, and syntenic blocks from outer to inner) across chromosomes of the genome assemblies of three *Ormosia* species. Repeat densities in each Circos plot was quantified by all repetitive elements. For *O. purpureiflora*, the Circos plot also includes SNP density results between the results of gene density and syntenic blocks. All densities were estimated using a 1 Mbp sliding window. (C) The inferred phylogenetic tree, divergence time, and contracted (-) and expanded (+) gene families in *O. purpureiflora* and other species. (D) The density distribution of synonymous nucleotide substitutions (Ks) in the whole genome duplication analysis for *Ormosia* species and their sister species, *L. albus*. (E) Syntenic blocks among *Ormosia* species and *L. albus*. (F) Intrachromosomal structural variations observed among the three *Ormosia* species.

gous contigs and regions, effectively mitigating challenges associated with high repeat content and heterozygosity in these genomes. However, ONT reads generally have high sequencing error rates, ranging from 5% to 20% [105–107]. For *O. purpureiflora*, the error rate of ONT reads was 15.18%, as assessed using the 20 kb ONT read set used for genome assembly (see Methods in "Genome assembly" section). Similarly, error rates of 17.75% and 16.82% were observed in the 10 kb ONT read sets of *O. emarginata* and *O. semicastrata*, respectively. Given these limitations, haplotype-resolved de novo genome assembly was not performed for the 3 *Ormosia* species. Future studies should incorporate highly accurate Hi-Fi long-read sequencing technology and phasing steps to optimize the current assemblies and improve their overall quality.

### Repeat and gene annotation

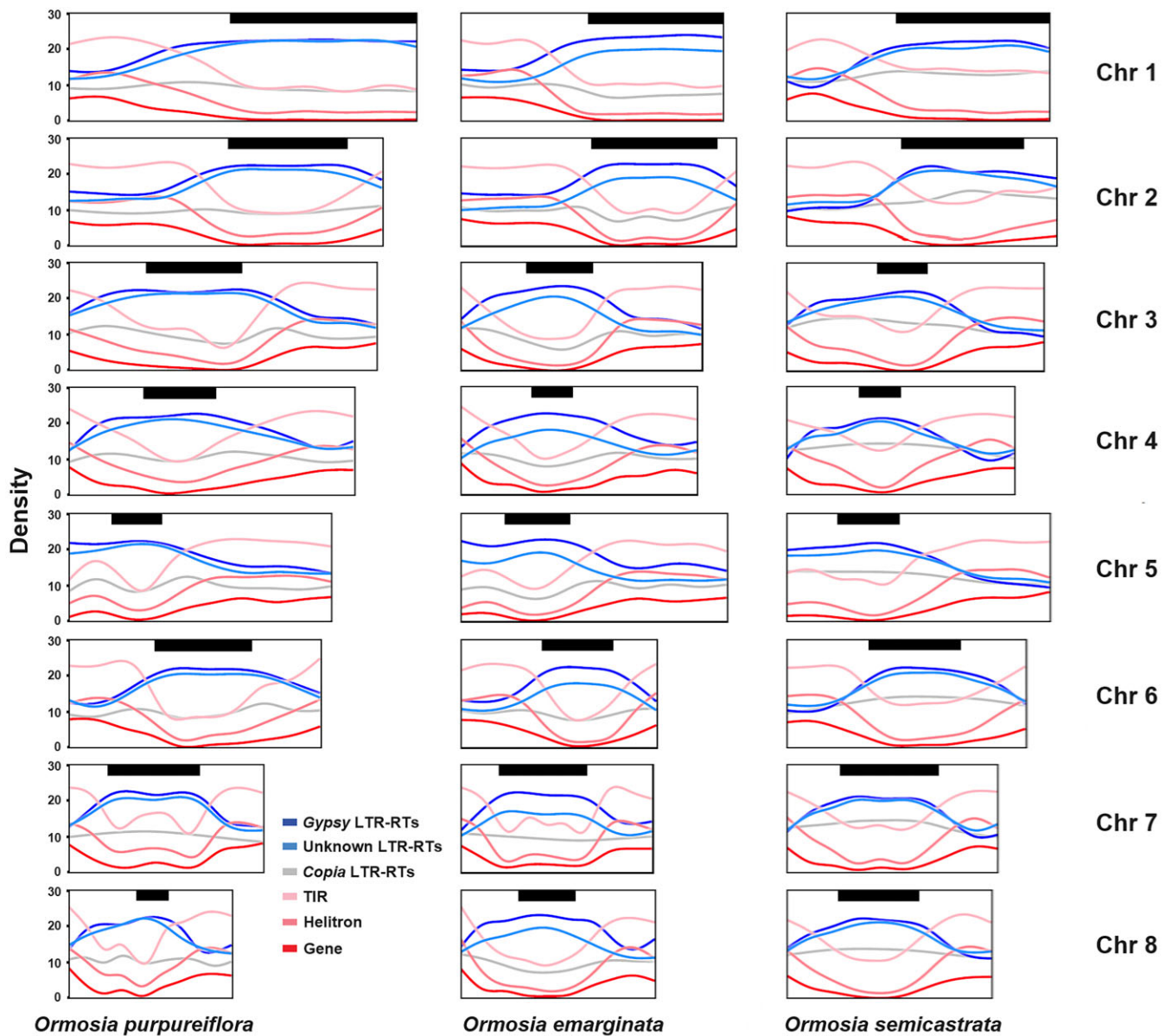
RED analyses identified 1,037,006,095 bp (65.5%), 885,912,252 bp (62.4%), and 968,176,023 bp (64.1%) of repetitive sequences in *O. purpureiflora*, *O. emarginata*, and *O. semicastrata*, respectively. EDTA analyses revealed higher percentages, that is, 1,139,417,595 bp (71.9%), 989,514,254 bp (69.6%), and 1,074,353,470 bp (71.1%) of repetitive sequences in *O. purpureiflora*, *O. emarginata*, and *O. semicastrata*, respectively.

*icastrata*, respectively (Supplementary Table S5). After combining the results from RED and EDTA, the total repetitive components were found to be 1,209,324,791 bp (76.3%) in *O. purpureiflora*, 1,051,218,280 bp (74.0%) in *O. emarginata*, and 1,135,447,010 bp (75.2%) in *O. semicastrata*. According to EDTA analyses, the Gypsy-like long terminal repeat retrotransposon (LTR-RT) family represents the most abundant repetitive sequence, comprising 33.51%, 35.45%, and 27.73% of the genome assemblies for *O. purpureiflora*, *O. emarginata*, and *O. semicastrata*, respectively.

Overall, a "complementary" pattern was observed in the density distributions of repetitive elements between their length proportions and numbers on the chromosomes in *Ormosia*, namely two types of distribution concentrated on different parts of the same chromosomes (Fig. 3B). In numbers, a closer examination revealed that the types of Helitrons and terminal inverted repeats (TIRs) generally distributed disparately from the types of Gypsy and unknown LTR-RTs (Fig. 4 and Supplementary Fig. S4), whereas the other LTR-RT type, *Copia*, was generally evenly distributed along the chromosomes in *Ormosia*.

Gene prediction identified 55,061 genes encoding 59,809 proteins in *O. purpureiflora*. For *O. emarginata* and *O. semicastrata*, the





**Figure 4:** Smoothing lines for gene and repeat density distributions (bin size, 100,000 bp) along chromosomes in *Ormosia* species. The repeat densities were measured by the number of each repetitive element. Scatterplots for the gene density and repeat distribution are presented in [Supplementary Fig. S4](#). Box sizes correspond to chromosome sizes in *Ormosia* species, whereas the black bars on the upper part of each box (chromosome) represent the hot structural rearrangement region in the chromosomes.

predictions revealed 50,517 and 51,220 genes encoding 54,456 and 55,363 proteins, respectively (Table 3). Table 3 also provides statistics on various gene features in the 3 species' assemblies. Overall, *O. purpureiflora* exhibited the lowest average number of exons and introns per gene as well as the shortest average gene and CDS lengths. Approximately 70.81%, 76.43%, and 72.43% of protein-coding genes in *O. purpureiflora*, *O. emarginata*, and *O. semicastrata*, respectively, were functionally annotated in at least one database (Table 3). Comparatively low annotation rates were also observed in other genomes, such as *Senna tora* (67.16%), *Pisum sativum* (72.70%), and *Sesbania bispinosa* (78.15%; [Supplementary Table S6](#)). By contrast, higher annotation rates were reported in agriculturally important species such as *Glycine max* (99.03%), *Cajanus cajan* (98.97%), *Cicer arietinum* (98.84%), and *Vigna unguiculata* (98.55%), whose genomes have received greater research attention, contributing to more functional information in annotation databases.

The low annotation rates in *Ormosia* species may be attributed to the presence of novel genes with unknown functions, which are less represented in the current annotation databases.

Gene prediction completeness, as assessed using BUSCO, indicated a completeness score of 96.1% in *O. purpureiflora* (88.8% complete and single-copy, 7.3% complete but duplicated), with 1.5% fragmented and 2.4% missing genes. For *O. emarginata*, the BUSCO analysis revealed 95.1% completeness (88.6% complete and single-copy, 6.5% complete but duplicated), with 1.5% fragmented and 3.4% missing genes. In *O. semicastrata*, the completeness score was 96.3% (89.6% complete and single-copy, 6.7% complete but duplicated), with 1.4% fragmented and 2.3% missing genes.

OMark evaluations reported a completeness score of 97.9% for *O. purpureiflora* (67.0% single-copy, 30.0% duplicated), with 2.1% missing genes. Of the predicted genes, 60.4% were con-

**Table 3:** Statistics of predicted genes for three *Ormosia* species<sup>a</sup>

Species	<i>O. purpureiflora</i>	<i>O. emarginata</i>	<i>O. semicastrata</i>
Predicted gene information			
No. of protein-coding genes	55,061	50,517	51,220
No. of mRNAs	59,809	54,456	55,363
No. of exons	254,087	241,766	245,306
No. of CDSs	242,624	232,457	235,681
No. of 5' UTRs	32,588	25,531	27,197
No. of introns	193,432	186,540	189,213
No. of 3' UTRs	30,110	23,826	25,220
Genes			
Average gene length (bp)	2,991.06	3,277.60	3,356.76
Largest length of genes (bp)	423,361	494,687	267,325
Minimum length of genes (bp)	141	153	153
50% cumulative length of genes (bp)	1,467	1,636	1,710
90% cumulative length of genes (bp)	7,312	7,422	7,533
Exons in genes			
Average exons per gene	3.81	4.03	4.03
Average exon length (bp)	226.87	218.64	222.94
Largest length of exons (bp)	8,728	7,959	7,959
Minimum length of exons (bp)	3	3	3
50% cumulative length of exons (bp)	159	149	151
90% cumulative length of exons (bp)	658	581	603
Introns in genes			
Average introns per gene	2.81	3.03	3.03
Average intro length (bp)	757.16	790.41	811.46
Largest length of introns (bp)	422,767	783,472	318,269
Minimum length of introns (bp)	11	11	11
50% cumulative length of introns (bp)	229	234	246
90% cumulative length of introns (bp)	1,375	1,419	1,463
CDS in genes			
Average CDS length (bp)	864.15	881.45	898.37
Largest length of CDSs (bp)	16,359	15,351	16,323
Minimum length of CDSs (bp)	141	150	144
50% cumulative length of CDSs (bp)	609	621	642
90% cumulative length of CDSs (bp)	1941	1953	1917
Gene functional annotations using different databases			
dbCAN	1,671	1,538	1,596
EggNOG	41,143	38,192	38,955
KEGG	20,284	19,163	19,545
GO	29,006	27,167	27,824
InterPro	35,255	32,777	33,548
MEROPS	1,335	1,269	1,292
Pfam	28,007	26,150	27,091
SignalP	4,143	3,778	3,952
UniProt	9,517	8,988	9,234
Total	42,348	39,147	40,100

<sup>a</sup>Using all transcripts.

sistent, 3.2% were inconsistent, and 36.39% are unknown (Supplementary Table S7). For *O. emarginata*, OMArk indicated 97.1% completeness (67.0% single-copy, 30.0% duplicated), with 2.9% missing; 62.8% consistent, 2.5% inconsistent, and 34.66% unknown genes. For *O. semicastrata*, OMArk reported 98.1% completeness (67.8% single-copy, 30.3% duplicated), with 1.9% missing, 62.8% consistent, 2.8% inconsistent, and 34.37% unknown genes. No contamination was detected in the gene sets of any of the 3 *Ormosia* species. Compared with other species, *Ormosia* exhibited a higher proportion of duplicated and unknown genes and lower consistency. Similarly high levels of duplicated genes were observed in *Ormosia*'s sister species, *L. albus* (37.77%), which may be attributed to lineage-specific WGD events (see below the results in "Gene duplications, synteny, and structural variation analysis" section). The low consistency scores are likely linked to the high proportion of unknown genes. The high proportion of un-

known genes in *Ormosia* may result from newly identified genes that lack homologs in OMArk's reference databases, reflecting the limited genomic information available for this lineage. A similar trend of high unknown gene rates (39.07%) and low consistency (56.99%) in *S. tora* may also be explained by the same factor.

The *Ormosia* genes showed high matching rates with Fabaceae representative genes (Supplementary Table S8), ranging from 73.41% to 73.91%. These rates were only slightly lower than those observed for *Medicago truncatula* (75.68%) and *Pisum sativum* (75.36%), supporting the completeness of the predicted *Ormosia* genes.

According to InterPro functional annotation, we found that some photosynthesis-related genes were not annotated in *O. purpureiflora* when compared with *O. emarginata* and *O. semicastrata* (Supplementary Table S9). Specifically, the number of genes associated with Photosystem I PsA/PsB (IPR001280) in *O. pur-*



*purpureiflora* was 3, which was lower than the number of genes in *O. emarginata* (8) and *O. semicastrata* (9). In addition, the InterPro database showed the absence of annotation in several genes related to plant–pathogen interaction (EDS1-like, IPR044214), plant reproduction (DBP10, C-terminal, IPR012541), pyrimidine/nucleotide metabolism (deoxyuridine triphosphate nucleotidohydrolase, IPR008181; dUTPase-like, IPR029054/IPR036157), regeneration (Thioredoxin DCC1, IPR044691), seed maturation protein 1 (SMP1, IPR044984), and nodulin (IPR003387) in *O. purpureiflora*.

### Gene family

A total of 47,608 gene families were identified using OrthoFinder. In *O. purpureiflora*, 50,275 genes (91.3%) were assigned to 27,347 gene families. Among these, 454 families were specific to *O. purpureiflora* (Supplementary Table S10). The genes in these families were mainly enriched in processes such as endoplasmic reticulum to Golgi vesicle-mediated transport and non-membrane-bounded organelle assembly in GO's BP category (Supplementary Table S11) and ribosome biogenesis in eukaryotes in the KEGG analysis (Supplementary Table S12).

The phylogenetic tree (Fig. 3C) indicated that *O. purpureiflora* was sister to *O. emarginata* and that *Ormosia* was sister to *L. albus*. The estimated divergence time between *O. purpureiflora* and *O. emarginata* was approximately 2.94 million years ago (95% CI: 1.19–5.00), whereas the divergence time between *Ormosia* and *Lupinus* was 45.90 million years ago (95% CI: 32.66, 56.44). In *O. purpureiflora*, 1,020 gene families were expanded and 623 were contracted. Among these, the expansion and contraction were significant in 205 and 84 gene families ( $P < 0.05$ ). Significantly expanded gene families were mainly enriched in DNA integration and regulation of amino acid transmembrane transport in GO's biological process (BP) category (Supplementary Table S13) and alkaloid, polyketide, and zeatin biosynthesis in the KEGG analysis (Supplementary Table S14). The significantly contracted gene families were mainly associated with transcription by lipid transport and lipid localization in the GO's BP category (Supplementary Table S15) and with terpenoid biosynthesis in the KEGG analysis (Supplementary Table S16).

The genes in the contracted gene families related to terpenoid biosynthesis were primarily cytochrome P450 (CYP450) genes, which are responsible for downstream activities in the final terpenoid products [108, 109]. However, terpenoids were mostly represented by two conserved domains with Pfam IDs of PF01397 and PF03936 [110]. A comparison showed that the *O. purpureiflora* assembly annotated 23 and 25 of these genes, slightly fewer than those in *O. emarginata* (31 and 26) and *O. semicastrata* (26 and 28). Nevertheless, the number of genes in *Ormosia* species was much higher than that in their sister species, *L. albus* (8 and 10).

### Gene duplications, synteny, and structural variation analysis

WGD analysis indicated that *O. purpureiflora* has undergone 1 WGD event (Fig. 3D), which was shared with the other 2 *Ormosia* species and *L. albus*. Therefore, this WGD event is not specific to *Ormosia* but instead may be specific to the Genistoid lineage in Fabales [16, 111]. Future studies, including newly published Fabaceae genomes, will help confirm this hypothesis.

Gene duplication analysis revealed that the 3 *Ormosia* species exhibited similar numbers of genes across different duplication types (Supplementary Table S17). In *O. purpureiflora*, enrichment analysis showed that WGD-duplicated genes were primarily asso-

ciated with the processes related to calcium ion, blue light, flower, and development, and cytokinin biosynthetic process in the GO's BP category (Supplementary Table S18). In KEGG analysis, these genes were linked to signaling proteins, glycosylphosphatidylinositol (GPI)-anchored proteins, GTP-binding proteins, and SNARE interactions in vesicular transport (Supplementary Table S19). Tandem-duplicated genes were mainly associated with phloem development, glutathione metabolic process, and the biosynthesis of monoterpenoid, anthocyanin, zeatin, and flavonoid (Supplementary Tables S20 and S21). Proximal-duplicated genes were predominantly involved in diterpenoid and triterpenoid biosynthetic process, arginine biosynthetic process, phloem development, and flavone and flavone biosynthesis (Supplementary Tables S22 and S23). These results were consistent with those of a previous study on *O. emarginata* and *O. semicastrata* by Liu et al. [15], which showed that tandem and proximal duplicated genes were relevant to various (secondary) biosynthetic and metabolic processes, including the biosynthesis of alkaloid, flavonoid, and terpenoid.

Synteny analysis within *Ormosia* revealed 48, 42, and 45 syntenic blocks in *O. purpureiflora*, *O. emarginata*, and *O. semicastrata*, respectively (Supplementary Table S24). The longest syntenic blocks identified in these species were between chromosomes 2 and 3. These blocks measured 39,614,256 bp and contained 427 gene pairs in *O. purpureiflora*, 33,895,706 bp with 383 gene pairs in *O. emarginata*, and 36,266,649 bp with 424 gene pairs in *O. semicastrata*. The syntenic relationships were illustrated in a Circos plot (Fig. 3B).

Overall, the *O. purpureiflora* genome exhibited highly syntenic relationships with the other two *Ormosia* genomes, as shown by both synteny analysis (Fig. 3E) and dot plots (Supplementary Fig. S5). However, further genetic variation analysis revealed extensive intrachromosomal rearrangements among the *Ormosia* species (Fig. 3F). These rearrangements were primarily concentrated in specific “hot” chromosomal regions, where the gene density was low, indicating unstable genome architecture in these regions, while gene-rich regions maintained a more conserved genome structure in *Ormosia*.

*O. semicastrata* exhibited greater divergence from *O. purpureiflora* and *O. emarginata*, as evidenced by the high unaligned proportions (52.36% unalignment with *O. purpureiflora* assembly and 50.54% unalignment with *O. emarginata* assembly) compared with lower unaligned proportions between *O. emarginata* and *O. purpureiflora* (24.88% and 28.51%, respectively, Supplementary Table S25). Furthermore, *O. semicastrata* exhibited fewer translocations and duplications than *O. purpureiflora* and *O. emarginata*. These findings align with those of our phylogeny analysis (Fig. 3C) as well as previous results, which have reported that *O. emarginata* and *O. semicastrata* belong to different clades [1]. Although *O. purpureiflora* was sister to *O. emarginata* and thus in the same clade, the structural rearrangements observed between *O. emarginata* and *O. semicastrata* were not preserved in *O. purpureiflora*.

*O. purpureiflora* and *O. emarginata* exhibited the largest inversion on Chromosome 1, spanning from 166,804,741 to 222,962,103 bp in *O. purpureiflora* and from 127,118,909 to 186,491,244 bp in *O. emarginata*. Extensive duplications were also observed on the same chromosome. Notably, a *O. purpureiflora*-specific inverted region was identified on Chromosome 2, spanning from 52,506,652 to 61,757,520 bp (Fig. 3F and Supplementary Fig. S5), which was located away from the rearrangement hot regions. This inverted region was 9,250,868 bp in length and contained 577 genes. The enrichment analysis of these genes revealed their involvement in osmotic stress and temperature regulation

(Supplementary Table S26), which may contribute to *O. purpureiflora*'s adaptation to rocky environments (Fig. 1F), thin soil layers (with low soil moisture content), and relatively high elevation (400–750 m in altitude) [11].

### Identification of nucleotide-binding leucine-rich repeat and other R genes

Compared with other species in our phylogenetic analysis (Fig. 3C), *Ormosia* species were found to have a higher number of R genes (Supplementary Table S27). For the nucleotide-binding leucine-rich repeat (NLR) genes, their numbers were 1,269–1,346 in 3 *Ormosia* species by the Intrepro and Pfam databases searching, and 276–298 by the Resistify program. The results were higher than those of 10–12 compared species. However, when looking at the percentages of the NLR genes in *Ormosia*, they were not high in both databases searching and Resistify identifying results.

Specifically, the number and percentage of NLR genes in the *Ormosia* species were higher than those in the sister species *L. albus*. However, in terms of the other R genes, *L. albus* displayed a higher number and percentage than the *Ormosia* species. Among the *Ormosia* species, *O. purpureiflora* had a higher number and percentage of other R genes than *O. emarginata* and *O. semicastrata*. The distribution of R genes across the chromosomes of each *Ormosia* species is shown in Supplementary Fig. S6. R genes were spread across all 8 chromosomes, following a distribution pattern consistent with the overall gene distribution in *Ormosia* species.

### Transcription factor

Identified with PlantTFDB, *Ormosia* species had a higher number of TF genes than all the other species in our phylogenetic analysis (Fig. 3C), except for *Glycine max*, *Sesbania bispinosa*, *Acacia pycnantha*, and their sister species *L. albus*. However, the percentage of TF genes in *Ormosia* species was relatively low when compared with the species in the phylogeny, particularly in *O. purpureiflora* (3.96%, the lowest one; Supplementary Table S28). Detection with TransFacPred confirmed the high numbers but low percentage of TF genes in *Ormosia*.

### SNP calling

The raw SNPs called by NGSEP identified 37,875,127 loci, comprising 24,941,612 SNPs and 12,933,515 InDels. After quality filtering and InDel removal, 358,992 SNPs were retained. In total, 272,941 loci were identified as deviations from HWE and removed. Further LD filtering retained 40,146 loci. PCAdapt analysis indicated that five main components were suitable to account for population structure, as shown in the Scree plot (Supplementary Fig. S7). PCAdapt identified 5,814 outlier SNPs, whereas BayPass revealed 1,097 candidate SNPs. Across both analyses, 109 SNPs were identified as outliers. By removing these outlier SNPs, 40,037 SNPs remained as neutral loci for population genetics analyses.

### Genetic diversity and genetic structure

Genetic diversity analyses in *O. purpureiflora* (sub)populations indicated that all (sub)populations displayed similar levels of genetic diversity, and LFS4 exhibited the highest genetic diversity for 3 parameters ( $H_o$ ,  $H_e$ , and  $\pi$ ) (Table 1). Four (sub)populations showed close to zero  $F_{is}$ , indicating random mating status in them. Two (sub)populations, LFS1 and LFS4, displayed low negative  $F_{is}$  values, indicating excess heterozygosity. The overall  $F_{st}$  was 0.107, suggesting relatively high population differentiation. Compared with *O. henryi*, a species more widely distributed in southern China, *O. purpureiflora* displayed lower genetic diversity

in the values of  $H_o$  and  $H_e$ , but not in  $\pi$ . For *O. henryi*, the genetic diversity measures were  $H_o$ : 0.228–0.287,  $H_e$ : 0.237–0.290,  $\pi$ : 0.122–0.143, and  $F_{is}$ : –0.023–0.022 in) [112].

Heterozygosity excess in plants may be attributed to several factors, including polyploidy, reproduction mode (such as outcrossing, self-incompatibility systems, and clonal growth), demographic history (such as population bottlenecks), and natural selection (e.g., the overdominant phenomenon, where heterozygous individuals have high survival rates) [113–120]. *O. purpureiflora* is a diploid species, as mentioned earlier. Its flowers are insect-pollinated, exhibiting a typical outcrossing reproductive system. Given the small size of its populations, the limited reproduction among individuals reduces the likelihood of inbreeding and leads to a decrease in inbred offspring. Therefore, the reproductive system may favor heterozygous individuals in the population. *O. purpureiflora* also reproduces asexually through suckering [11], which contributes to the observed heterozygosity excess in some (sub)populations. However, the effects of other factors, such as a historical bottleneck, cannot be overlooked and warrant further investigation.

PCA revealed that the first principal component generally divided LFS4 from the other (sub)populations (Fig. 2B). The second principal component further separated NKS from the others. The third principal component indicated the divergence in some LFS5 individuals. In the ADMIXTURE analysis, the cross-validation (CV) error decreased consistently from  $K = 1$  to  $K = 6$  (Supplementary Fig. S8), but from  $K = 4$  onward, the decrease slowed down. Therefore,  $K = 4$  was identified as the optimal number of genetic groups. Given the limited number of (sub)populations in *O. purpureiflora*, the present study reports the results for  $K = 2$  to  $K = 6$  (Fig. 2C). When  $K = 2$ , individuals in LFS4 were separated from the others. At  $K = 3$ , individuals in NKS were further separated. At  $K = 4$ , LFS5 was separated as a distinct group. When  $K$  was increased further, LFS2 was separated from the others and extensive admixture was observed in all LFS (sub)populations. Both PCA and ADMIXTURE analyses highlighted the distinctiveness of LFS4, although the reasons for this distinctiveness remain unclear.

### Conclusion

Fabaceae play a crucial role in biological nitrogen fixation and serve as a source of nutrition for wild fauna, contributing to the health and balance of ecosystems. The same holds true for *Ormosia* species. Previous studies have shown that *Ormosia* species are rich in secondary metabolites, including alkaloids, terpenes, and flavonoids, which warrant further exploration, particularly from a genomic perspective. The genomes of the two previous *Ormosia* species and the current *O. purpureiflora* genome indicate that genes involved in the biosynthesis of these metabolites are often found in tandem duplications, proximal duplications, or are expanded. The association between gene distribution and repeats suggests that these repeats play a role in gene duplication, highlighting the need for future research on this topic. Thus, the high-quality *Ormosia* genomes serve as a valuable resource for understanding the efficiency of metabolite biosynthesis and identifying potentially useful chromosome regions (such as syntenic regions and structural rearrangements) for future study.

### Additional Files

**Supplementary Table S1:** Protein sequences of the species used for gene prediction.

**Supplementary Table S2:** Species used for comparative genomics.



**Supplementary Table S3:** Species pairs and their estimated divergence times used for time calibration points to infer time-calibrated phylogeny of *O. purpureiflora*.

**Supplementary Table S4:** nQuire results for the ploidy level assessment in *Ormosia* species.

**Supplementary Table S5:** Repeat contents in *Ormosia* assemblies.

**Supplementary Table S6:** Summary of gene functional annotations of the *O. purpureiflora* assembly performed using different databases.

**Supplementary Table S7:** OMARk gene quality assessment.

**Supplementary Table S8:** Number of the genes matching to representative genes in Fabaceae.

**Supplementary Table S9:** Comparison of gene number in part of the InterPro accession among 3 *Ormosia* species.

**Supplementary Table S10:** Summary of gene families.

**Supplementary Table S11:** GO enrichment results for specific gene families in *O. purpureiflora*.

**Supplementary Table S12:** KEGG enrichment results for specific gene families in *O. purpureiflora*.

**Supplementary Table S13:** GO enrichment results for significantly expanded gene families in *O. purpureiflora*.

**Supplementary Table S14:** KEGG enrichment results for significantly expanded gene families in *O. purpureiflora*.

**Supplementary Table S15:** GO enrichment results for significantly contracted gene families in *O. purpureiflora*.

**Supplementary Table S16:** KEGG enrichment results for significantly contracted gene families in *O. purpureiflora*.

**Supplementary Table S17:** Number of different gene duplication in *Ormosia*.

**Supplementary Table S18:** GO enrichment results for *O. purpureiflora* WGD genes.

**Supplementary Table S19:** KEGG enrichment results for *O. purpureiflora* WGD genes.

**Supplementary Table S20:** GO enrichment results for tandem-duplicated genes in *O. purpureiflora*.

**Supplementary Table S21:** KEGG enrichment results for tandem-duplicated genes in *O. purpureiflora*.

**Supplementary Table S22:** GO enrichment results for proximal-duplicated genes in *O. purpureiflora*.

**Supplementary Table S23:** KEGG enrichment results for proximal-duplicated genes in *O. purpureiflora*.

**Supplementary Table S24:** Syntenic block analysis results for *Ormosia* species.

**Supplementary Table S25:** Structural variations between the *Ormosia* species.

**Supplementary Table S26:** GO enrichment results for *O. purpureiflora* specific inversion on Chromosome 2.

**Supplementary Table S27:** R genes in the species.

**Supplementary Table S28:** Summary of transcription factor genes in *Ormosia* species relative to those in the other species.

**Supplementary Figure S1:** Schematic showing the steps of merging repeats measured in RED and EDTA.

**Supplementary Figure S2:** Chromosome numbers observed in *O. purpureiflora* (scale bar, 10  $\mu$ m).

**Supplementary Figure S3:** Genome size estimation using GenomeScope.

**Supplementary Figure S4:** Gene and repeat number density (bin size, 100,000 bp) along chromosomes in *Ormosia* species. The smoothing line is added for each density distribution by the `geom_smooth()` function in the `ggplot2` program of the R package. Box sizes correspond to chromosome sizes in the *Ormosia* species.

**Supplementary Figure S5:** Dot plots among 3 *Ormosia* species. The red arrow indicates a specific inversion on Chromosome 2 of *O. purpureiflora*.

**Supplementary Figure S6:** Distribution of R genes on each chromosome in the 3 *Ormosia* species.

**Supplementary Figure S7:** Scree plot from PCAadapt.

**Supplementary Figure S8:** Cross-validation error plot of admixture analysis.

## Funding

The study is supported by the Guangdong Science and Technology Plan Project (2023A1111110001), the Key-Area Research and Development Program of Guangdong Province (2022B1111230001) and its sub-project (2022B1111230001-2-5), the 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), and the Guangdong Science and Technology Plan Project (grant no. 2023B1212060046).

## Data Availability

Raw sequenced reads have been uploaded to the NCBI Sequence Read Archive under accession numbers SRR24060960 for short WGS reads, SRR24061088 and SRR24061087 for long WGS reads, SRR24085385 for ultra-long WGS reads, SRR24112497 for Hi-C reads, SRR24044811 for fruit RNA-seq reads, SRR24044812 for seed RNA-seq reads, SRR24085891 for leaf RNA-seq reads, SRR24085890 for flower RNA-seq reads in *O. purpureiflora*; SRR25460826 for Hi-C reads of *O. emarginata*; SRR25460825 for Hi-C reads for *O. semicastrata*; SRR29820911–SRR29820936 for resequencing reads of LFS1, SRR29824870–SRR29824895 for resequencing reads of LFS2, SRR29837260–SRR29837285 for resequencing reads of LFS3, SRR29856316–SRR29856341 for resequencing reads of LFS4, SRR29887191–SRR29887216 for resequencing reads of LFS5, SRR29761002–SRR29761004, SRR29761010–SRR29761017, SRR29761028–SRR29761030, SRR29761107, SRR29761108, SRR29761114, SRR29761115, SRR29761118, SRR29761123, SRR29761124, SRR29761126, SRR29761139 for resequencing reads of NKS in *O. purpureiflora*. Assembled genomes are available under accession numbers GCA\_04095595.1 for *O. purpureiflora*, GCA\_029884595.2 for *O. semicastrata*, and GCA\_029884605.2 for *O. emarginata*. Annotations, SNPs, and the other files have been submitted to figshare [122]. All additional supporting data are available in the GigaScience repository, GigaDB [123], with separate datasets for *O. purpureiflora* [124], *O. semicastrata* [125], and *O. emarginata* [126].

## Competing Interests

The authors declare that they have no competing interests.

## References

1. Torke BM, Cardoso D, Chang H, 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.* 2022;166:107329. <https://doi.org/10.1016/j.ympev.2021.107329>.

2. Niu M, Jiang K-W, Song Z-Q, et al. Two new synonyms of *Ormosia semicastrata* (Fabaceae, Papilionoideae, Ormosieae). *Phytotaxa*. 2023;613(2):140–52. <https://doi.org/10.11646/phytotaxa.613.2.3>.
3. Wang Z, Shi G, Sun B, et al. A new species of *Ormosia* (Leguminosae) from the Middle Miocene of Fujian, Southeast China and its biogeography. *Rev Palaeobot Palynol*. 2019;270:40–47. <https://doi.org/10.1016/j.revpalbo.2019.07.003>.
4. Li L, Lei M, Wang H, et al. First report of dieback caused by *Lasiodiplodia pseudotheobromae* on *Ormosia pinnata* in China. *Plant Dis*. 2020;104:2551–55. <https://doi.org/10.1094/PDIS-03-20-0647-RE>.
5. Wei L, Wang G, Xie C. Predicting suitable habitat for the endangered tree *Ormosia microphylla* in China. *Sci Rep*. 2024;14:10330. <https://doi.org/10.1038/s41598-024-61200-5>.
6. Zhang L-J, Zhou W-J, Ni L, et al. A review on chemical constituents and pharmacological activities of *Ormosia*. *Chin Tradit Herbal Drugs*. 2021;52(14):4433–42. <https://doi.org/10.7501/j.issn.0253-2670.2021.14.035>.
7. Zhou Q-Q, Xie X-Y, Zhu J-W, et al. Hosimosines A-E, structurally diverse cytosine derivatives from the seeds of *Ormosia hosiei* Hemsl. et Wils. *Fitoterapia*. 2023;170:105661. <https://doi.org/10.1016/j.fitote.2023.105661>.
8. Zhou W, Quan Y, Chen Y, et al. A new lignan from leaves of *Ormosia xylocarpa*. *Rec Nat Prod*. 2023;17(1):189–94. <https://doi.org/10.25135/rnp.338.2203.2386>.
9. Wang J, Li L, Wang Z, 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*. 2023;246:125601. <https://doi.org/10.1016/j.ijbiomac.2023.125601>.
10. Wang J, Wang X, Deng X, et al. Analysis of candidate genes for terpene synthesis in *Ormosia henryi* based on metabolome and transcriptome. *J Zhejiang A&F Univ*. 2023;40(5):970–81. <https://doi.org/10.11833/j.issn.2095-0756.2022073>.
11. Yu E-P. Preliminary study on conservation ecology of the rare and endemic plant *Ormosia purpureiflora* to Guangdong. Master's thesis. Beijing: University of Chinese Academy of Sciences; 2024.
12. Tang J, Zou R, Wei X, et al. Complete chloroplast genome sequences of five *Ormosia* species: molecular structure, comparative analysis, and phylogenetic analysis. *Horticulturae*. 2023;9(7):796. <https://doi.org/10.3390/horticulturae9070796>.
13. Wang Z-F, Yu E-P, Zeng QS, et al. The complete chloroplast genome of *Ormosia purpureiflora* (Fabaceae). *Mitochondrial DNA B*. 2012;6(12):3327–28. <https://doi.org/10.1080/23802359.2021.1994901>.
14. Wang Z-F, Zhang Y, Zhong X-J, et al. The complete mitochondrial genome of *Ormosia boluensis*. *Mitochondrial DNA B*. 2021;6(8):2109–11. <https://doi.org/10.1080/23802359.2021.1920503>.
15. Liu P-P, Yu E-P, Tan Z-J, et al. Genome assemblies of two *Ormosia* species: gene duplication related to their evolutionary adaptation. *Agronomy*. 2023;13:1757. <https://doi.org/10.3390/agronomy13071757>.
16. Zhao Y, Zhang R, Jiang K-W, et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Mol Plant*. 2021;14(5):748–73. <https://doi.org/10.1016/j.molp.2021.02.006>.
17. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. *PLoS One*. 2021;16(10):e0257521. <https://doi.org/10.1371/journal.pone.0257521>.
18. Wang Z-F, Rouard M, Droc G, et al. Genome assembly of *Musa beccarii* shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes. *GigaScience*. 2023;12:giad005. <https://doi.org/10.1093/gigascience/giad005>.
19. Wang Z-F, Fu L, Yu E-P, et al. Chromosome-level genome assembly and demographic history of *Euryodendron excelsum* in monotypic genus endemic to China. *DNA Res*. 2024;31(1):dsad028. <https://doi.org/10.1093/dnares/dsad028>.
20. Oxford Nanopore's Basecaller. <https://github.com/nanoporetech/h/dorado>. Accessed 15 August 2023.
21. Joshi NA, Fass JN. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). 2011; <https://github.com/najoshi/sickle>. Accessed 3 September 2021.
22. Długosz M, Deorowicz S. RECKONER: read error corrector based on KMC. *Bioinformatics*. 2017;33:1086–89. <https://doi.org/10.1093/bioinformatics/btw746>.
23. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27:764–70. <https://doi.org/10.1093/bioinformatics/btr011>.
24. Vurture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33:2202–04. <https://doi.org/10.1093/bioinformatics/btx153>.
25. Weiß CL, Pais M, Cano LM, et al. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinform*. 2018;19:122. <https://doi.org/10.1186/s12859-018-2128-z>.
26. Porechop v. 0.2.4. <https://github.com/rrwick/Porechop/release/tag/v0.2.4>. Accessed 8 January 2019.
27. Hu J, Wang Z, Sun Z, et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol*. 2024;25:107. <https://doi.org/10.1186/s13059-024-03252-4>.
28. Pseudohaploid. <https://github.com/schatzlab/pseudohaploid>. Accessed 28 August 2020.
29. Guan DF, McCarthy SA, Wood J, et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36:2896–98. <https://doi.org/10.1093/bioinformatics/btaa025>.
30. Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27(5):737–46. <https://doi.org/10.1101/gr.214270.116>.
31. Aury JM, Istace B, Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genom Bioinform*. 2021;3(2):lqab034. <https://doi.org/10.1093/nargab/lqab034>.
32. Wick RR, Holt KE. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol*. 2022;18(1):e1009802. <https://doi.org/10.1371/journal.pcbi.1009802>.
33. Depthcharge v. 0.2.0. <https://github.com/slimsuite/depthcharge>. Accessed 28 January 2023.
34. Scaffhic v. 1.1. <https://github.com/wtsi-hpag/scaffHiC>. Accessed 7 December 2022.
35. Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. 2016;3(1):95–98. <https://doi.org/10.1016/j.cels.2016.07.002>.
36. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. 2017;356(6333):92–95. <https://doi.org/10.1126/science.aal3327>.



37. Xu M, Guo L, Gu S, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience*. 2020;9(9):giaa094. <https://doi.org/10.1093/gigascience/giaa094>.
38. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*. 2016;44(12):e11. <https://doi.org/10.1093/nar/gkw294>.
39. Teloclip v. 0.0.3. <https://github.com/Adamtaranto/teloclip>. Accessed 28 March 2023.
40. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol*. 2019;1962:227–45. <https://doi.org/10.1093/bioinformatics/btv351>.
41. Rashid U, Wu C, Shiller J, et al. AssemblyQC: a Nextflow pipeline for reproducible reporting of assembly quality. *Bioinformatics*. 2024;40(8):btac477. <https://doi.org/10.1093/bioinformatics/btae477>.
42. Zhang Y, Lu H-W, Ruan J. GAEP: a comprehensive genome assembly evaluating pipeline. *J Genet Genomics*. 2023;50(10):747–54. <https://doi.org/10.1016/j.jgg.2023.05.009>.
43. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res*. 2018;46(21):e126. <https://doi.org/10.1093/nar/gky730>.
44. Rhie A, Walenz BP, Koren S, et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245. <https://doi.org/10.1186/s13059-020-02134-9>.
45. Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 2019;20:275. <https://doi.org/10.1186/s13059-019-1905-y>.
46. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinform*. 2015;16(1):227. <https://doi.org/10.1186/s12859-015-0654-5>.
47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–42. <https://doi.org/10.1093/bioinformatics/btq033>.
48. Gu Z, Gu L, Eils R, et al. circlize implements and enhances circular visualization in R. *Bioinformatics*. 2014;30(19):2811–12. <https://doi.org/10.1093/bioinformatics/btu393>.
49. Brůna T, Hoff KJ, Lomsadze A, et al. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*. 2021;3(1):lqaa108. <https://doi.org/10.1093/nargab/lqaa108>.
50. Funannotate v. 1.8.16. <https://github.com/nextgenusfs/funannotate>. Accessed 12 March 2023.
51. Zhang H, Tanner Y, Huang L, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46:W95–101. <https://doi.org/10.1093/nar/gky418>.
52. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. 2017;34:2115–22. <https://doi.org/10.1093/molbev/msx148>.
53. The Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47(D1):D330–38. <https://doi.org/10.1093/nar/gky1055>.
54. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–29. <https://doi.org/10.1038/75556>.
55. Kanehisa M, Soto Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44(D1):D457–62. <https://doi.org/10.1093/nar/gkv1070>.
56. Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. 2019;47(D1):D351–60. <https://doi.org/10.1093/nar/gky1100>.
57. Rawlings ND, Barrett AJ, Thomas PD, et al. The merops database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2018;46(D1):D624–32. <https://doi.org/10.1093/nar/gkx1134>.
58. El-Gebali S, Mistry J, Bateman A., et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47(D1):D427–32. <https://doi.org/10.1093/nar/gky995>.
59. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37(4):420–23. <https://doi.org/10.1038/s41587-019-0036-z>.
60. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506–15. <https://doi.org/10.1093/nar/gky1049>.
61. Nevers Y, Vesztrócy AW, Rossier V, et al. Quality assessment of gene repertoire annotations with OMARK. *Nat Biotechnol*. 2025;43:124–133. <https://doi.org/10.1038/s41587-024-02147-w>.
62. Fernandez CGT, Bayer PE, Peterleit J, et al. The conservation of gene models can support genome annotation. *Plant Genome*. 2023;16(3):e20377. <https://doi.org/10.1002/tpg2.20377>.
63. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinf*. 2009;10:421. <https://doi.org/10.1186/1471-2105-10-421>.
64. Emms DM, OrthoFinder KS: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157. <https://doi.org/10.1186/s13059-015-0721-2>.
65. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20:238. <https://doi.org/10.1186/s13059-019-1832-y>.
66. Emms DM, Kelly S. STAG: species tree inference from all genes. *bioRxiv* 2018; <https://doi.org/10.1101/267914>.
67. Emms DM, Kelly S. STRIDE: species tree root inference from gene duplication events. *Mol Biol Evol*. 2017;34:3267–78. <https://doi.org/10.1093/molbev/msx259>.
68. Han MV, Thomas GWC, Jose LM, et al. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using cafe 3. *Mol Biol Evol*. 2013;30(8):1987–97. <https://doi.org/10.1093/molbev/mst100>.
69. dos Reis M, Zhu T, Yang Z. The impact of the rate prior on bayesian estimation of divergence times with multiple loci. *System Biol*. 2014;63:555–65. <https://doi.org/10.1093/sysbio/syu020>.
70. Chen CJ, Chen H, Zhang Y, et al. TBtools—an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*. 2020;13(8):1194–1202. <https://doi.org/10.1016/j.molp.2020.06.009>.
71. Zwaenepoel A, de Peer YV. wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics*. 2019;35:2153–55. <https://doi.org/10.1093/bioinformatics/bty915>.
72. Almeida-Silva F, Van de Peer Y. doubletrouble: identification and classification of duplicated genes. R package version 0.99.1. 2022. <https://github.com/almeidasilvaf/doubletrouble>. Accessed 28 March 2023.
73. Qiao X, Li QH, Yin H, et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants.

- Genome Biol. 2019;20:38. <https://doi.org/10.1186/s13059-019-1650-2>.
74. MCScanX. <https://github.com/wyp1125/MCScanX>. Accessed 28 July 2023.
  75. Wang Y, Jia L, Tian G, et al. shinyCircos-V2.0: leveraging the creation of Circos plot with enhanced usability and advanced features. *iMeta*. 2023;2(2):e109. <https://doi.org/10.1002/imt2.109>.
  76. Bandi V, Gutwin C. Interactive exploration of genomic conservation. In: Proceedings of the 46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20). Waterloo, Canada: Canadian Human-Computer Communications Society, 2020. <https://doi.org/10.20380/GI2020.09>.
  77. Pérez-Wohlfeil E, Diaz-del-Pino S, Trelles O. Ultra-fast genome comparison for large-scale genomic experiments. *Sci Rep*. 2019;9:10274. <https://doi.org/10.1038/s41598-019-46773-w>.
  78. Goel M, Schneeberger K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics*. 2022;38(10):2922–26. <https://doi.org/10.1093/bioinformatics/btac196>.
  79. Shao Z-Q, Xue J-Y, Wu P, et al. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol*. 2016;170(4):2095–2109. <https://doi.org/10.1104/pp.15.01487>.
  80. Chou WC, Jha S, Linhoff MW, et al. The NLR gene family: from discovery to present day. *Nat Rev Immunol*. 2023;23:635–54. <https://doi.org/10.1038/s41577-023-00849-x>.
  81. MdL S, Resende MLV, Alves GSC, et al. Genome-wide identification, characterization, and comparative analysis of NLR resistance genes in *Coffea* spp. *Front Plant Sci*. 2022;13:868581. <https://doi.org/10.3389/fpls.2022.868581>.
  82. Liu Y, Zhang Y-M, Tang Y, et al. The evolution of plant NLR immune receptors and downstream signal components. *Curr Opin Plant Biol*. 2023;73:102363. <https://doi.org/10.1016/j.pbi.2023.102363>.
  83. De-la-Cruz IM, Hallab A, Olivares-Pinto U, et al. Genomic signatures of the evolution of defence against its natural enemies in the poisonous and medicinal plant *Datura stramonium* (Solanaceae). *Sci Rep*. 2021;11:882. <https://doi.org/10.1038/s41598-020-79194-1>.
  84. Smith M, Jones JT, Hein I. Resistify: A novel NLR classifier that reveals Helitron-associated NLR expansion in solanaceae. *Bioinform Biol Insights*. 2025;19:11779322241308944. <https://doi.org/10.1177/11779322241308944>.
  85. Tian F, Yang DC, Meng YQ, et al. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res*. 2019;48(D1):D1104–13. <https://doi.org/10.1093/nar/gkz1020>.
  86. Patiyal S, Tiwari P, Ghai M, et al. A hybrid approach for predicting transcription factors. *Front Bioinform*. 2024;4:1425419. <https://doi.org/10.3389/fbinf.2024.1425419>.
  87. Tello D, Gil J, Loaiza CD, et al. NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics*. 2019;35(22):4716–23. <https://doi.org/10.1093/bioinformatics/btz275>.
  88. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
  89. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156–58. <https://doi.org/10.1093/bioinformatics/btr330>.
  90. SNP Filtering Tutorial. <http://www.ddocent.com/filtering/>. Accessed 10 January 2021.
  91. Purcell S, Chang C. PLINK. [www.cog-genomics.org/plink/1.9/](http://www.cog-genomics.org/plink/1.9/). Accessed 10 January 2021.
  92. Chang CC, Chow CC, Tellier LCAM, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1):7. <https://doi.org/10.1186/s13742-015-0047-8>. Date deposit on 23 February 2015.
  93. Gaunt T, Rodríguez S, Day I. Cubic exact solutions for the estimation of pairwise haplotype frequencies: implications for linkage disequilibrium analyses and a web tool 'CubeX'. *BMC Bioinform*. 2007;8:428. <https://doi.org/10.1186/1471-2105-8-428>.
  94. Luu K, Bazin E, Blum MG. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Resour*. 2017;17:67–77. <https://doi.org/10.1111/1755-0998.12592>.
  95. Privé F, Luu K, Vilhjálmsson BJ, et al. Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. *Mol Biol Evol*. 2020;37:2153–54. <https://doi.org/10.1093/molbev/msaa053>.
  96. Gautier M. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*. 2015;201:1555–79. <https://doi.org/10.1534/genetics.115.181453>.
  97. Korunes KL, Samuk K. pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour*. 2021;21:1359–68. <https://doi.org/10.1111/1755-0998.13326>.
  98. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19:1655–64. <https://doi.org/10.1101/gr.094052.109>.
  99. Zheng X, Levine D, Shen J, et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012;28(24):3326–28. <https://doi.org/10.1093/bioinformatics/bts606>. Accessed 4 June 2024.
  100. Musmann S, Douglas MR, Chafin T, et al. AdmixPipe: population analyses in Admixture for non-model organisms. *BMC Bioinform*. 2020;21:337. <https://doi.org/10.1186/s12859-020-03701-4>.
  101. Kopelman NM, Mayzel J, Jakobsson M, et al. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*. 2015;15:1179–91. <https://doi.org/10.1111/1755-0998.12387>.
  102. World Flora Online (WFO). *Ormosia Macrocalyx* Ducke. 2024. UK: London. <http://www.worldfloraonline.org/taxon/wfo-0000168275>. Accessed 10 October 2024.
  103. Bandel G. Chromosome numbers and evolution in the Leguminosae. *Caryologia*. 1974;27(1):17–32. <https://doi.org/10.1080/00087114.1974.10796558>.
  104. Chen Y-L, Wang Z-F, Jian S-G, et al. Genome assembly of *Cordia subcordata*, a coastal protection species in Tropical Coral Islands. *Int J Mol Sci*. 2023;24:16273. <https://doi.org/10.3390/ijms242216273>.
  105. Mochizuki T, Sakamoto M, Tanizawa Y, et al. A practical assembly guideline for genomes with various levels of heterozygosity. *Brief Bioinform*. 2023;24(6):bbad337. <https://doi.org/10.1093/bib/bbad337>.
  106. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to base-pair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*. 2018;19:90. <https://doi.org/10.1186/s13059-018-1462-9>.
  107. Chen Y, Nie F, Xie SQ, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun*. 2021;12:60. <https://doi.org/10.1038/s41467-020-20236-7>.

108. Xiao H, Zhang Y, Wang M. Discovery and engineering of cytochrome P450s for terpenoid biosynthesis. *Trends Biotechnol.* 2019;37:618–31. <https://doi.org/10.1016/j.tibtech.2018.11.008>.
109. Zheng X, Li P, Lu X. Research advances in cytochrome P450-catalysed pharmaceutical terpenoid biosynthesis in plants. *J Exp Bot.* 2019;70:4619–30. <https://doi.org/10.1093/jxb/erz203>.
110. Liu H-L, Harris AJ, Wang Z-F, et al. The genome of the paleogene relic tree *Bretschneidera sinensis*: insights into trade-offs in gene family evolution, demographic history, and adaptive SNPs. *DNA Res.* 2022;29(1):dsac003. <https://doi.org/10.1093/dnares/dsac003>.
111. Xu W, Zhang Q, Yuan W, et al. The genome evolution and low-phosphorus adaptation in white lupin. *Nat Commun.* 2020;11:1069. <https://doi.org/10.1038/s41467-020-14891-z>.
112. Zhou C, Xia S, Wen Q, et al. Genetic structure of an endangered species *Ormosia henryi* in southern China, and implications for conservation. *BMC Plant Biol.* 2023;23:220. <https://doi.org/10.1186/s12870-023-04231-w>.
113. Balloux F. Heterozygote excess in small populations and the heterozygote-excess effective population size. *Evolution.* 2004;58(9):1891–900. <https://doi.org/10.1111/j.0014-3820.2004.tb00477.x>.
114. Stoeckel S, Grange J, Fernández-Manjarres JF, et al. Heterozygote excess in a self-incompatible and partially clonal forest tree species—*Prunus avium* L. *Mol Ecol.* 2006;15(8):2109–18. <https://doi.org/10.1111/j.1365-294X.2006.02926.x>.
115. Stevens L, Salomon B, Sun G. Microsatellite variability and heterozygote excess in *Elymus trachycaulus* populations from British Columbia in Canada. *Biochem Syst Ecol.* 2007;35(11):725–36. <https://doi.org/10.1016/j.bse.2007.05.017>.
116. Campoy JA, Lerigoleur-Balsemin E, Christmann H, et al. Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of *Prunus avium* L. landraces and bred cultivars. *BMC Plant Biol.* 2016;16:49. <https://doi.org/10.1186/s12870-016-0712-9>.
117. Ruiz Mondragon KY, Aguirre-Planter E, Gasca-Pineda J, et al. Conservation genomics of *Agave tequilana* Weber var. Azul: low genetic differentiation and heterozygote excess in the tequila agave from Jalisco, Mexico. *PeerJ.* 2022;10:e14398. <https://doi.org/10.7717/peerj.14398>.
118. Cisternas-Fuentes A, Koski MH. Drivers of strong isolation and small effective population size at a leading range edge of a widespread plant. *Heredity.* 2023;130:347–57. <https://doi.org/10.1038/s41437-023-00610-z>.
119. Depecker J, Verleysen L, Asimonyio JA, et al. Genetic diversity and structure in wild robusta coffee (*Coffea canephora* A. Froehner) populations in Yangambi (DR Congo) and their relation to forest disturbance. *Heredity.* 2023;130:145–53. <https://doi.org/10.1038/s41437-022-00588-0>.
120. Le Veve A, Burghgraeve N, Genete M, et al. Long-term balancing selection and the genetic load linked to the self-incompatibility locus in *Arabidopsis halleri* and *A. lyrata*. *Mol Biol Evol.* 2023;40(6):msad120. <https://doi.org/10.1093/molbev/msad120>.
121. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
122. Wang Z-F, Yu E-P, Fu L, et al. Chromosome-scale assemblies of three *Ormosia* species. *FigShare Database.* 2025. <https://doi.org/10.6084/m9.figshare.26826466.v15>
123. Wang Z-F, Yu E-P, Fu L, et al. Supporting data for “Chromosome-scale assemblies of three *Ormosia* species: repetitive sequences distribution and structural rearrangement”. *GigaScience Database.* 2025; <https://doi.org/10.5524/102671>. Date of deposit on 24 March 2025.
124. Wang Z-F, Yu E-P, Fu L, et al. The genomic data of *Ormosia purpureiflora*. *GigaScience Database.* 2025; <https://doi.org/10.5524/102672>. Date of deposit on 24 March 2025.
125. Wang Z-F, Yu E-P, Fu L, et al. The genomic data of *Ormosia semicastrata*. *GigaScience Database.* 2025; <https://doi.org/10.5524/102673>. Date of deposit on 24 March 2025.
126. Wang Z-F, Yu E-P, Fu L, et al. The genomic data of *Ormosia emarginata*. *GigaScience Database.* 2025; <https://doi.org/10.5524/102674>. Date of deposit on 24 March 2025.