Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Chromosome-level genome assembly of Camellia yunnanensis and mining of key candidate genes for white flower color and floral scent based on the phenylpropanoid pathway

Yujia Zeng
Southwest Forestry University
Hongxing Xiao
Southwest Forestry University
Fujun Yan
Southwest Forestry University
Xueqin Wu
Southwest Forestry University
Kaihua Jia
Shandong Academy of Agricultural Sciences
Wei Zheng
Southwest Forestry University
Huizhen Hu
Southwest Forestry University
Zhonglang Wang
Chinese Academy of Sciences
Fang Geng
fanggeng20@swfu.edu.cn

Southwest Forestry University

Research Article

Keywords: Camellia yunnanensis, chromosome-level genome, Hi-C assembly, white flower and floral scent traits, phenylpropanoid metabolic pathway

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11386089/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Chromosome-level genome assembly of *Camellia yunnanensis* and mining of key candidate genes for white flower color and floral scent based on the phenylpropanoid pathway

Yujia Zeng[1, 2+], Hongxing Xiao[1, 2+], Fujun Yan[1, 2], Xueqin Wu[1, 2], Kaihua Jia[3], Wei Zheng[1, 2], Huizhen Hu[1, 2], Zhonglang Wang[4], Fang Geng[1, 2*]

[1] College of Landscape Architecture and Horticulture, Southwest Forestry University, Kunming 650224, China;
[2] Yunnan Key Laboratory of Landscape Plant Resource Cultivation and Application, Kunming 650224, China;
[3] Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan, 250100, China;
[4] Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China.

* Correspondence: Fang Geng, fanggeng20@swfu.edu.cn.
+ Yujia Zeng and Hongxing Xiao contributed equally to this work.

Abstract

Background *Camellia yunnanensis,* a rare endemic woody ornamental plant in Yunnan, possesses unique white fragrant flowers and strong stress tolerance, serving as a vital germplasm resource for the evolutionary research and molecular breeding of *Camellia*. Despite being listed as a least-concern species by IUCN, its wild populations are sparse and protected in Yunnan, and the genomic basis and molecular regulatory mechanisms of its floral traits remain poorly understood.

Results We performed a chromosome-level de novo genome assembly of *C. yunnanensis* using integrated PacBio HiFi, Oxford Nanopore, and Hi-C sequencing technologies. The assembled genome size was 5.61 Gb, consisting of two haplotypes (2.89 Gb and 2.72 Gb) with contig N50 values of 79.14 Mb and 114.74 Mb, respectively. A total of 30 pseudochromosomes were anchored with a mounting rate of 99.52 % and a scaffold N50 of 190.51 Mb, and canonical plant telomeric repeats were identified at all chromosome terminals. Repetitive sequences accounted for 67.03 % of the genome, dominated by LTR retrotransposons (55.38 %), and 96,898 protein-coding genes were predicted. Transcriptomic analysis based on the bud, initial flowering,

and full-bloom stages revealed that 37 structural genes involved in the phenylpropanoid pathway (map00940) were significantly differentially expressed among the three stages. At the full-bloom stage, the significant downregulation of *DFR*/*ANS* together with the upregulation of *FLS* inhibited anthocyanin biosynthesis, redirecting metabolic precursors preferentially toward the flavonol branch and resulting in the white flower phenotype; the sustained high expression of *C3H*/*CCoAOMT* from the bud stage to the initial flowering stage ensured the synthesis of benzenoid aroma precursors. Compared with other *Camellia* species, *C. yunnanensis* redistributes metabolic precursors through transcriptional silencing of terminal anthocyanin genes while maintaining an active floral scent branch, a regulatory pattern consistent with its unique white-flowered and fragrant phenotype.

Conclusions In summary, coordinated regulation of core genes in the phenylpropanoid pathway mediates the distinctive white flower color and floral scent traits of *C. yunnanensis*. Furthermore, the high-quality genome obtained in this study provides important resources for mining candidate genes related to the species' distinctive ornamental traits, and lays a foundation for *Camellia* evolution, germplasm conservation, and breeding of aromatic ornamental plants.

Keywords: *Camellia yunnanensis*; chromosome-level genome; Hi-C assembly; white flower and floral scent traits; phenylpropanoid metabolic pathway

Background

The genus *Camellia* (Theaceae) is a woody plant group with great evolutionary significance and considerable horticultural exploitation value. This genus exhibits rich phenotypic variation and harbors abundant wild germplasm resources characterized by distinctive flower color, unique floral fragrance, and superior stress tolerance, serving as pivotal

materials for elucidating the evolutionary patterns of ornamental traits and facilitating the innovative breeding of camellia cultivars[1, 2]. In recent years, omics studies on *Camellia* plants have achieved substantial progress. Economically vital cultivated species, including *C. sinensis* and *C. oleifera*, have been well-resolved with high-quality chromosome-level reference genomes and systematic population genetic analyses, laying a solid foundation for in-depth functional gene exploration and genetic improvement[3-6]. In contrast, numerous wild ornamental *Camellia* species endemic to Southwest China possess excellent breeding potential, yet their genomic resources remain extremely scarce. The molecular regulatory mechanisms governing key ornamental traits (e.g., flower color and floral fragrance) in these wild species are poorly characterized, which severely limits the scientific evaluation, in situ conservation, and targeted utilization of elite wild *Camellia* germplasms.

Camellia yunnanensis, commonly known as Monkey wood, is an endemic evergreen shrub or small tree in the mountainous regions of Southwest China. It is predominantly distributed in Yunnan Province, with sporadic populations occurring in southern Sichuan and Guizhou, and represents a typical protected native germplasm of Yunnan[7]. Distinct from most cultivated camellias that are fragrance-free and lack white-flower germplasm resources, *C. yunnanensis* produces fragrant white flowers, rendering it a precious wild parental material for breeding fragrant white camellia varieties. In terms of conservation status, although *C. yunnanensis* is classified as Least Concern (LC) in the IUCN Red List[8], persistent habitat fragmentation and anthropogenic disturbances have caused the shrinkage of wild populations and impaired natural regeneration. Accordingly, it is listed as a Class III Key Protected Wild Plant of Yunnan Province. Long-term adaptation to high-altitude and barren habitats has endowed *C.*

yunnanensis with prominent stress tolerance, making it an essential genetic resource for stress resistance improvement in *Camellia*. Despite its high germplasm and ecological value, no chromosome-level high-quality genome of *C. yunnanensis* has been reported to date. Its genomic background and the molecular regulatory networks underlying key floral traits remain unclarified, hindering comprehensive insights into its phylogenetic position and adaptive evolutionary mechanisms.

High-quality chromosome-level genomes are indispensable for elucidating plant evolutionary processes, mining functional genes, conducting conservation genomic research, and promoting molecular breeding of woody species[9, 10]. Currently, the genomic resources of *Camellia* show an uneven distribution pattern. Most available chromosome-level genome assemblies focus on economically important species of the sections Thea and Oleifera, whereas the majority of endemic wild *Camellia* species with unique ornamental characteristics lack high-resolution reference genomes. This research gap impedes the systematic exploration of their phylogenetic relationships, genomic variation features, and genetic basis of specific phenotypic traits, which has become a key bottleneck restricting the in-depth study of *Camellia* biodiversity. Therefore, performing chromosome-level genome assembly and multi-omics analysis of rare wild *Camellia* species to dissect trait regulatory mechanisms is critical for enriching the genomic resource repository of *Camellia* and advancing the conservation and innovative utilization of wild germplasms.

Flower color and floral fragrance are core ornamental traits that determine the horticultural value of camellias, and their underlying molecular regulatory mechanisms have long been research hotspots in ornamental plant biology[11]. Accumulated studies have verified that the

biosynthesis of anthocyanins and phenylpropanoid-derived volatile aromatic compounds shares the upstream phenylpropanoid metabolic pathway, and differential transcriptional regulation of downstream branch pathways dominates the phenotypic differentiation of flower color and floral fragrance in plants[12, 13]. *C. yunnanensis* exhibits a rare combination of white petals and distinct floral fragrance, making it an ideal natural model for exploring the synergistic regulatory mechanism of multiple branches in the phenylpropanoid pathway. To fill the genomic research gap of *C. yunnanensis* and identify key genes responsible for its white-flower and floral scent traits, this study integrated PacBio HiFi, Oxford Nanopore, and Hi-C sequencing technologies to generate the first high-quality chromosome-level genome assembly and gene annotation of *C. yunnanensis*. Combined with transcriptomic data from three flower developmental stages (flower bud, initial flowering, and full blooming), we elucidated the spatiotemporal expression patterns of the phenylpropanoid metabolic pathway and screened core candidate genes underlying the formation of white flower color and floral scent. The genomic and transcriptomic resources obtained in this study enrich the omic database for wild *Camellia* germplasms, providing solid theoretical basis and fundamental data support for future research on *Camellia* evolutionary, conservation of rare wild germplasms, and molecular breeding of fragrant white -flowered *Camellia* cultivars.

Methods

Plant materials

Fresh leaf samples were collected from healthy adult individuals of wild *C. yunnanensis* in Zixi Mountain, Chuxiong Yi Autonomous Prefecture, Yunnan (101°22′ E, 25°00′ N) for

whole-genome sequencing (Fig. 1). Flower, root, and young stem tissues were also harvested for full-length transcriptome sequencing. To identify key genes regulating white flower color and floral scent traits, flower samples were collected at three critical developmental stages: the bud stage (S1, unopened flower buds), the initial flowering stage (S2, petals beginning to unfold with undehisced anthers), and the full-bloom stage (S3, flowers fully open). All samples were immediately wrapped in aluminum foil, snap-frozen in liquid nitrogen, and stored at −80 °C for subsequent nucleic acid extraction and sequencing analyses. The collection and use of these materials fully complied with institutional, national, and local laws and regulations, as well as the provisions of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). The plant species was formally identified by Prof. Zhonglang Wang from the Kunming Institute of Botany, Chinese Academy of Sciences, and no voucher specimen was deposited as the fresh wild materials were collected directly for sequencing.

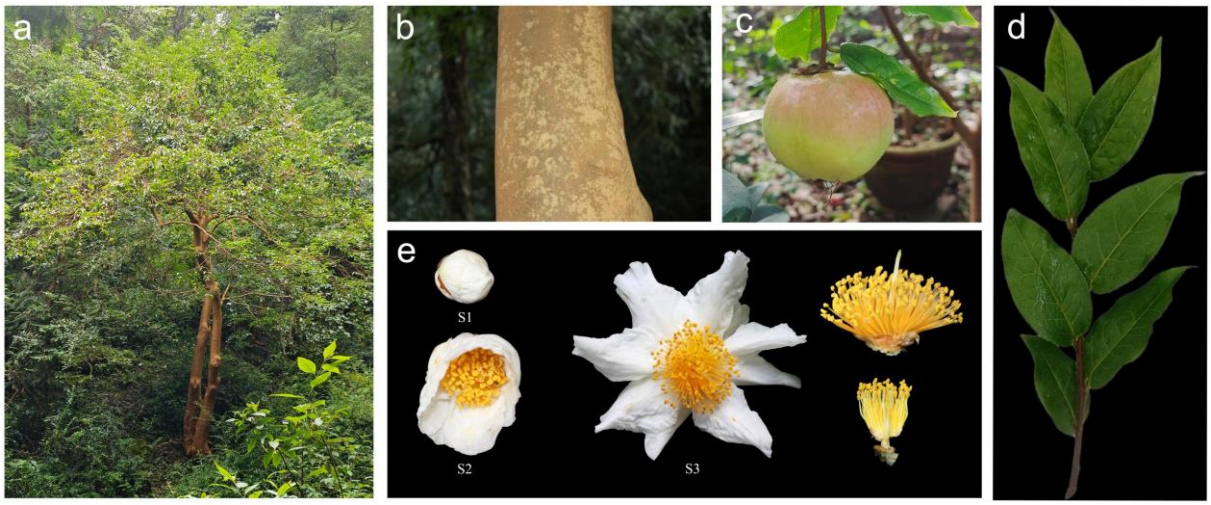


Fig. 1 (a) The morphology of wild individual of *C. yunnanensis.* (b) Trunk. (c) Fruit. (d) Leaf and stem. (e) Three flowering stages (S1: bud stage, S2: initial flowering stage, S3: full-bloom stage); stamens and pistils.

DNA/RNA extraction

High-quality genomic DNA was isolated from fresh leaf tissues using the modified CTAB method. The purity, concentration, and integrity of genomic DNA were assessed via 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (Thermo Fisher Scientific, USA), and Qubit 3.0 fluorometer (Life Technologies, USA). Total RNA of *C. yunnanensis* was extracted using TRIzol® Reagent (Invitrogen, USA). The concentration and quality of total RNA were evaluated using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) to ensure compliance with library construction standards.

Genome and transcriptome sequencing

PacBio HiFi sequencing

PacBio single-molecule real-time (SMRT) libraries were constructed following the manufacturer's standard protocols. A sequencing library with an insert size of approximately 15 kb was generated using 3 μg of high-quality genomic DNA and sequenced on the PacBio Revio system. A total of 81.69 Gb HiFi clean reads (approximately 4.35 million reads) were obtained for subsequent genome assembly (Table S1).

Oxford Nanopore Technologies (ONT) sequencing

Genomic DNA was randomly fragmented using a Covaris ultrasonic crusher, and sequencing libraries were prepared with the SQK-LSK109 ligation kit according to standard procedures. Purified libraries were loaded onto R9.4 Spot-On Flow Cells and sequenced on the PromethION platform (Oxford Nanopore Technologies, UK). Raw sequencing reads were base-called using Guppy software (v0.3.0). Ultimately, 59.55 Gb of ONT clean data

(approximately 3.59 million reads) were generated, with an average read length of 16.57 kb and an N50 length of 68.15 kb (Table S2).

Hi-C sequencing

The Hi-C library was enriched, tailed with A bases, and amplified by PCR with 12–14 cycles. Qualified libraries were sequenced on the DNBSEQ-T7 platform (BGI Inc., China) under the PE150 mode. In total, 327.72 Gb of Hi-C clean data (approximately 2.18 billion reads) were acquired to support chromosome-level genome scaffolding (Table S3).

Full-length transcriptome sequencing

High-quality total RNA from multiple tissues was mixed equally for full-length cDNA library construction following the manufacturer's instructions. The constructed library was sequenced on the Oxford Nanopore PromethION platform. Finally, 16.46 Gb of full-length transcriptome clean data (approximately 11.99 million reads) were obtained for subsequent gene prediction and functional annotation (Table S4).

Genome size estimation and *de novo* assembly

Genome size, heterozygosity, and repeat content estimation

The 19-mer frequency distribution was calculated using KMC (v3.2.4)[14]. Genome size, heterozygosity, and repetitive sequence content were preliminarily estimated using GCE (v1.0.0)[15]. The estimation results were further verified and corrected using findGSE (v0.1.0)[16] and GenomeScope 2.0[17]. Comprehensive analysis showed that the haploid genome size of *C. yunnanensis* was approximately 2.82 Gb, with a heterozygosity rate of 1.54% (Table S5).

De novo genome assembly and optimization

The initial contig assembly was generated using hifiasm (v0.19.8 - r602) with PacBio HiFi reads, and the haplotype-resolved assembly was selected for subsequent genomic analysis[18]. Hi-C clean reads were aligned to the assembled contigs using Juicer[19], and chromosome-level scaffolding was performed with 3D-DNA[20]. Potential assembly errors and chromosome boundary abnormalities were manually inspected and corrected using Juicebox[21]. After manual curation, the final chromosomal frameworks and unanchored contigs were obtained. Additionally, the chloroplast and mitochondrial genomes of *C. yunnanensis* were assembled using the GetOrganelle pipeline[22].

To improve the continuity and completeness of the genome assembly, gap filling was performed using quarTeT[23] based on HiFi and ONT reads. Chromosome terminal sequences were extended, and telomere regions were recovered by aligning HiFi reads to chromosomal ends. Redundant sequences and exogenous contaminant contigs were removed using Redundans[24], with a focus on eliminating redundant fragments derived from organelle genomes and rDNA repeats. Low-coverage fragments and overlapping haplotype sequences were further filtered to ensure the purity and accuracy of the final assembly.

Referring to the published chromosome-level genome of *C. sinensis*[25], the assembled chromosomes were numbered chr01–15[ab], in which the a/b subgroups do not represent parental haplotypes. The subgroup-a assembly was retained as the primary genome result, and all chromosomal sequences were standardized to a consistent orientation. In total, 30 chromosomes were assembled, covering 99.52% of the total genome length (Table S6). Four chromosomes (chr1b, chr8b, chr13b, and chr15b) were gap-free with complete terminal sequences, representing the highest-quality assembled units in this genome (Table S7).

Genome annotation

Repetitive sequence annotation

Repetitive sequences and transposable elements (TEs) were annotated via a combination of *de novo* prediction and homology-based searching. First, a species-specific TE library was constructed using EDTA[26] with parameters "-sensitive 1 -anno 1". RepeatMasker was subsequently used to systematically identify and mask repetitive regions across the whole genome. In total, 3.76 Gb of sequences (accounting for 67.03% of the entire genome) were annotated as repetitive elements, among which LTR (55.38%) and TIR (7.17%) superfamilies were the most abundant TE categories (Tables S8, S9).

Homology-based protein evidence

To support accurate gene structure annotation, publicly available protein sequences of 11 representative species, including *C. sinensis*, *Diospyros lotus*, *Rhododendron ovatum*, *Gilia yorkii*, *Vitellaria paradoxa*, *Aegiceras corniculatum*, *Coffea canephora*, *Solanum lycopersicum*, *Cornus controversa*, *Vitis vinifera*, and *Arabidopsis thaliana*, were collected. A total of 339,202 non-redundant protein sequences were retained as homology evidence for gene prediction.

Transcriptome-based evidence

Multiple transcriptome datasets were integrated to generate comprehensive transcript evidence. For short-read RNA-seq data, *de novo* transcript assembly was performed using Trinity[27], and genome-guided assembly was conducted by mapping clean reads to the assembled genome with HISAT2[28] followed by transcript reconstruction using StringTie[29] (Table S10). For full-length transcriptome data, reads were aligned to the reference genome using minimap2[30], and transcripts were further assembled with StringTie[29] (Table S11).

Gene structure prediction and integration annotation

The PASA pipeline[31] was used to polish transcript structures and identify full-length genes. High-confidence full-length gene models were used to train the ab initio prediction tools AUGUSTUS[32] and SNAP[33]. The MAKER2 pipeline[34] was applied to integrate three types of evidence: ab initio prediction, homologous protein alignment, and transcriptome assembly. During MAKER2 annotation, RepeatMasker was used to mask repetitive sequences, while BLASTN, BLASTX, and Exonerate[35] were used to optimize gene model alignment.

To improve annotation accuracy, EvidenceModeler (EVM)[36] was adopted to integrate and consolidate gene models generated from MAKER2 and PASA. TEsorter[37] was used to identify TE-derived protein domains, which were masked in the EVM integration process to avoid false gene predictions from transposable elements.

Non-coding RNA annotation and functional annotation of protein-coding genes

Non-coding RNAs (ncRNAs) were annotated using multiple specialized tools. tRNAScan-SE[38] was used to predict tRNA genes, Barrnap was employed for rRNA identification, and RfamScan was used to annotate other ncRNAs (e.g., miRNA, snRNA). All ncRNA annotations were merged and redundant loci were removed, with priority given to retaining protein-coding gene coordinates to resolve overlapping annotations.

Functional annotation of protein-coding genes was performed via three complementary strategies: (1) Functional classification, including GO and KEGG terms, was assigned using eggNOG-mapper[39] based on homologous alignment with the eggNOG database. (2) Homology searching was conducted using DIAMOND[40] against public protein databases (Swiss-Prot, TrEMBL, NR, and Arabidopsis protein database), with a cutoff of 30% sequence

identity and E-value < 1e-5. (3) Conserved domains and functional motifs were identified using InterProScan[41] by searching against multiple sub-databases (PRINTS, Pfam, SMART, PANTHER).

Transcriptome data processing and differential expression analysis

Total RNA was extracted and qualified for cDNA library construction, which were then sequenced on the Illumina NovaSeq 6000 platform using the PE150 mode to obtain gene expression data across different floral developmental stages.

Raw sequencing reads were quality-filtered using fastp (v0.21.0) to remove reads containing more than five N bases, those with a proportion of low-quality bases ($Q < 15$) exceeding 40%, reads with adapter contamination, and reads shorter than 30 bp, yielding clean reads. The clean reads were aligned to the *C. yunnanensis* reference genome assembled in this study using STAR (v2.7.9a), and transcript assembly and quantification were performed using StringTie (v2.1.4), with gene expression levels measured in FPKM (Fragments Per Kilobase of transcript per Million mapped reads).

Differential expression analysis was conducted using DESeq2 (v1.36.0) for pairwise comparisons among the S1, S2, and S3 groups, with $|\log_2 (\text{fold change})| \geq 1$ and adjusted *P* value (padj) < 0.05 as the thresholds for screening significantly differentially expressed genes (DEGs). GO functional annotation and KEGG pathway enrichment analysis of DEGs were performed using clusterProfiler (v3.14.3). In addition, three software tools—CNCI, CPC2, and PLEK—were integrated to predict the coding potential of newly assembled transcripts, and candidate lncRNAs with a length ≥ 200 nt were screened.

Results

Assessment of the assembled genome

A high-quality chromosome-scale reference genome of *C. yunnanensis* was assembled using multi-omics datasets of 81.69 Gb PacBio HiFi reads, 59.55 Gb Oxford Nanopore long reads, and 327.72 Gb Hi-C reads (Fig. 2; Tables S1, S2). Two haplotype-resolved genome assemblies were obtained with a combined total length of 5.61 Gb. Haplotype A was 2.89 Gb in size with a contig N50 of 79.14 Mb, while haplotype B reached 2.72 Gb with a contig N50 of 114.74 Mb (Table 1). The total assembled genome size was slightly smaller than the k-mer-based genome estimate (Table S5), which most likely results from the incomplete reconstruction of large tandem-repeat arrays at chromosomal termini. The primary assembly yielded a scaffold N50 of 190.51 Mb and an overall contig N50 of 87.92 Mb, with only 84 assembly gaps remaining. Ultimately, contigs were anchored to 30 pseudochromosomes, achieving an anchoring rate of 99.52%.

Canonical plant telomeric repeat motifs $(TTTAGGG)_n$ were identified at the terminal regions of most pseudochromosomes. Tandem arrays of 18S and 5S rDNA were also detected across chromosomes (Fig. 3a). Furthermore, a highly abundant tandem-repeat family, designated cent2 and characterised by distinct monomer sequences, was present on all pseudochromosomes. This repeat family is hypothesised to contribute to centromere formation and functional regulation (Fig. 3a). The identification of these key chromosomal structural elements provides valuable sequence resources for future studies of chromosome segregation, centromere evolution, and karyotypic diversification in *C. yunnanensis*.

We adopted multiple complementary approaches to evaluate the accuracy, completeness, and reliability of the genome assembly (Table S12). PacBio HiFi, Iso-Seq, and RNA-Seq reads were mapped against the assembled contigs using BWA and Minimap2, respectively. Read-depth statistics revealed that 98.46% of the assembled genome was covered by HiFi reads at a depth of $\geq 5\times$, which greatly reduced the risks of misassembly and sequence loss. Read-depth distributions showed uniform genome-wide coverage, with no obvious signatures of sequence redundancy or haplotype collapse (Fig. 3b), confirming high accuracy and effective haplotype separation for both assemblies.

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis indicated high completeness of the *C. yunnanensis* genome (Table S13). Up to 99.20% of conserved core orthologs (including single-copy and multi-copy genes) were recovered, with merely 0.40% of core genes missing. No prominent heterozygous peaks were observed within regions containing BUSCO single-copy and multi-copy genes, pointing to homogeneous gene-model distributions (Fig. 4a, b). Hi-C chromatin-interaction analysis further corroborated assembly quality. The Hi-C contact heatmap generated with Juicer exhibited strong intra-chromosomal interactions concentrated along the diagonal, accompanied by low background noise and high chromosomal resolution (Fig. 4c). Clear pairing signals between homologous chromosomes were observed, and no large-scale misplacement, misanchoring, or inter-chromosomal misjoins were detected (Fig. 4d).

Taken together, the chromosome-scale genome assembly of *C. yunnanensis* features few gaps, well-assembled structural elements (telomeres and rDNA), high read-mapping coverage, excellent BUSCO completeness, absence of haplotype collapse, and high-precision

pseudochromosome anchoring. Multiple lines of evidence demonstrate that this assembly is high-quality and robust. It provides solid genetic resources for subsequent gene-model annotation, dissection of the phenylpropanoid metabolic pathway, evolutionary investigations, and molecular-breeding applications in *Camellia*.

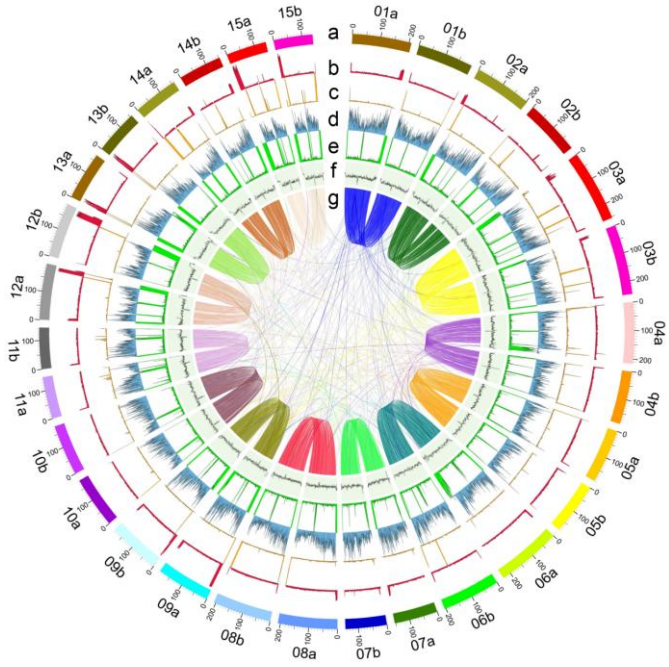


Fig. 2 The genomic landscape of 30 pseudo-chromosomes of *C. yunnanensis* (labeled as Chr01–15 [a, b]). Circles from outside to inside: (a) Chromosome length. (b) Density of Class I transposable elements (TEs), including long terminal repeats (LTRs) and long/short dispersed elements. (c) Density of Class II TEs (DNA and Helitrons). (d) Density of coding genes (mRNA). (e) Proportion of tandem repeat sequences. (f) GC content. (g) Synteny blocks.

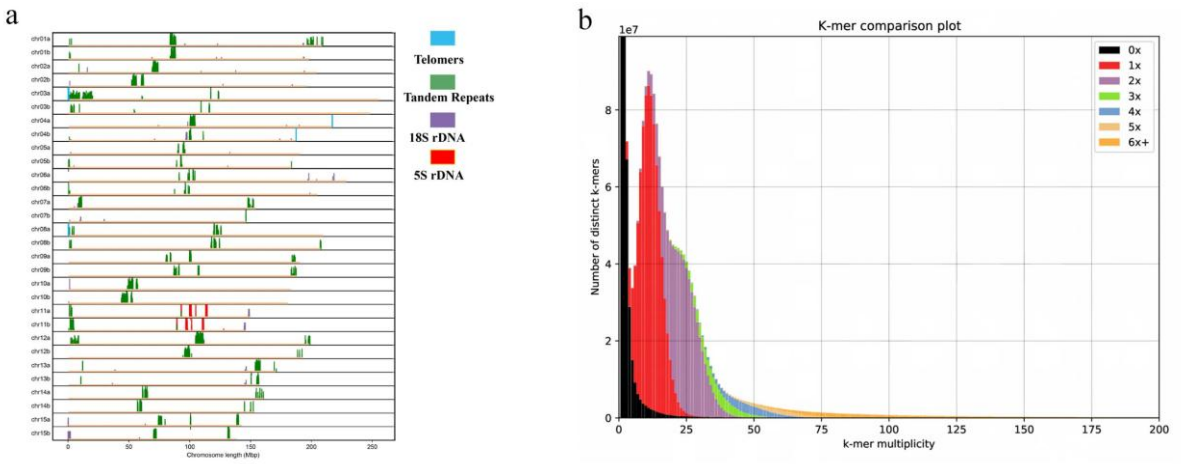
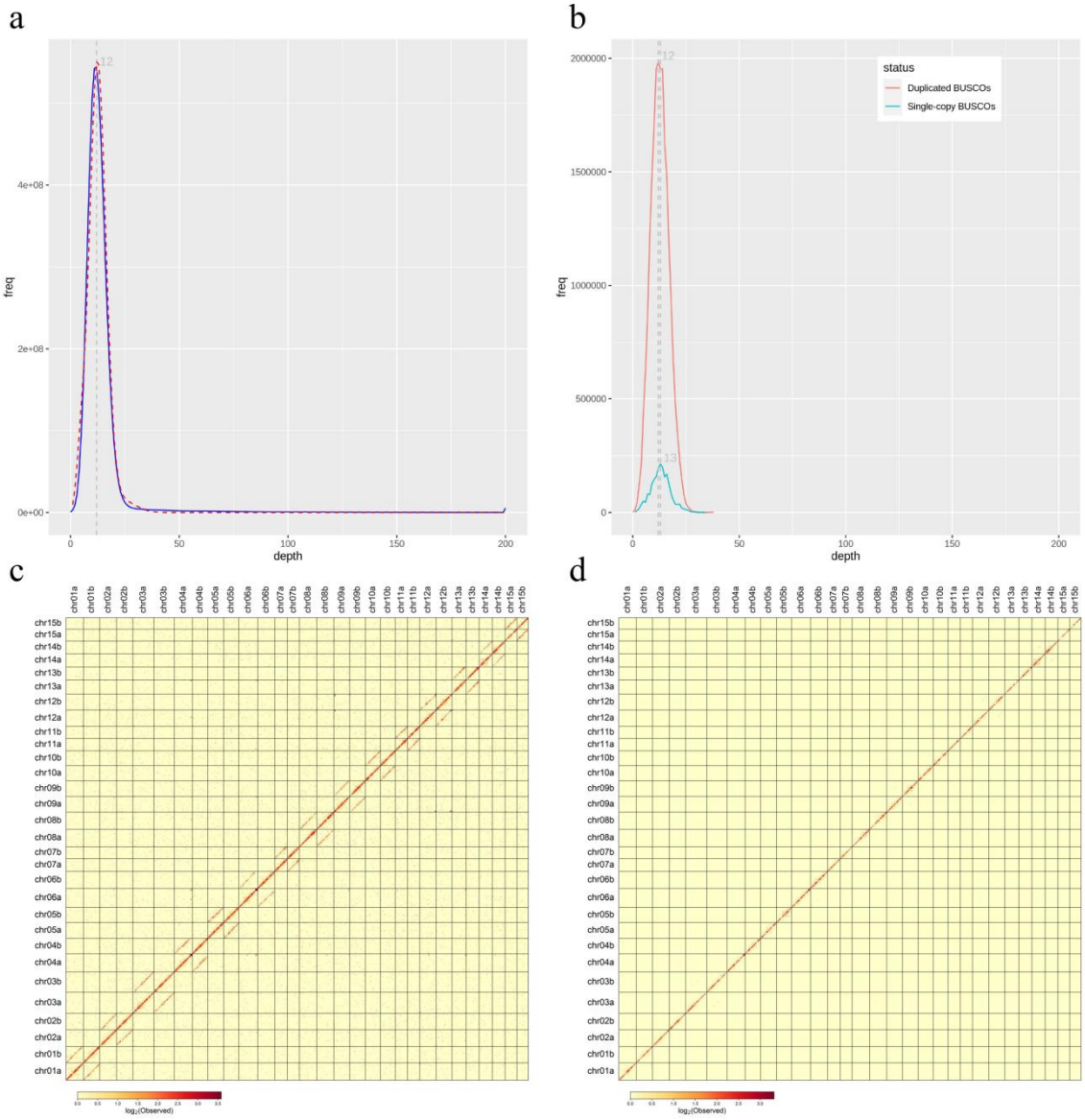


Fig. 3 (a) Profiles of repetitive sequence distribution across the chromosomes of *C. yunnanensis*, highlighting telomeric TTTAGGG sequences, tandem repeats, 18S rDNA, and 5S rDNA. The vertical scale denotes the frequency of repetitive sequences in 20 kb segments. Black triangles signify the positions of gaps. (b) Graphical representation of copy number variation spectra for genome assembly validation against HiFi sequencing data.



315 **Fig. 4** Distribution of genome-wide. (a) BUSCO core gene regions. (b) Coverage depth evaluated using HiFi

316 data; Hi-C interaction heatmaps for haplotypes A and B, depicting interactions. (c) Reads with mapping

317 quality of 0 or higher (including duplicates). (d) Mapping quality of 1 or higher (excluding duplicates).

318 Note: The color scale represents interaction intensity, with yellow signifying weak and red indicating strong

319 interactions.

320

Table 1 Statistics of the *C. yunnanensis* genome assembly.

Parameter	Genome	Haplotype A	Haplotype B
Total assembly size (bp)	5,605,676,745	2,885,072,616	2,720,604,129
GC content (%)	38.63	38.61	38.65
Total number of contigs	174	124	50
Maximum contig length (bp)	207,122,663	179,847,349	207,122,663
Minimum contig length (bp)	50,031	50,031	1,746,982
Contig N50 (bp)	87,919,763	79,142,960	114,742,480
Contig N90 (bp)	25,051,436	27,884,668	25,051,436
Total number of scaffolds	90	75	15
Maximum scaffold length (bp)	254,891,473	254,891,473	247,685,858
Minimum scaffold length (bp)	50,031	50,031	138,012,220
Scaffold N50 (bp)	190,507,856	197,953,991	187,390,795
Scaffold N90 (bp)	145,843,665	149,186,312	145,706,149
Gap number	84	49	35

Quality assessment of gene annotation

Building upon the high-quality chromosome-scale genome assembly, we conducted comprehensive gene annotation for *C. yunnanensis* by combining *de novo* gene prediction, homology-based protein evidence, and transcriptome-supported evidence. Repetitive-element annotation showed that repeats occupy 67.03 % of the whole-genome sequence. Long-terminal-repeat retrotransposons (LTRs) constitute the predominant repeat fraction, representing 55.38 % of the assembled genome (Tables S8, S9).

A total of 96,898 protein-coding genes were predicted, and functional assignments were generated for all predicted gene models (Tables S14, S15). We further assessed the completeness and accuracy of the annotated gene set using the BUSCO ortholog dataset (Table S16). The BUSCO evaluation recovered 99.30 % of complete core orthologs, of which 2.20 % were single-copy orthologs and 97.10 % were multi-copy orthologs. Only 0.20 % of gene models were fragmented, with a missing-gene proportion of merely 0.50 %, reflecting very low fragmentation and gene loss in this annotation. Collectively, these metrics confirm that our

gene-model annotation is complete, precise, and robust, and is well-suited for downstream analyses, including functional-gene mining and metabolic-pathway dissection.

Transcriptome sequencing analysis

Quality assessment of transcriptome data

To dissect the molecular regulatory mechanisms governing flower-color and floral-scent formation during floral development in *C. yunnanensis*, we performed transcriptome sequencing for three representative flower-developmental stages: flower bud (S1), initial bloom (S2), and full bloom (S3) (Table S17). A total of 390,912,918 raw reads were generated. After rigorous quality filtering, 390,682,666 clean reads were retained, corresponding to an effective read retention rate of 99.9 %. Quality metrics showed that Q20 base accuracy exceeded 97 % and Q30 accuracy was greater than 93 % across all samples. The GC content was stable, ranging from 44.23 % to 44.50 %, reflecting consistent base composition. Correlation analysis of biological replicates demonstrated high sample consistency and good technical reproducibility. Clean reads were mapped to the chromosome-scale reference genome, with mapping rates varying from 84.83 % to 97.81 % among samples. The overall high mapping rates validate the suitability of our reference genome for downstream differential-expression profiling and functional-gene mining across floral developmental stages (Table S18).

Transcript assembly and functional annotation analysis

Transcript reconstruction was carried out using StringTie based on quality-filtered RNA-seq reads, yielding 26,556 valid assembled transcripts with a transcript N50 of 1,157 bp, which indicates satisfactory assembly completeness and continuity. In total, 78.3 % of assembled transcripts were functionally annotated (Table S19). Gene-ontology (GO) enrichment revealed

that newly assembled transcripts were predominantly enriched in core functional categories, including "catalytic activity" and "metabolic process" (Fig. S1), consistent with active secondary-metabolic activity accompanying floral development. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment showed significant enrichment in fundamental metabolic pathways such as carbohydrate metabolism, amino-acid metabolism, and lipid metabolism (Fig. S2), which supply metabolic substrates for floral-organ growth and substance accumulation. Eukaryotic Orthologous Groups (KOG) classification further annotated numerous transcripts associated with the biosynthesis, transport, and catabolism of secondary metabolites (Fig. S3), suggesting active secondary-metabolic regulation during flower development in *C. yunnanensis*.

Furthermore, we integrated three coding-potential prediction tools (CNCI, CPC2, and PLEK) to identify long non-coding RNAs (lncRNAs) from reconstructed transcripts. We recovered 1,287 putative lncRNAs with length ≥ 200 nt and no protein-coding potential (Fig. S4). This lncRNA dataset provides valuable resources for future investigations of non-coding regulatory mechanisms underlying floral development.

Global differential-expression analysis across flowering stages

Sample-correlation and principal-component analysis (PCA) showed that Pearson correlation coefficients between biological replicates within each stage were all higher than 0.99, demonstrating excellent intra-group reproducibility. Samples from distinct floral stages were well separated in PCA space (PC1 = 45.6 %, PC2 = 29.4 %), indicating substantial transcriptomic divergence among petal samples at different developmental stages and validating the rationality of sample grouping for differential-gene analysis (Fig. 5a, 5b).

Pairwise differential-expression gene (DEG) comparisons across the three floral stages revealed extensive transcriptomic reprogramming during flower development. We identified 15,505 DEGs for the S1 versus S3 comparison (6,528 up-regulated, 8,977 down-regulated); 8,779 DEGs for S2 versus S1 (3,756 up-regulated, 5,023 down-regulated); and 12,685 DEGs for S3 versus S2 (7,714 up-regulated, 4,972 down-regulated). The S1-versus-S3 comparison yielded the largest number of DEGs and the greatest expression divergence, implying that petal regulatory networks undergo profound remodelling from flower-bud differentiation to full anthesis, accompanied by complex physiological, metabolic, and developmental transitions (Fig. 5c – f).

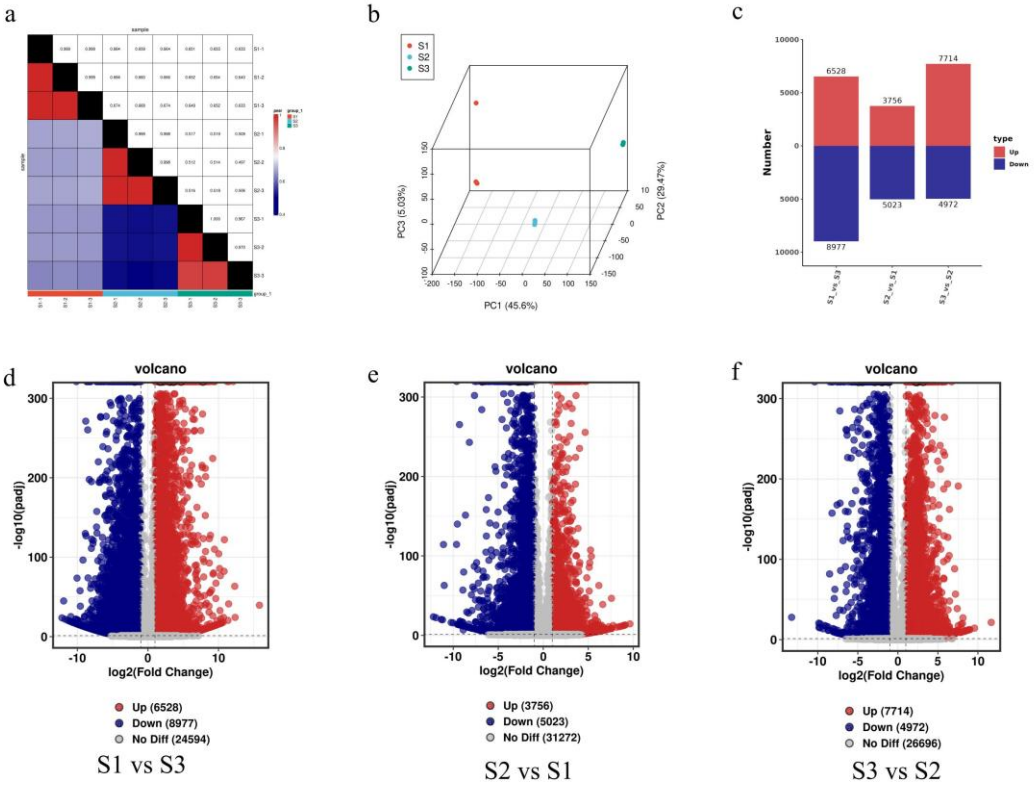


Fig. 5 Transcriptomic analysis of *C. yunnanensis* at different flowering stages. (a) Pearson correlation heatmap among the three groups of samples. (b) Principal component analysis (PCA) plot among the three groups of samples. (c) Venn diagram of differentially expressed genes (DEGs) among the three comparison

groups. (d) Volcano plot of S1 vs S3. (e) Volcano plot of S2 vs S1. (f) Volcano plot of S3 vs S2.

Note: The group before "vs" represents the control group, and the group after "vs" represents the experimental group. The same applies below.

Functional enrichment analysis of differentially expressed genes

GO enrichment uncovered distinct temporal functional signatures for DEGs derived from different pairwise comparisons (Fig. S5). For S1 vs S3, DEGs were enriched in terms related to ribosomal structural components, transporter activity, and hormone-response pathways, reflecting marked shifts in protein biosynthesis and hormonal signalling from bud to full-bloom stages. For S2 vs S1, enrichment was observed for oxidoreductase activity and amino-acid transmembrane transport, highlighting predominant metabolic and redox regulation at the initial-bloom stage. For S3 vs S2, terms associated with chloroplasts, thylakoids, and photosystems were significantly enriched, pointing to activated photosynthetic-related pathways during mid-to-late-flower development. In addition, carbohydrate-metabolism and hormone-response terms were consistently enriched across all three comparisons, suggesting that sugar metabolism and hormone-signalling cascades operate throughout flower development and constitute core regulators of petal morphogenesis and metabolite accumulation.

KEGG enrichment further illustrated core metabolic regulatory dynamics during floral development (Fig. 6; Tables S20–22). DEGs from all three comparisons were significantly enriched in the ribosome (map03010), biosynthesis of secondary metabolites (map01110), and plant-hormone signal-transduction (map04075) pathways. This result demonstrates that protein synthesis, secondary-metabolite modulation, and hormone signalling form the metabolic

foundation supporting continuous petal development. Notably, the phenylpropanoid-biosynthesis pathway (map00940) was significantly enriched in S2 vs S1 ($p < 0.05$), together with concomitant enrichment of the flavonoid-biosynthesis pathway (map00941). This indicates specific activation of phenylpropanoid-derived secondary metabolism during the transition from bud to initial bloom, representing a key metabolic feature of early floral development. Meanwhile, photosynthesis-related pathways (map00195) were strongly up-regulated in S3 vs S2, consistent with GO enrichment results and confirming enhanced photosynthetic metabolism in mid-to-late-flower development to supply energy and substrates for flower opening.

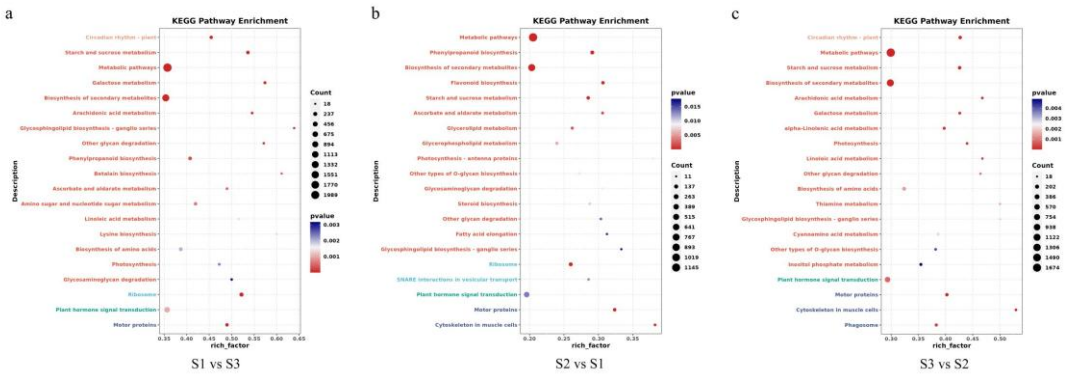


Fig. 6 KEGG enrichment analysis of differentially expressed genes.

Transcriptional features of the phenylpropanoid pathway underlying flower-color and floral-scent divergence

The phenylpropanoid metabolic pathway (map00940) is a core shared upstream pathway governing flower color and floral scent secondary metabolism in plants. Based on stage-resolved transcriptome profiles across the flowering period, this study systematically analyzed the expression dynamics of core structural genes in the phenylpropanoid pathway across the flower bud, initial bloom, and full bloom stages, and, combined with differential expression

analysis, screened key candidate genes associated with the white flower and floral scent traits of *C. yunnanensis* (Fig. 7). The results showed that the 37 structural genes involved in the phenylpropanoid biosynthesis pathway were significantly differentially expressed among the three stages (Table S23).

Upstream rate-limiting genes of the phenylpropanoid pathway, namely *PAL*, *C4H*, and *4CL*, were generally highly expressed at S1, consistent with the KEGG enrichment results. This suggests global activation of the phenylpropanoid pathway in early floral development, providing abundant precursors for downstream biosynthesis of flavonoids, anthocyanins, and volatile aromatic compounds. Among them, *PAL* (*Cayun03aG0056200*) peaked at S1 and was persistently down-regulated afterwards; *C4H* (*Cayun11aG0102500, Cayun02aG0036600, Cayun01aG0242600*) maintained high transcript abundance at S1. Members of the *4CL* family exhibited divergent expression patterns, with *Cayun09aG0003300* and *Cayun08aG0365300* specifically highly expressed at S1, suggesting functional divergence among different copies in the regulation of flower development.

Within the anthocyanin-biosynthesis branch, upstream structural genes including *CHS* (*Cayun10aG0253700, Cayun04aG0253500, Cayun10aG0253800*), *CHI* (*Cayun03aG0112400, Cayun09aG0118900, Cayun09aG0080500*), *F3H* (*novel1145*, *Cayun10aG0026800*), and *F3'H* (*Cayun15aG0051500*) were highly expressed at S1, demonstrating that the upstream transcriptional programme for anthocyanin biosynthesis is already initiated in early petal development. Nevertheless, the terminal anthocyanin-biosynthesis genes *DFR* (*Cayun08aG0100700, Cayun01aG0032500*) and *ANS* (*Cayun13aG0162200, Cayun14aG0209400, Cayun14aG0209500*) showed brief expression at S1 followed by

sustained and pronounced down-regulation, reaching extremely low transcript levels at S3. This directly blocks the biosynthesis and accumulation of anthocyanin end-products, which aligns perfectly with the pure-white petal phenotype of *C. yunnanensis*. Meanwhile, the key flavonol-branch gene *FLS* (*Cayun08aG0238500*, *Cayun10aG0167400*) was significantly up-regulated at S3 and displayed an antagonistic expression pattern relative to *DFR*. Since *DFR* and *FLS* compete for the common substrate dihydroflavonols, such transcriptional antagonism implies that phenylpropanoid precursors are preferentially channelled toward flavonol biosynthesis at S3, thereby suppressing anthocyanin metabolic output.

Contrasting with the stage-specific repression of the anthocyanin branch, the floral-scent biosynthetic branch exhibited sustained activation in early flower development. The key benzenoid-aroma gene *C3H* (*Cayun12aG0052900*) showed high expression at S2, and multiple *CCoAOMT* family members were co-upregulated from S1 to S2, among which *Cayun06aG0009000* reached its expression maximum at S2. These observations indicate that structural genes for volatile benzenoid biosynthesis remain transcriptionally active from S1 to S2, providing stable enzymatic machinery and precursor supply for the production of characteristic benzenoid volatiles such as eugenol and isoeugenol, and ultimately contributing to the distinctive floral-scent trait of *C. yunnanensis*.

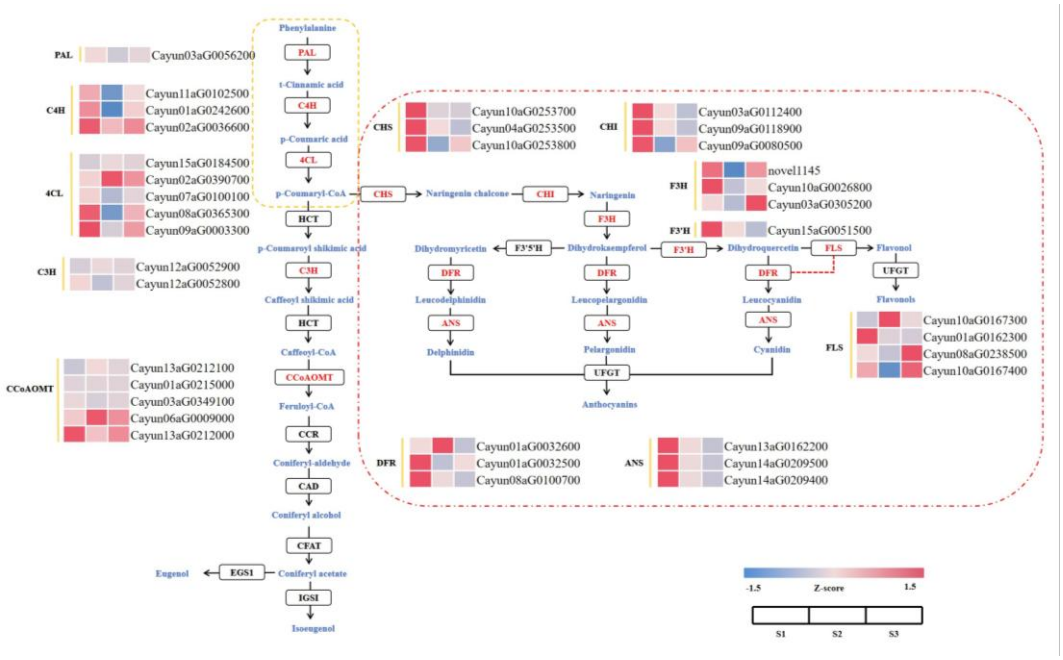


Fig. 7 Expression profiles of key enzyme genes in the phenylpropanoid metabolic pathway across three stages of flower development.

Discussion

High-quality chromosome-scale genome assembly of *C. yunnanensis*

Here, we report the first chromosome-scale, haplotype-resolved genome assembly for *C. yunnanensis*, comprising 30 anchored pseudochromosomes with a total assembled size of 5.61 Gb. The assembly exhibits completeness and continuity comparable to previously published high-quality reference genomes of other *Camellia* species. The genome of *C. yunnanensis* is substantially larger than those of cultivated relatives *C. sinensis* (3.1 Gb)[42] and *C. oleifera* (2.95 Gb)[4], exhibiting genomic structural features distinctly different from those of cultivated *Camellia* species. Genome structure analysis indicated that this genome expansion was largely driven by the massive proliferation of repetitive sequences, with total repeats constituting 67.03% of the genome, among which LTR retrotransposons accounted for 55.38% and represented the

core elements driving genome expansion. However, plant genome size is not a direct criterion for inferring the antiquity of a species' evolutionary history, but rather an evolutionary outcome of the long-term dynamic equilibrium between transposon burst amplification and host genome suppression and deletion mechanisms[43]. Therefore, the exceptionally large genome of *C. yunnanensis* suggests that it may have experienced a unique history of transposon amplification, harboring germplasm specificity and genetic resource uniqueness worthy of further exploration.

Notably, the amplification of LTR retrotransposons is a known driver of plant genome size variation, as reported in large-genome taxa such as *Lilium sargentiae* and *Allium cepa*[44, 45] , and recent LTR expansion events in tea plants have also been demonstrated to be closely associated with dynamic genome evolution[46]. However, whether the large-scale LTR expansion has influenced the expression regulation of key genes in the phenylpropanoid pathway—by providing novel cis-regulatory elements or altering the chromatin environment of neighboring genes—thereby participating in the formation of white flower color and floral scent traits, remains an open question. This hypothesis warrants further testing through systematic identification of LTR-adjacent genes and integrated analysis with flower developmental transcriptomic data.

In terms of assembly quality, the scaffold N50 of our assembly reached 190.51 Mb with a BUSCO completeness of 99.20%, outperforming several published wild *Camellia* genomes, including *C. nitidissima* (scaffold N50 = 187.52 Mb, BUSCO = 98.80%), *C. fascicularis* (scaffold N50 = 191.08 Mb, BUSCO = 99.13%), and *C. pitardii* (contig N50 = 146.18 Mb, BUSCO = 96.34%)[47-49]. This high-accuracy, haplotype-resolved genomic resource not only fills the gap in genomic research on *C. yunnanensis*, but also provides a reliable sequence

foundation for subsequent elucidation of the regulatory mechanisms of the phenylpropanoid pathway and for mining key functional genes associated with white flower color and floral scent traits.

Phenylpropanoid-pathway reprogramming underpinning white-fragrant flower traits in *C. yunnanensis*

Reprogramming of phenylpropanoid pathway branches is a core regulatory mechanism underlying the differentiation of flower color and floral scent phenotypes in ornamental plants [11, 50]. Based on temporal transcriptomic data from three key floral developmental stages of *C. yunnanensis*, this study systematically dissected the expression dynamics of 37 core structural genes in the phenylpropanoid pathway and preliminarily screened core candidate genes—including *PAL*, *C4H*, *4CL*, *DFR*, *ANS*, *FLS*, *C3H*, and *CCoAOMT*—potentially regulating white flower color and floral scent traits, revealing pronounced temporal differentiation of branch-specific regulation during flower development in *C. yunnanensis*. The phenylpropanoid pathway was broadly activated at the bud stage (S1), with upstream genes and early anthocyanin biosynthetic genes synchronously expressed, conferring transcriptional capacity for anthocyanin synthesis in young petals. Pathway divergence began to emerge from S2 onward: terminal anthocyanin biosynthetic genes were gradually silenced, whereas floral scent-related genes maintained sustained activity. By the full-bloom stage (S3), the anthocyanin branch was largely shut down, while the flavonol branch was strongly activated, resulting in a significant redistribution of metabolic flux. This regulatory pattern of "early global activation followed by stage-specific branch partitioning" has also been reported in other ornamental plants such as *Narcissus tazetta* and *Lilium cernuum* [51, 52], suggesting that secondary metabolic

reprogramming during flower development may adopt a conserved regulatory strategy.

The pure white petal phenotype of *C. yunnanensis* is primarily mediated by the significant transcriptional repression of *DFR* and *ANS* at the full-bloom stage. *FLS* was significantly upregulated and competes with *DFR* for the shared substrate dihydroflavonol, channeling metabolic flux preferentially toward the flavonol branch and markedly suppressing pigment biosynthesis. Comparison with published studies on flower color in *Camellia* reveals that silencing of the anthocyanin pathway is a shared regulatory feature underlying white petal formation in this genus. For instance, in white petals of *C. oleifera*, the anthocyanin structural genes *Co4CL1*, *CoF3H1*, and *CoANS* were significantly downregulated compared with those in colored petals[53]; the white flower phenotype of *C. reticulata* 'Tongzimian' is associated with upstream CHS downregulation and high ANR expression[54]. Interestingly, the repression in 'Tongzimian' occurs upstream in the pathway, whereas in *C. yunnanensis* it is concentrated at the downstream *DFR*/*ANS* node. This difference suggests that white flower evolution in *Camellia* may involve a regulatory strategy of "multi-site repression with convergent phenotypes." Furthermore, studies on flower color regulation in *C. nitidissima* further support the metabolic redistribution model proposed in this study: in *C. nitidissima* petals, FLS is significantly upregulated while anthocyanin pathway genes are suppressed, promoting substantial flavonol accumulation[55], which is highly consistent with the antagonistic *DFR*/*FLS* expression pattern identified in this study.

With respect to floral scent formation, the sustained high expression of *C3H* and multiple *CCoAOMT* members from S1 to S2 ensures a continuous supply of phenylpropanoid aroma precursors. Typical phenylpropanoid-derived aroma components, such as phenylethanol,

551 eugenol, and methyl salicylate, are commonly detected in fragrant germplasm of *Camellia* sect.
552 Theopsis, suggesting a degree of conservation in phenylpropanoid pathway-mediated floral
553 scent regulation among fragrant *Camellia* species[56]. A three-stage transcriptomic study of
554 floral scent in *C. japonica* 'High Fragrance' further provides transcript-level evidence,
555 identifying 91 candidate differentially expressed genes related to phenylpropanoid biosynthesis,
556 most of which exhibited upregulated expression patterns during flower development, consistent
557 with the progressive accumulation of volatile compounds[57]. However, the purple-flowered tea
558 mutant exhibits an opposite pattern: anthocyanins and volatile benzenoid compounds
559 (benzaldehyde, benzyl alcohol, acetophenone, 2-phenylethanol, etc.) are simultaneously
560 enhanced, with structural genes for benzenoid biosynthesis induced and FLS suppressed[13].
561 This indicates that the "white flower-floral scent coexistence" phenotype of *C. yunnanensis* is
562 not simply achieved by "weakening anthocyanin synthesis to enhance fragrance"; rather, it may
563 be achieved by silencing key terminal genes of the anthocyanin pathway to enable redistribution
564 of metabolic precursors, while simultaneously preserving the activated state of the floral scent
565 branch. By integrating stage-dependent transcriptomic dynamics, this study proposes a
566 regulatory model for the white flower color and floral scent traits of *C. yunnanensis* (Fig 8):
567 early broad activation of the phenylpropanoid pathway generates abundant secondary metabolic
568 precursors; subsequent differential regulation of downstream branches suppresses terminal
569 anthocyanin synthesis, resulting in white flowers, while preserving the activity of the floral
570 scent pathway to maintain volatile biosynthesis, collectively shaping this distinctive
571 horticultural phenotype.

572 Future studies could detect the dynamic accumulation of anthocyanins and benzenoid

aroma compounds in petals at different flowering stages by GC-MS, validate the expression patterns of candidate genes by qRT-PCR, and verify the functions of key genes such as *DFR*, *FLS*, and *CCoAOMT* through heterologous expression or transient silencing experiments, thereby providing genetic resources and a theoretical basis for molecular breeding of fragrant white-flowered *Camellia*.

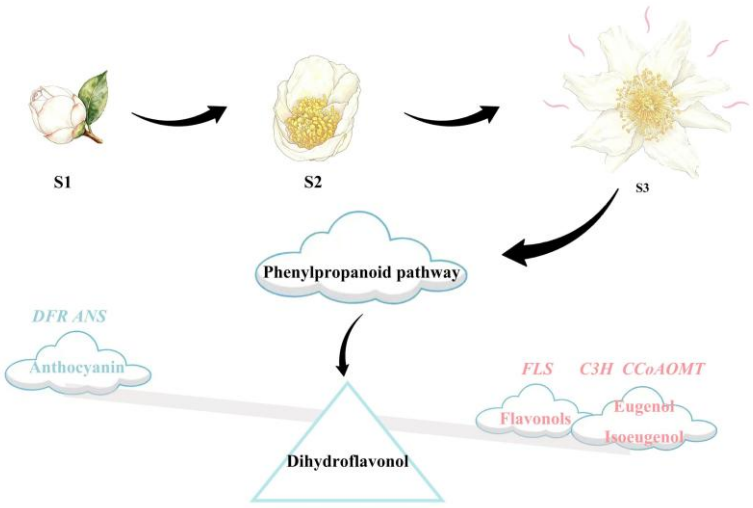


Fig. 8 Dynamic partitioning model of the phenylpropanoid metabolic branch point during flower development in *C. yunnanensis.*

Note: S1, S2, and S3 represent the bud stage, initial flowering stage, and full bloom stage, respectively. The phenylpropanoid pathway serves as the common upstream origin, and dihydroflavonol is the shared substrate for anthocyanins (*DFR*/*ANS*) and flavonols. The anthocyanin branch is active at S1, weakens at S2, and is blocked at S3, consistent with the white flower phenotype; whereas the flavonol and eugenol/isoeugenol branches remain continuously active from S1 to S3 and become dominant at S3. Blue indicates downregulation, and red indicates upregulation.

Conclusion

In this study, we constructed for the first time a chromosome-level, high-quality,

haplotype-resolved reference genome of *C. yunnanensis,* and integrated with transcriptomic data from three key floral developmental stages, dissected the molecular regulatory basis of its white-flowered and fragrant traits. The assembled diploid genome has a total size of 5.61 Gb, comprising two haplotypes anchored to 30 pseudochromosomes. The massive amplification of LTR retrotransposons (accounting for 55.38% of the genome) is the primary factor driving genome expansion, with repetitive sequences constituting 67.03% in total. The genome assembly has a scaffold N50 of 190.51 Mb and a BUSCO completeness of 99.20%, with assembly quality superior to most published wild *Camellia* genomes. Transcriptomic analysis identified a total of 37 differentially expressed structural genes of the phenylpropanoid pathway. The phenylpropanoid pathway was broadly activated at the bud stage; by the full-bloom stage, the specific downregulation of *DFR* and *ANS* blocked terminal anthocyanin biosynthesis, while *FLS* was significantly upregulated and competed with *DFR* for the shared substrate dihydroflavonol, driving metabolic precursors preferentially toward the flavonol branch and resulting in pure white petals; *C3H* and *CCoAOMT* maintained sustained high expression from the bud stage to the initial flowering stage, providing precursor assurance for the synthesis of phenylpropanoid aroma volatiles. This study enriches the omics resources of wild *Camellia* germplasm and identifies key candidate genes such as *DFR*, *ANS*, *FLS*, *C3H*, and *CCoAOMT*, providing data resources and a theoretical basis for evolutionary studies of *Camellia*, conservation and utilization of rare germplasm resources, and molecular breeding of fragrant ornamental *Camellia*.

Supplementary Information

Supplementary Table 1–23; Supplementary Fig 1–5

Acknowledgments

Not applicable.

Authors' contributions

YZ and HX contributed equally to this work. YZ: Writing – original draft, Data curation, Formal analysis. HX: Writing – original draft, Investigation, Methodology, Data curation. FY: Investigation. XW: Investigation. KJ: Investigation. WZ: Investigation. HH: Investigation. ZW: Investigation. FG: Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Visualization, Conceptualization. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (32260416), the Yunnan Province Xingdian Talent Project "Youth Talent Special Project" Support Project (YNQR-QNRC-2-19-104), and Yunnan Agricultural Basic Research Joint Special General Project (202101BD070001-095). Funders have no role in the design and publication of the manuscript.

Data availability

The raw genome and transcriptome sequencing data for *Camellia yunnanensis* (houzimu) have been deposited in the NGDC (National Genomics Data Center: https://ngdc.cncb.ac.cn/) under the BioProject number PRJCA059027, and are available via a confidential reviewer access link (https://ngdc.cncb.ac.cn/gwh/Assembly/reviewer/iAjEvTAyNFaoKaKDCAWfUAfktiUoPZaGPiWBuDjmLOkTThkNEqlZkYycaCzTTQxy). The data will be fully open for access upon

632 publication of the manuscript.

633 ## Declarations

634 ## Ethics approval and consent to participate

635 The wild *Camellia yunnanensis* plant materials used in this study were collected from Chuxiong
636 City, Yunnan Province, China. The collection and use of plant materials fully complied with
637 institutional, national, and local laws and regulations. This study conforms to the provisions of
638 the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES).

639 ## Consent for publication

640 Not applicable.

641 ## Competing interests

642 The authors declare no competing interests.

643 ## References

644 [1] Xia LF. Camellia flower. Beijing: China Architecture and Building Press; 2003.
645 [2] Zhang XQ. The classification of Chinese Camellia cultivars, test guideline and construction
646 of database for the known cultivars [master's thesis]. Beijing: Chinese Academy of
647 Forestry; 2008.
648 [3] Xia EH, Zhang HB, Sheng J, Li K, Zhang QJ, Kim C, et al. The tea tree genome provides
649 insights into tea flavor and independent evolution of caffeine biosynthesis. Mol Plant.
650 2017;10(6):866-77.
651 [4] Lin P, Wang K, Wang Y, Hu Z, Yan C, Huang H, et al. The genome of oil-Camellia and

population genomics analysis provide insights into seed oil domestication. Genome Biol. 2022;23(1):14.

[5] Shen TF, Huang B, Xu M, Zhou PY, Ni ZX, Gong C, et al. The reference genome of *Camellia chekiangoleosa* provides insights into *Camellia* evolution and tea oil biosynthesis. Hortic Res. 2022;9:uhab083.

[6] Zhang F, Feng LY, Lin PF, Jia JJ, Gao LZ. Chromosome-scale genome assembly of oil-tea tree *Camellia crapnelliana*. Sci Data. 2024;11(1):599.

[7] Peng CS, Li SJ, Wang J. Preliminary report on the selection of superior provenances of *Camellia yunnanensis* in Longling County. South China Agric. 2023;17(4):244-7.

[8] Rivers MC. *Camellia yunnanensis*. The IUCN Red List of Threatened Species 2015: e.T62059596A62059604.

[9] Lan L, Leng L, Liu W, Ren Y, Reeve W, Fu X, et al. The haplotype-resolved telomere-to-telomere carnation (*Dianthus caryophyllus*) genome reveals the correlation between genome architecture and gene expression. Hortic Res. 2024;11(1):uhad244.

[10] Wang B, Zhang R, Sun W, Yang J. A nearly telomere-to-telomere diploid genome assembly of *Firmiana kwangsiensis*, a threatened species in China. Sci Data. 2024;11(1):1394.

[11] Dudareva N, Pichersky E. Metabolic engineering of plant volatiles. Curr Opin Biotechnol. 2008;19(2):181-9.

[12] Ke Y, Wang D, Fang Z, Zou Y, Hussain Z, Iqbal S, et al. Elucidating scent and color variation in white and pink-flowered *Hydrangea arborescens* ‘Annabelle’ through multi-omics profiling. Plants. 2026;15(1):155.

[13] Mei X, Wan S, Lin C, Zhou C, Hu L, Deng C, et al. Integration of metabolome and

674 transcriptome reveals the relationship of benzenoid–phenylpropanoid pigment and aroma
675 in purple tea flowers. Front Plant Sci. 2021;12:762330.
676 [14] Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics.
677 Bioinformatics. 2017;33(17):2759-61.
678 [15] Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics by
679 analyzing k-mer frequency in de novo genome projects. arXiv:1308.2012. 2013.
680 [16] Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation
681 within human and Arabidopsis using k-mer frequencies. Bioinformatics. 2018;34(4):550-
682 7.
683 [17] Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for
684 reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.
685 [18] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly
686 using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170-5.
687 [19] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer
688 provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst.
689 2016;3(1):95-8.
690 [20] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo
691 assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.
692 Science. 2017;356(6333):92-5.
693 [21] Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox
694 provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst.
695 2016;3(1):99-101.

[22] Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21(1):241.

[23] Lin Y, Ye C, Li X, Chen Q, Wu Y, Zhang F, et al. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. Horticulture Research. 2023;10(8):uhad127.

[24] Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Research. 2016;44(2):e113.

[25] Chen JD, Zheng C, Ma JQ, Jiang CK, Ercisli S, Yao MZ, et al. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. Horticulture Research. 2020;7:63.

[26] Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology. 2019;20(1):275.

[27] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29(7):644-52.

[28] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12(4):357-60.

[29] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;33(3):290-5.

718 [30] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
719 2018;34(18):3094-100.
720 [31] Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving
721 the Arabidopsis genome annotation using maximal transcript alignment assemblies.
722 Nucleic Acids Research. 2003;31(19):5654-66.
723 [32] Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped
724 cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637-44.
725 [33] Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5(1):59.
726 [34] Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use
727 annotation pipeline designed for emerging model organism genomes. Genome Research.
728 2008;18(1):188-96.
729 [35] Slater G, Birney E. Automated generation of heuristics for biological sequence comparison.
730 BMC Bioinformatics. 2005;6(1):31.
731 [36] Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic
732 gene structure annotation using EVidenceModeler and the Program to Assemble Spliced
733 Alignments. Genome Biology. 2008;9(1):R7.
734 [37] Zhang RG, Li GY, Wang XL, Dainat J, Wang ZX, Ou S, et al. TEsorter: an accurate and
735 fast method to classify LTR-retrotransposons in plant genomes. Horticulture Research.
736 2022;9:uhac017.
737 [38] Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA
738 genes in genomic sequence. Nucleic Acids Research. 1997;25(5):955-64.
739 [39] Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al.

740 Fast genome-wide functional annotation through orthology assignment by eggNOG-
741 mapper. Molecular Biology and Evolution. 2017;34(8):2115-22.
742 [40] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.
743 Nature Methods. 2015;12(1):59-60.
744 [41] Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-
745 scale protein function classification. Bioinformatics. 2014;30(9):1236-40.
746 [42] Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, et al. Draft genome sequence of Camellia
747 sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality.
748 Proceedings of the National Academy of Sciences. 2018;115(18):E4151-8.
749 [43] Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of
750 the increase/decrease model. Cytogenetic and Genome Research. 2005;110(1-4):91-107.
751 [44] Liang Y, Gao Q, Li F, Du Y, Wu J, Pan W, et al. The giant genome of lily provides insights
752 into the hybridization of cultivated lilies. Nature Communications. 2025;16(1):45.
753 [45] Lei P, Zhang Y, Ma Y, Zhao D, Ni K, Fang F, et al. A near-complete genome assembly of
754 onion resolves its genomic architecture, origin, and evolution. Plant Communications.
755 2026;7(8):101825.
756 [46] Zhang QJ, Li W, Li K, Nan H, Shi C, Zhang Y, et al. The chromosome-level reference
757 genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in
758 driving genome size evolution. Molecular Plant. 2020;13(7):935-8.
759 [47] Wang XF, Liu TJ, Feng T, Huang HR, Zou P, Wei X, et al. A telomere-to-telomere genome
760 assembly of *Camellia nitidissima*. Scientific Data. 2025;12(1):815.
761 [48] Xiao HX, Wu XQ, Yan FJ, Zeng YJ, Jia KH, Zhang GL, et al. High quality chromosome

level genome assembly of *Camellia fascicularis*, an endangered plant in China. Scientific Data. 2025;12(1):1573.

[49] Lv Z, Wang J, Zhou L, Zou S, Ai L. Near telomere-to-telomere genome assembly of Camellia pitardii. Scientific Data. 2025;12(1):1422.

[50] Qiao ZL, Hu HZ, Yan B, Chen LQ. Advances of researches on biosynthesis and regulation of floral volatile benzenoids/phenylpropanoids. Acta Horticulturae Sinica. 2021;48(9):1815-26.

[51] Yang J, Ren Y, Zhang D, Chen X, Huang J, Xu Y, et al. Transcriptome-based WGCNA analysis reveals regulated metabolite fluxes between floral color and scent in *Narcissus tazetta* flower. International Journal of Molecular Sciences. 2021;22(15):8249.

[52] Chen S, Chen Z, Zhuang Q, Chen H. Multi-omics joint analysis reveals the mechanism of flower color and fragrance variation in *Lilium cernuum*. Frontiers in Plant Science. 2025;16:1489918.

[53] Zeng HT, Zheng T, Tang Q, Xu H, Chen M. Integrative metabolome and transcriptome analyses reveal the coloration mechanism in *Camellia oleifera* petals with different color. BMC Plant Biology. 2024;24(1):19.

[54] Qu Y, Ou Z, Yong QQ, Yao X, Luo J. Coloration differences in three *Camellia reticulata* Lindl. cultivars: 'Tongzimian', 'Shizitou' and 'Damanao'. BMC Plant Biology. 2024;24(1):18.

[55] Feng Y, Li J, Yin H, Shen J, Liu W. Multi-omics analysis revealed the mechanism underlying flavonol biosynthesis during petal color formation in *Camellia nitidissima*. BMC Plant Biology. 2024;24(1):847.

784 [56] Fan Z, Li J, Li X, Yin H. Composition analysis of floral scent within genus Camellia
785 uncovers substantial interspecific variations. Scientia Horticulturae. 2019;250:207-13.
786 [57] Chen X, Zhang X, Li Y, Tian X, Tian X, Zhao H, et al. Molecular mechanisms underlying
787 floral fragrance in *Camellia japonica* 'high fragrance': a time-course assessment. Frontiers
788 in Plant Science. 2024;15:1461442.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx