Chromosome-Level Genome Assembly of Ventilago leiocarpa Benth. (Rhamnaceae)

Song Guo
Guangxi Science & Technology Normal University
Shengzhen Chen
Fujian Agriculture and Forestry University
Ruoyu Liu
Fujian Agriculture and Forestry University
Yanan Wang
Fujian Agriculture and Forestry University
Peng Zhang
Guangxi Science & Technology Normal University
Qinglong Yao
Fujian Agriculture and Forestry University
Yousheng Que
Fujian Agriculture and Forestry University
Linwei Yang
Fujian Agriculture and Forestry University
Ping Zheng
zhengping13@mails.ucas.ac.cn

Fujian Agriculture and Forestry University

data-descriptor

Keywords:

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-7925269/v1

Additional Declarations: No competing interests reported.

Abstract

Ventilago leiocarpa Benth.is a medicinal plant in the Rhamnaceae family, whose dried roots and rhizomes are used in traditional Chinese medicine. In this study, we generated a high-quality chromosome-level genome assembly of *V. leiocarpa* using PacBio HiFi long reads and Hi-C data. The final assembly was 227.7 Mb in size, consisting of 27 scaffolds, of which 12 pseudochromosomes anchored 97.94% of the total genome, with a contig N50 of 18.9 Mb and a GC content of 35.31%. BUSCO analysis showed 98.8% completeness, indicating high genome integrity. A total of 24,973 protein-coding genes were predicted, with 95.8% complete BUSCOs in the annotated gene set. Transcriptomic data from roots, stems, and leaves exhibited > 96% mapping rates, supporting annotation accuracy. Comparative genomic analysis revealed close evolutionary relationships with *Sageretia thea*[1] and *Rhamnella rubrinervis* [2]. This high-quality genome provides an essential resource for investigating the biosynthesis of bioactive compounds, functional genomics, and molecular breeding of *V. leiocarpa* and related medicinal plants.

Background & Summary

Ventilago leiocarpa Benth. (Rhamnaceae), commonly known as Zi Jiuniu, is a climbing shrub native to tropical and subtropical regions of Asia. In southern China, it is an important medicinal plant, with its roots and rhizomes traditionally used in Yao ethnomedicine to treat inflammation, bruises, and rheumatism [3,4]. Phytochemical studies have identified a range of bioactive compounds in *V. leiocarpa*, including flavonoids, triterpenoids, and naphthopyran derivatives, which are believed to contribute to its anti-inflammatory and antioxidant properties. However, the molecular mechanisms underlying the biosynthesis and regulation of these secondary metabolites remain largely unexplored.

From a phylogenetic perspective, *V. leiocarpa* belongs to the genus *Ventilago* within the Rhamnaceae family. However, *Ventilago* is underrepresented in phylogenetic studies of this family. Currently, the Rhamnaceae phylogenetic tree includes chloroplast genome data from only 13 out of the 24 genera, with just one chloroplast genome sequenced from the genus *Ventilago*. As a result, the evolutionary relationships within this genus and its closely related species remain poorly understood. Recent research has utilized chloroplast and mitochondrial genome data to explore the phylogenetic relationships between *V. leiocarpa* and other Rhamnaceae species [5], such as *Hovenia* and *Rhamnus*, contributing to a more comprehensive understanding of Rhamnaceae taxonomy. A high-quality nuclear genome of *V. leiocarpa* could provide a valuable resource for investigating the genetic mechanisms of secondary metabolite biosynthesis, adaptive evolution, and the taxonomic relationships within the Rhamnaceae family.

Here, we present the first chromosome-level reference genome of *V. leiocarpa*, assembled using PacBio HiFi long reads and Hi-C chromatin interaction data. The genome spans 227.7 Mb with a contig N50 of 18.9 Mb, and 96.7% of sequences are anchored to 12 pseudochromosomes. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment indicated 98.8% completeness, and 28,743 protein-coding

genes were predicted, with 94.71% annotated across functional databases. This high-quality genome provides a solid foundation for future functional genomics and molecular breeding efforts, aimed at enhancing the medicinal and ecological potential of *V. leiocarpa*.

Methods

Sample collection and sequencing. Fresh samples of *V. leiocarpa* were collected from Laibin City, Guangxi, China (23°46'15.845" N, 109°12'51.995" E). Healthy individuals were selected, and young leaves were immediately frozen in liquid nitrogen and stored at − 80°C until DNA extraction. High-molecular-weight genomic DNA was extracted using a modified CTAB method, and its purity and integrity were examined using agarose gel electrophoresis and a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA).

For PacBio HiFi sequencing, genomic DNA was sheared into large fragments (~ 20 kb) using a g-TUBE (Covaris), and a SMRTbell library was constructed with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) [6,7]. Sequencing was performed on the PacBio Revio platform, yielding 33.99 Gb of high-fidelity (HiFi) reads, corresponding to approximately 146× genome coverage (Table 1). The reads had an average length of 21.65 kb and an N50 of 17.39 kb, with 99.67% longer than 10 kb, ensuring sufficient long-read continuity for *de novo* assembly.

To obtain chromosome-level scaffolds, a Hi-C library was constructed following a standard proximity-ligation protocol. Fresh leaf tissues were crosslinked with formaldehyde, digested with the restriction enzyme HindIII, and ligated to capture spatially adjacent chromatin fragments. The resulting Hi-C library was sequenced on the MGI DNBSEQ platform (2 × 150 bp), generating 30.13 Gb of raw reads. After quality filtering with fastp (v0.24.0), 30.01 Gb of clean reads (Q30 = 95.87%, GC = 35.93%) were retained, representing approximately 133× coverage of the genome (Table 1).

In addition, to support gene structure annotation and transcriptomic analysis, total RNA was extracted from roots, stems, and leaves, each with three biological replicates, using the TRIzol reagent (Invitrogen). RNA quality was checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Nine RNA-seq libraries were constructed and sequenced on the Illumina NovaSeq 6000 platform (2 × 150 bp). Each library produced no less than 5 Gb of high-quality clean data (Q30 > 92%) (Table 1). These transcriptome reads were subsequently aligned to the assembled genome to assist gene prediction and functional annotation.

Table 1
Summary of sequencing data for *V. leiocarpa*.

Sequencing type	Platform	Library insert size / read length	Total data (Gb)	Coverage (×)
PacBio HiFi	PacBio Revio	~ 20 kb	33.99	146×
Hi-C	MGI DNBSEQ	2 × 150 bp	30.01	133×
RNA-seq	Illumina NovaSeq 6000	2 × 150 bp	≥ 5 per library (9 libraries)	–

Genome survey and de novo assembly. A genome survey of *V. leiocarpa* was conducted using PacBio HiFi reads to estimate basic genomic characteristics. Based on 21-mer frequency distribution analysis conducted with Jellyfish (v2.3.0) [8] and GenomeScope (v1.0)[9], the genome size was estimated at approximately 199.70 Mb, with a heterozygosity rate of 1.23% (Fig. 1A). These results indicated a relatively small but moderately heterozygous genome suitable for long-read sequencing assembly.

The genome of *V. leiocarpa* was assembled using Hifiasm (v0.24.0-r702)[10] based on PacBio HiFi long reads, yielding a primary assembly of 227.69 Mb in 29 contigs, with a contig N50 of 18.90 Mb and a maximum contig length of 24.93 Mb. For chromosome-level scaffolding, Hi-C clean reads were aligned to the draft assembly using BWA (v0.7.18) and processed through Juicer (v1.6), 3D-DNA, and manual curation in Juicebox (v1.11.08) [11]. The final assembly contained 27 scaffolds, of which 12 were anchored to pseudochromosomes (representing 222.99 Mb, 97.94% of the total genome), while 15 short scaffolds (4.70 Mb, 2.06%) remained unanchored (Table 2, Fig. 1B). The chromosome-level assembly retained a contig N50 of 18.90 Mb and exhibited a genome-wide GC content of 35.31%, consistent with the Hi-C sequencing statistics (GC = 35.93%). These results indicate a well-anchored genome with stable base composition and accurate chromosomal organization. To evaluate assembly accuracy, PacBio HiFi reads were realigned to the final chromosome-level genome using Minimap2 (v2.26) [12]. A total of 99.98% of reads were successfully mapped, with an average sequencing depth of 146.6×, confirming high sequence accuracy and uniform coverage across the genome.

Table 2
Summary statistics of the *V. leiocarpa* genome assembly.

Content	*V. leiocarpa*
Total assembly size (bp)	227,685,732
GC content (%)	35.31
Number of contigs	29
Contig N50 (bp)	18,895,906
Longest contig (bp)	24,931,219
Contig L50 / L90	6/11
Mean contig length (bp)	7,851,198
Number of scaffolds	27
Scaffold N50 (bp)	18,895,906
Longest scaffold (bp)	24,931,219
Mean scaffold length (bp)	8,432,805
Number of pseudochromosomes	12
Anchored sequence length (bp)	222,988,997 (97.94%)
Unanchored sequence length (bp)	4,696,735 (2.06%)
BUSCO completeness (eudicots_odb10)	C:98.8%[S:97.4%,D:1.4%],F:0.9%,M:0.3%

Assembly completeness was further assessed using BUSCO (v5.8.0)[13,14] with the eudicots_odb10 dataset, revealing 98.8% complete and 0.9% fragmented single-copy orthologs. Among the complete BUSCOs, 97.4% were single-copy and 1.4% were duplicated (Table 2). Together, these results demonstrate that the *V. leiocarpa* genome assembly is of high continuity and completeness, providing a robust foundation for downstream annotation and comparative analyses.

Repeat and gene annotation. Repetitive elements in the *V. leiocarpa* genome were identified using EDTA (v2.2.2) [15,16], which integrates multiple structure-based and homology-based tools to predict various transposable element (TE) types, including LTR, TIR, MITE, and Helitron elements. After filtering out low-confidence and redundant TE predictions, a comprehensive non-redundant TE library was generated and used for genome-wide annotation. A total of 93,370 interspersed repeats were identified, occupying 51.59 Mb, which accounts for 22.66% of the assembled genome (227.68 Mb). Among these, LTR retrotransposons represented the largest class, contributing 14.31% of the genome, followed by terminal inverted repeat (TIR) transposons (5.77%) and non-TIR elements such as Helitrons (1.65%). The LTR/Gypsy and LTR/Copia superfamilies were the most abundant, covering 4.51% and 2.58% of the genome, respectively. DNA transposons were dominated by Mutator, PIF/Harbinger, and CACTA families,

together accounting for approximately 4.9% of the genome (Table 3, Fig. 2A). The overall TE landscape suggests that retrotransposons are the predominant repetitive components shaping the *V. leiocarpa* genome.

Table 3
Statistics of repetitive elements annotation of *V. leiocarpa*.

Class	Superfamily	Number of elements	Length (bp)	Percentage of genome (%)
Class I (Retrotransposons)	LTR/Copia	7,276	5,875,017	2.58
	LTR/Gypsy	9,805	10,258,080	4.51
	LTR/unknown	21,415	16,435,144	7.22
	LINE/L1	1,407	922,567	0.41
Class II (DNA transposons)	TIR/CACTA	7,417	2,654,134	1.17
	TIR/Mutator	14,462	4,905,583	2.15
	TIR/PIF-Harbinger	9,333	3,537,387	1.55
	TIR/Tc1-Mariner	680	359,484	0.16
	TIR/hAT	4,693	1,673,613	0.74
Non-TIR (Helitrons)	Helitron	10,658	3,756,622	1.65
Unclassified	Repeat fragments	6,224	1,216,211	0.53
Total	—	**93,370**	**51,593,842**	**22.66**

For gene structure annotation, transcriptome data from multiple tissues (root, stem, and leaf) were aligned to the assembled genome using Hisat2 (v2.2.1) [17] and StringTie (v3.0.0)[18] to obtain high-confidence transcript assemblies. The PASA pipeline (v2.5.3)[19] was employed to refine gene models based on the assembled transcripts, and Helixer (v0.3.4)[20], a deep learning–based gene prediction tool, was used for ab initio gene prediction. Gene models derived from these approaches were integrated and filtered to produce the final consensus gene set. A total of 24,973 protein-coding genes were predicted from the *V. leiocarpa* genome, with an average gene length of 1.83 kb and an average CDS length of 1.29 kb. The average protein length was 429 amino acids, and each gene contained on average 5.54 exons and 4.54 introns, with mean exon and intron lengths of 331 bp and 348 bp, respectively. Each gene had approximately 1.15 transcripts, indicating a relatively simple gene structure (Table 4). The completeness of the annotated gene set was evaluated using BUSCO (v5.8.0) with the eudicots_odb10 database, which identified 95.8% complete BUSCOs (94.4% single-copy and 1.4% duplicated), 3.0% fragmented, and 1.2%

missing orthologs, indicating that the predicted gene models are highly complete and of high quality (Table 4).

Table 4
Summary statistics of gene structure annotation in *V. leiocarpa.*

Content	*V. leiocarpa*
Number of protein-coding genes	24,973
Number of transcripts	28,743
Mean protein length (aa)	429.37
Mean gene length (bp)	1,833.31
Mean CDS length (bp)	1,288.11
Mean number of transcripts per gene	1.15
Mean number of exons per gene	5.54
Mean exon length (bp)	331.05
Mean number of introns per gene	4.54
Mean intron length (bp)	348.09
BUSCO completeness (eudicots_odb10)	C:95.8%[S:94.4%,D:1.4%],F:3.0%,M:1.2%

Gene functional annotation was conducted by aligning the representative protein set (longest isoform per gene) to the eggNOG database using DIAMOND (v2.1.11) [21,22]. The results showed that 93.58% of genes (21,427) were assigned to COG categories, 90.90% (20,813) were annotated with PFAM domains, 49.27% (11,281) were mapped to KEGG pathways [23], and 54.22% (12,414) were annotated with GO terms (Fig. 2A, Table S1-1) [24]. Overall, 94.71% (21,685) of genes were annotated with at least one functional entry, providing valuable insights into their biological functions, molecular activities, and involvement in metabolic pathways. Further analysis of the PFAM domains revealed pentatricopeptide repeat-containing proteins and LRR receptor-like kinases as the most abundant (Fig. 2B). KEGG pathway classification highlighted significant enrichment in categories related to membrane trafficking, ribosome function, and the ubiquitin system (Fig. 2C), while GO terms were predominantly associated with cellular component organization, metabolic processes, and catalytic activity (Fig. 2D). These findings offer a comprehensive overview of the genomic composition and functional landscape of *V. leiocarpa*.

To further verify the quality and completeness of the V. leiocarpa genome annotation, we collected samples of roots (R), stems (S), and leaves (L) (Fig. 3A), performed transcriptome sequencing, and aligned the resulting transcriptomic data from these tissues to the assembled reference genome. The mapping rates of all nine RNA-seq libraries ranged from 96.1% to 96.9%, with 93.3–94.9% uniquely mapped reads, confirming the accuracy and comprehensiveness of gene annotation. Most reads were

mapped to exonic regions (~ 95%), with a small proportion aligning to intronic (~ 4%) and intergenic (~ 1%) regions, further indicating well-defined gene models (Table S1-2). Correlation analysis among biological replicates showed high intra-tissue reproducibility (Pearson's r > 0.99), confirming the reliability of the RNA-seq data (Fig. 3B). In addition, moderate correlations were observed between root and stem samples (r = 0.73–0.93), suggesting transcriptional similarity between these two tissues. This pattern is consistent with the known medicinal use of *V. leiocarpa*, in which the dried roots and rhizomes (root–stem tissues) serve as the primary medicinal parts [25], reflecting their close physiological and metabolic functions. Differential expression analysis identified 4,406 upregulated and 5,424 downregulated differentially expressed genes (DEGs) in leaves vs. roots, 3,020 upregulated and 4,390 downregulated DEGs in leaves vs. stems, and 3,752 upregulated and 3,454 downregulated DEGs in stems vs. roots (Fig. 3C). KEGG and GO enrichment analysis was performed for DEGs identified among root, stem, and leaf tissues (Table S1-3, Fig. 3D). In the leaf vs. root comparison, DEGs were significantly enriched in pathways associated with secondary metabolite biosynthesis, including phenylalanine metabolism, terpenoid backbone biosynthesis, flavone and flavonol biosynthesis, isoflavonoid biosynthesis, and plant hormone signal transduction. In the leaf vs. stem comparison, enriched pathways included flavonoid biosynthesis, phenylpropanoid biosynthesis, diterpenoid biosynthesis, and MAPK signaling pathway. In the stem vs. root comparison, DEGs were mainly enriched in phenylpropanoid biosynthesis, flavonoid biosynthesis, cytochrome P450, and ABC transporter pathways (Fig. 3D). These results indicate that lots of DEGs identified across different tissues are involved in the biosynthesis and regulation of secondary metabolites, providing a molecular basis for understanding tissue-specific metabolic differentiation. Together, these transcriptomic results further support the high completeness and reliability of the *V. leiocarpa* genome annotation and provide important insights into the molecular mechanisms underlying the formation of medicinally active compounds in its roots and rhizomes.

Comparative genomics and phylogenetic analysis. Homologous gene clustering and comparative genomic analyses were performed using *V. leiocarpa* and 11 publicly available representative plant genomes, including six species within the Rhamnaceae family (Table S1-4). Orthologous gene families were identified using OrthoFinder (v2.5.4) [26] with protein sequences from the longest isoforms of each gene as input. All-vs-all protein sequence comparisons were conducted using DIAMOND with an *E*-value cutoff of $1e^{-5}$, followed by Markov Cluster Algorithm (MCL) clustering to define orthogroups. In total, 424,496 protein-coding genes from 12 species were analyzed, and 382,291 genes (92.7%) were assigned to 35,178 orthogroups, while 38,596 genes (7.3%) were unclustered. Among these, 10,892 orthogroups (11.8%) were species-specific, involving 50,284 genes. Furthermore, 271 single-copy orthogroups were identified across all species, whereas the six Rhamnaceae species shared 12,197 single-copy orthogroups (Fig. 4A).

A phylogenetic tree was constructed using single-copy orthologous genes aligned by MAFFT and inferred with FastTree under the JTT + CAT model[27,28]. Divergence times were estimated using r8s, with the calibration point between *Ziziphus jujuba* and *Arabidopsis thaliana* set at 121 million years ago (Mya), based on the TimeTree database (http://www.timetree.org, accessed on March 20, 2025) [29]. The

resulting phylogeny placed *V. leiocarpa* in a well-supported clade with *Sageretia thea* and *Rhamnella rubrinervis*, diverging approximately 29.07 Mya (Fig. 4B). Gene family expansion and contraction were analyzed using CAFE (v5.0) [30], which identified substantial gene family contractions in the ancestral Rhamnaceae lineage. Notably, *Z. jujuba* exhibited extensive gene family expansion, whereas *V. leiocarpa* and its close relatives showed moderate contraction events, reflecting lineage-specific genome evolution (Fig. 4B). Collinearity relationships among *V. leiocarpa*, *S. thea*, and *Z. jujuba* were examined using MCScanX [31]. The results revealed extensive collinearity among the 12 pseudochromosomes, with strong one-to-one correspondence across chromosomes 1–11. However, partial inversions were observed at the ends of chromosome 12 in all three genomes, suggesting structural rearrangements that may have occurred during the divergence of their common ancestor (Fig. 4C).

Data Records

The *V. leiocarpa* genome sequencing, assembly, and annotation data have been deposited in the China National GeneBank DataBase (CNGBdb) under accession number CNP0008168. The archived data include raw PacBio HiFi reads, Hi-C and RNA-seq data, the final chromosome-level assembly, and the corresponding gene structure (GFF3).

Technical Validation

The integrity and accuracy of the *V. leiocarpa* genome assembly were validated through multiple complementary assessments. The final assembly comprised 27 scaffolds, of which 12 were anchored to pseudochromosomes (222.99 Mb, 97.94% of the genome) and 15 unanchored scaffolds (4.70 Mb, 2.06%) remained (Table 2, Fig. 1B). The genome exhibited a contig N50 of 18.90 Mb and a GC content of 35.31%, consistent with Hi-C data (35.93%). PacBio HiFi reads were mapped back to the assembly using Minimap2 (v2.26), yielding a 99.98% mapping rate with an average sequencing depth of 146.6×, confirming the high accuracy and completeness of the assembly. BUSCO assessment (v5.8.0, eudicots_odb10) further demonstrated 98.8% complete (97.4% single-copy, 1.4% duplicated), 0.9% fragmented, and 0.3% missing orthologs (Table 2), supporting the high continuity and completeness of the genome. To validate genome annotation, gene models were manually inspected and functionally annotated against the PFAM, GO, and KEGG databases. Transcriptomic reads from root, stem, and leaf tissues were aligned to the genome, achieving > 96.1% overall mapping rates with > 93.3% uniquely mapped reads (Table S1-2), confirming the accuracy and completeness of gene prediction. The annotated gene set exhibited 95.8% complete BUSCOs (94.4% single-copy, 1.4% duplicated) (Table 4), indicating that the predicted genes are highly complete and suitable for downstream functional and comparative analyses.

Declarations

Code availability

All analyses were performed with standard bioinformatics tools according to their manuals, including Hifiasm (v0.24.0), Juicer, 3D-DNA, Minimap2 (v2.26), BUSCO (v5.8.0), OrthoFinder (v2.5.4), and MCScanX (v1.1) et al. No custom scripts were used.

Data availability

The genome sequencing, assembly, and annotation data of *V. leiocarpa* are readily available for public access. These data have been deposited in the China National GeneBank DataBase (CNGBdb)[32] with the accession number CNP0008168. The archived dataset encompasses raw PacBio HiFi reads, Hi-C and RNA-seq data, the final chromosome-level assembly, as well as the corresponding gene structure information in GFF3 format. Researchers can access and utilize these data for further studies in relevant fields, following the terms and conditions set by CNGBdb.

Funding

This research received financial support from multiple sources. The Guangxi Special Project for Science and Technology Bases and Talents provided funding under the grant number GuiKe AD22080058. Additionally, the Scientific Research and Technology Development Plan of Guangxi Laibin, China, contributed support through the project Laibin Sci-Tech Tackling 240102. Furthermore, the Fujian Provincial Natural Science Foundation offered financial assistance with the grant number 2025J01600.

Author contributions

S.G. and L.Y. conceived and designed the study. S.G. collected the samples and obtained project funding. S.C., Pi.Z. and R.L. performed genome assembly, data analysis, and drafted the manuscript. Y.W. analyzed the transcriptome data. Pe.Z. assisted with sample collection and phenotypic observation. Q.Y. and Y.Q. performed genomic data analysis. R.L., S.G. and Pi.Z. revised the manuscript and contributed to experimental design. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.G., L.Y. or P.Zh.

References

1. Jo, J. *et al.* The first Chromosomal-level genome assembly of *Sageretia thea* using Nanopore long reads and Pore-C technology. *Sci Data* **11**, 959, doi:10.1038/s41597-024-03798-9 (2024).
2. Griesmann, M. *et al.* Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science* **361**, doi:https://doi.org/10.1126/science.aat1743 (2018).

3. Qing, Q. & Zhijian, F. Description of the original plant of *Ventilago leiocarpa* drupe and morphological and histological study of its medicinal properties. *Journal of Guangdong University of Medicine*, 55-57, doi:https://link.cnki.net/doi/10.16720/j.cnki.tcyj.2024.170 (1989).
4. Xueqing, H. *et al.* Research progress on chemical composition and pharmacological activity of *Ventilago leiocarpa* plants. *Specialty Research* **46**, 149-156, doi:https://link.cnki.net/doi/10.16720/j.cnki.tcyj.2024.170 (2024).
5. Wanichthanarak, K. *et al.* Revisiting chloroplast genomic landscape and annotation towards comparative chloroplast genomes of Rhamnaceae. *BMC Plant Biol* **23**, 59, doi:https://doi.org/10.1186/s12870-023-04074-5 (2023).
6. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:https://doi.org/10.1126/science.1162986 (2009).
7. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461-465, doi:https://doi.org/10.1038/nmeth.1459 (2010).
8. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770, doi:https://doi.org/10.1093/bioinformatics/btr011 (2011).
9. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432, doi:https://doi.org/10.1038/s41467-020-14998-3 (2020).
10. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175, doi:https://doi.org/10.1038/s41592-020-01056-5 (2021).
11. Robinson, J. T. *et al.* Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256-258.e251, doi:https://doi.org/10.1016/j.cels.2018.01.001 (2018).
12. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:https://doi.org/10.1093/bioinformatics/bty191 (2018).
13. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245, doi:https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).
14. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212, doi:https://doi.org/10.1093/bioinformatics/btv351 (2015).
15. Ou, S. *et al.* Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome research* **34**, 1140-1153, doi:https://www.genome.org/cgi/doi/10.1101/gr.278131.123 (2024).
16. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology* **20**, 275, doi:https://doi.org/10.1186/s13059-019-1905-y (2019).

17. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:https://doi.org/10.1038/nmeth.3317 (2015).
18. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295, doi:https://doi.org/10.1038/nbt.3122 (2015).
19. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-5666, doi:https://doi.org/10.1093/nar/gkg770 (2003).
20. Stiehler, F. *et al.* Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* **36**, 5291-5298, doi:https://doi.org/10.1093/bioinformatics/btaa1044 (2021).
21. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:https://doi.org/10.1038/nmeth.3176 (2015).
22. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825-5829, doi:https://doi.org/10.1093/molbev/msab293 (2021).
23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:https://doi.org/10.1093/nar/28.1.27 (2000).
24. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:https://doi.org/10.1038/75556 (2000).
25. Qian Qing, F. Z. Description of the original plant of *Ventilago leiocarpa* and morphological and histological study of its medicinal properties. *Journal of Guangdong Medical College*, doi:https://link.cnki.net/doi/10.16809/j.cnki.1006-8783.1989.01.018 (1989).
26. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:https://doi.org/10.1186/s13059-019-1832-y (2019).
27. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:https://doi.org/10.1093/molbev/mst010 (2013).
28. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
29. Kumar, S. *et al.* TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* **39**, doi:https://doi.org/10.1093/molbev/msac174 (2022).
30. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516-5518, doi:https://doi.org/10.1093/bioinformatics/btaa1022 (2021).
31. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, e49, doi:https://doi.org/10.1093/nar/gkr1293 (2012).
32. Chen, F. Z. *et al.* CNGBdb: China National GeneBank DataBase. *Yi Chuan* **42**, 799-809, doi:https://doi.org/10.16288/j.yczz.20-080 (2020).

Figures

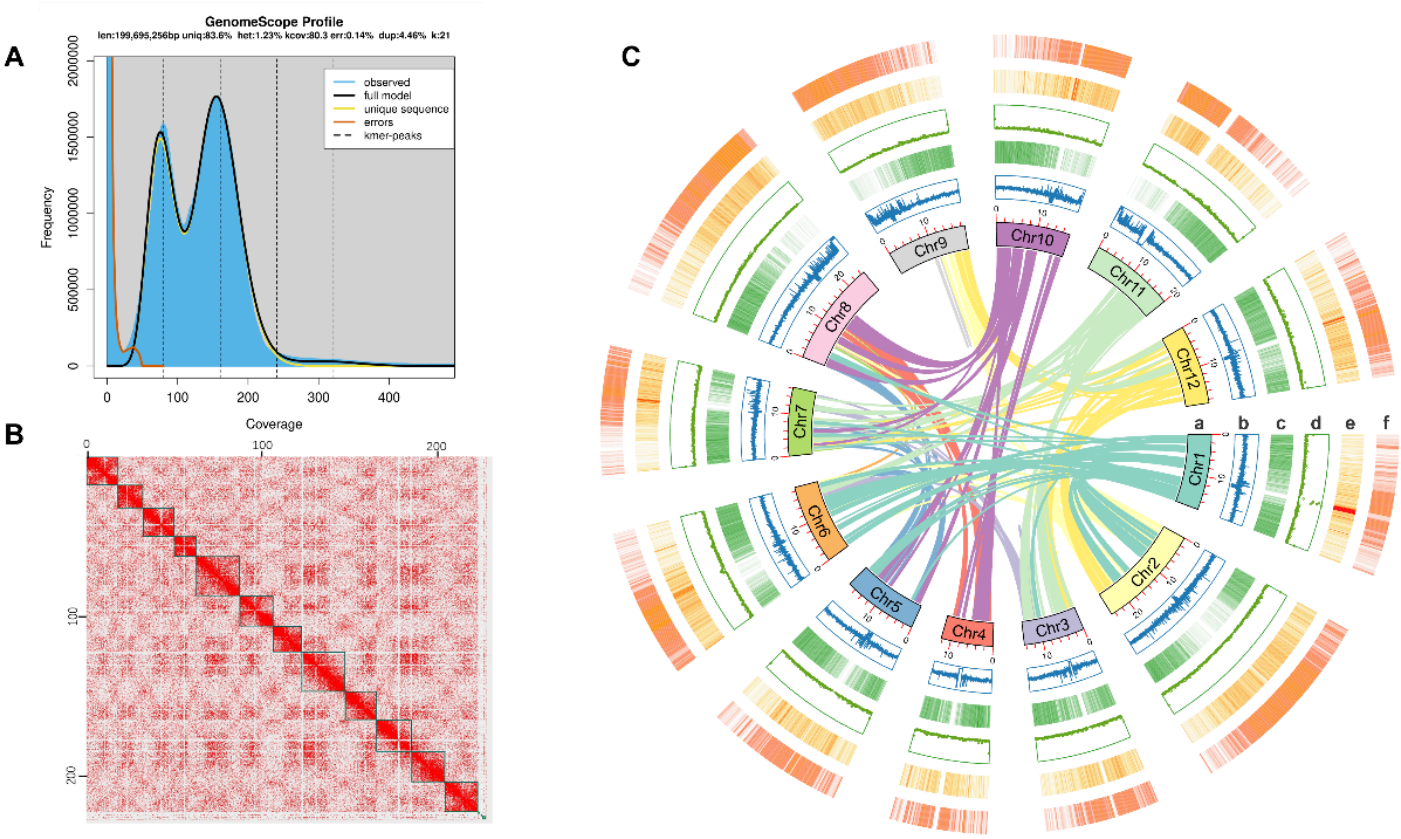

Figure 1

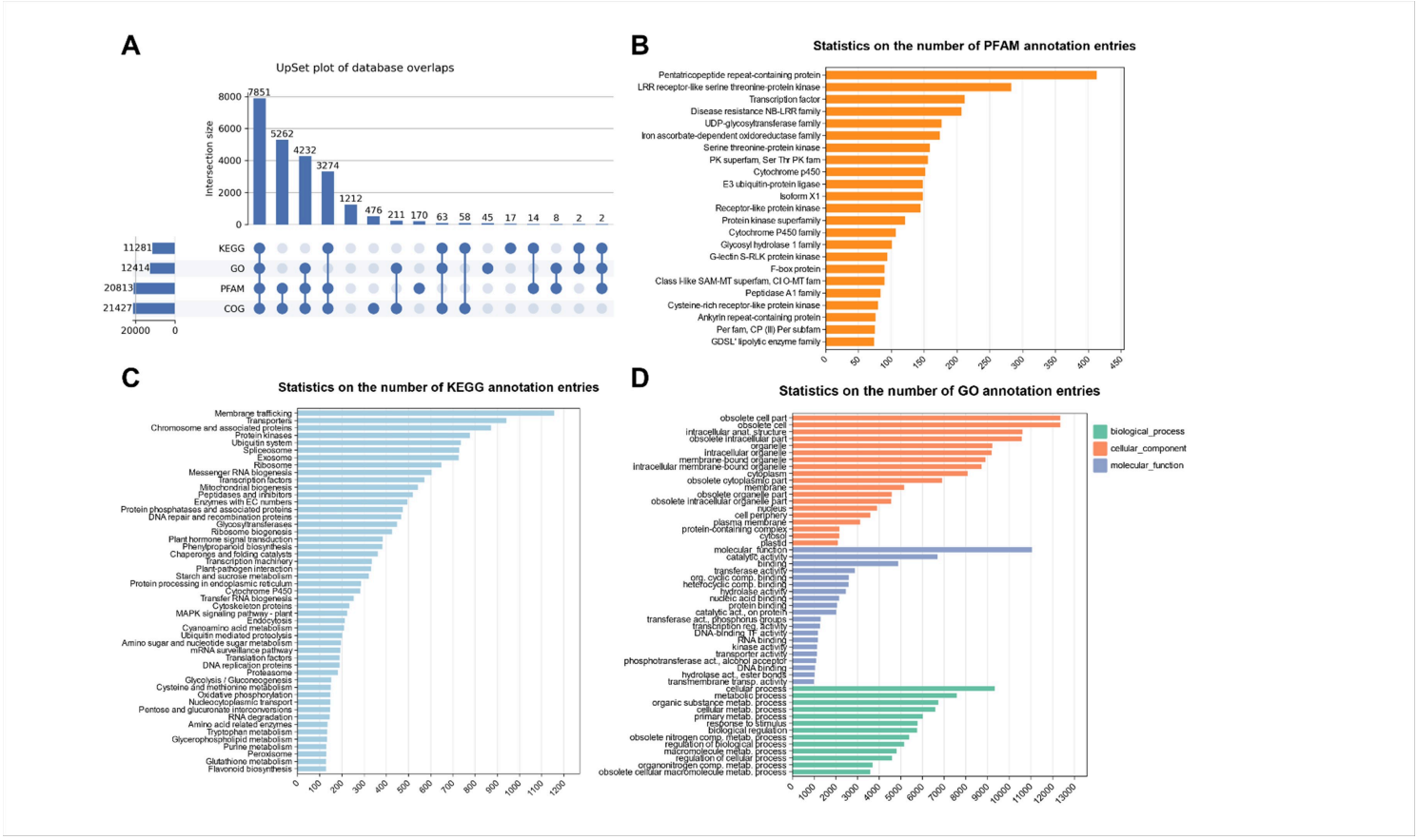
A
UpSet plot of database overlaps
Intersection size
7851
5262
4232
3274
1212
476
211
170
63
58
45
17
14
8
2
2
11281
12414
20813
21427
KEGG
GO
PFAM
COG
20000
0
B
Statistics on the number of PFAM annotation entries
Pentatricopeptide repeat-containing protein
LRR receptor-like serine threonine-protein kinase
Transcription factor
Disease resistance NB-LRR family
UDP-glycosyltransferase family
Iron ascorbate-dependent oxidoreductase family
Serine threonine-protein kinase
PK superfam, Ser Thr PK fam
Cytochrome p450
E3 ubiquitin-protein ligase
Isoform X1
Receptor-like protein kinase
Protein kinase superfamily
Cytochrome P450 family
Glycosyl hydrolase 1 family
G-lectin S-RLK protein kinase
F-box protein
Class I-like SAM-MT superfam, CIO-MT fam
Peptidase A1 family
Cysteine-rich receptor-like protein kinase
Ankyrin repeat-containing protein
Per fam, CP (III) Per subfam
GDSL' lipolytic enzyme family
C
Statistics on the number of KEGG annotation entries
Membrane trafficking
Transporters
Chromosome and associated proteins
Protein kinases
Ubiquitin system
Spliceosome
Exosome
Ribosome
Messenger RNA biogenesis
Transcription factors
Mitochondrial biogenesis
Peptidases and inhibitors
Enzymes with EC numbers
Protein phosphatases and associated proteins
DNA repair and recombination proteins
Glycosyltransferases
Ribosome biogenesis
Plant hormone signal transduction
Phenylpropanoid biosynthesis
Chaperones and folding catalysts
Transcription machinery
Plant-pathogen interaction
Starch and sucrose metabolism
Protein processing in endoplasmic reticulum
Cytochrome P450
Transfer RNA biogenesis
Cytoskeleton proteins
MAPK signaling pathway - plant
Endocytosis
Cyanoamino acid metabolism
Ubiquitin mediated proteolysis
Amino sugar and nucleotide sugar metabolism
mRNA surveillance pathway
Translation factors
DNA replication proteins
Proteasome
Glycolysis / Gluconeogenesis
Cysteine and methionine metabolism
Oxidative phosphorylation
Nucleocytoplasmic transport
Pentose and glucuronate interconversions
RNA degradation
Amino acid related enzymes
Tryptophan metabolism
Glycerophospholipid metabolism
Purine metabolism
Peroxisome
Glutathione metabolism
Flavonoid biosynthesis
0
100
200
300
400
500
600
700
800
900
1000
1100
1250
D
Statistics on the number of GO annotation entries
obsolete cell part
obsolete cell
intracellular anat. structure
obsolete intracellular part
organelle
Intracellular organelle
membrane-bound organelle
intracellular membrane-bound organelle
cytoplasm
obsolete cytoplasmic part
membrane
obsolete organelle part
obsolete intracellular organelle part
nucleus
cell periphery
plasma membrane
protein-containing complex
cytosol
plastid
molecular_function
catalytic activity
binding
transferase activity
org. cyclic comp. binding
heterocyclic comp. binding
hydrolase activity
nucleic acid binding
protein binding
catalytic act., on protein
transferase act., phosphorus groups
transcription reg. activity
DNA-binding TF activity
RNA binding
kinase activity
transporter activity
phosphotransferase act., alcohol acceptor
DNA binding
hydrolase act., ester bonds
transmembrane transp. activity
cellular process
metabolic process
organic substance metab. process
cellular metab. process
primary metab. process
response to stimulus
biological regulation
obsolete nitrogen comp. metab. process
regulation of biological process
macromolecule metab. process
regulation of cellular process
organonitrogen comp. metab. process
obsolete cellular macromolecule metab. process
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
biological_process
cellular_component
molecular_function

Figure 2

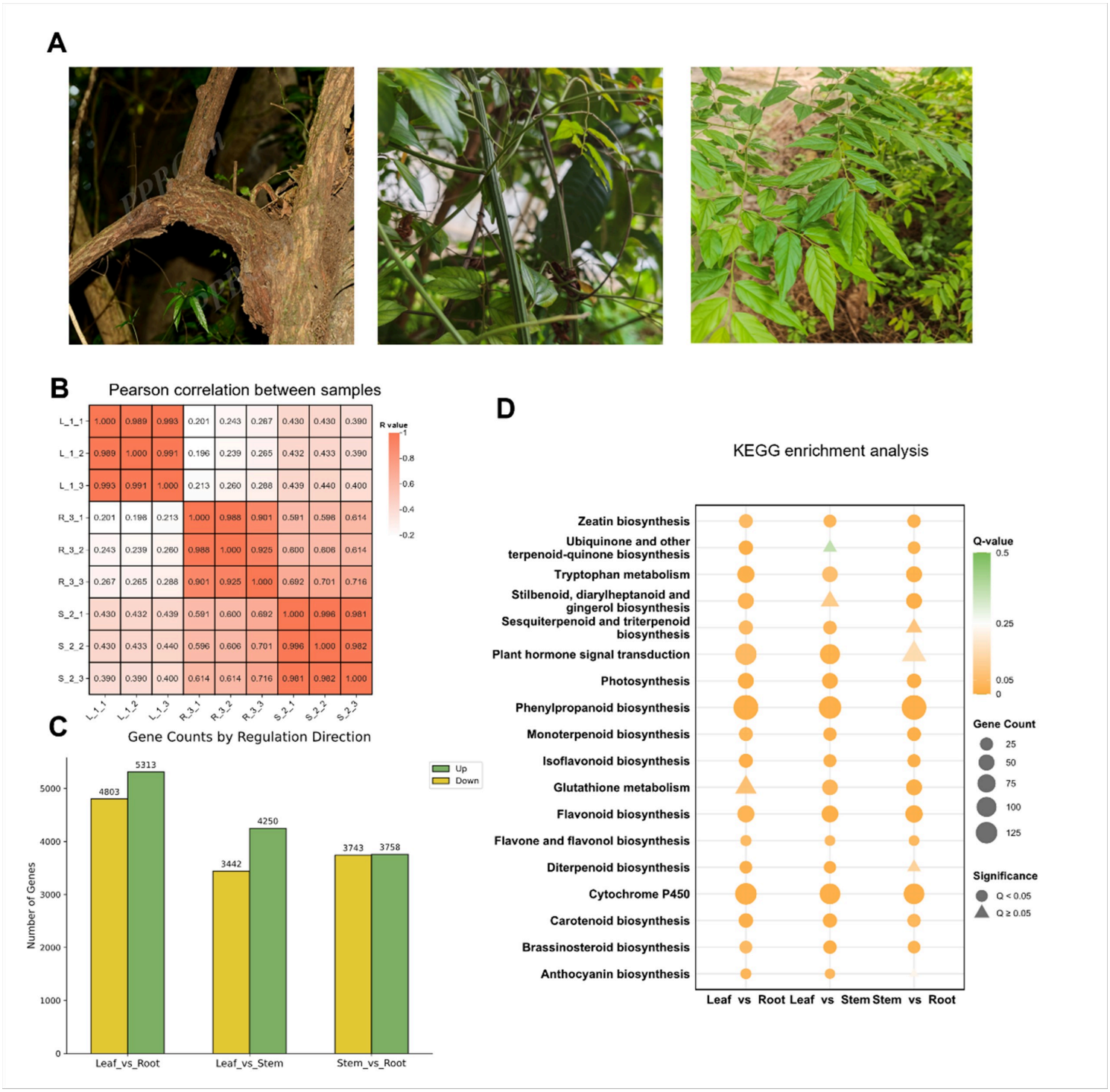
A
B
Pearson correlation between samples
R value
C
Gene Counts by Regulation Direction
Up
Down
Number of Genes
Leaf_vs_Root
Leaf_vs_Stem
Stem_vs_Root
D
KEGG enrichment analysis
Zeatin biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Tryptophan metabolism
Stilbenoid, diarylheptanoid and gingerol biosynthesis
Sesquiterpenoid and triterpenoid biosynthesis
Plant hormone signal transduction
Photosynthesis
Phenylpropanoid biosynthesis
Monoterpenoid biosynthesis
Isoflavonoid biosynthesis
Glutathione metabolism
Flavonoid biosynthesis
Flavone and flavonol biosynthesis
Diterpenoid biosynthesis
Cytochrome P450
Carotenoid biosynthesis
Brassinosteroid biosynthesis
Anthocyanin biosynthesis
Leaf vs Root
Leaf vs Stem
Stem vs Root
Q-value
Gene Count
Significance
Q < 0.05
Q ≥ 0.05

Figure 3

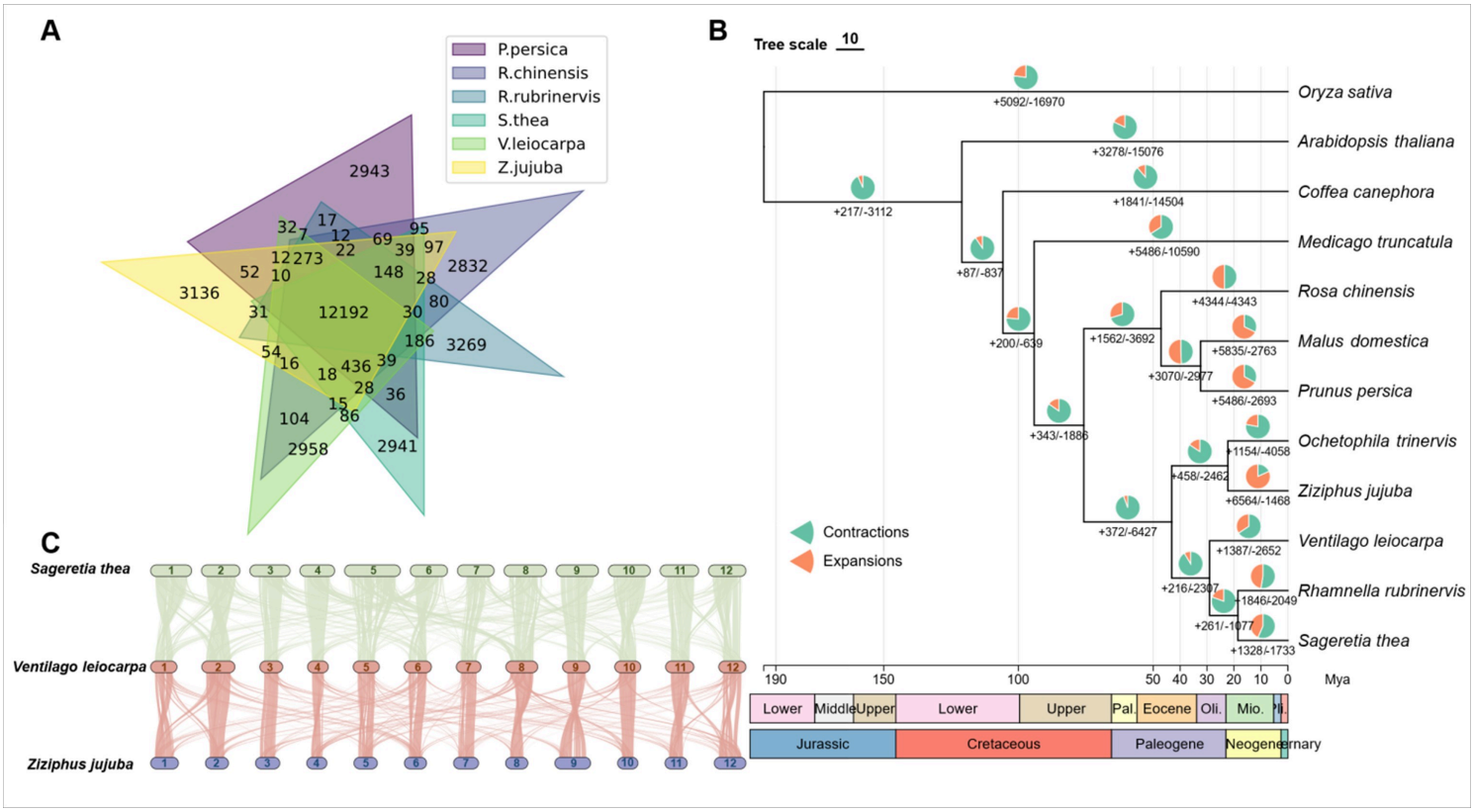

Figure 4

Supplementary Files

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

- TableS1.xlsx