scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Elaeocarpus petiolatus* (Elaeocarpaceae)

Huipeng Xi[1], Bingyang Liu[2], Mingcheng Wang[3] ✉ & Lei Wang[4] ✉

Elaeocarpus petiolatus **is an ecologically and economically important species in tropical and subtropical forests. Despite its significance, the lack of genomic resources has hindered research on the genetic diversity and adaptive traits of *E. petiolatus*. To address this gap, we present a comprehensive chromosome-level genome assembly of *E. petiolatus* generated using advanced PacBio high-fidelity (HiFi) long-read sequencing and Hi-C technology. The assembly spans 322.45 Mb, with a scaffold N50 of 20.58 Mb, indicating that 37.11% of the genome is composed of repetitive elements. We identified 25,295 protein-coding genes, of which 96.74% were functionally annotated. This high-quality genome provides a critical resource for understanding the genetic mechanisms underlying environmental adaptability and biosynthesis of bioactive compounds in *E. petiolatus*, thereby supporting conservation efforts and sustainable forest management. The assembled genome and associated sequencing data are publicly available, facilitating further evolutionary and functional studies on the Elaeocarpaceae family.**

Background & Summary

The family Elaeocarpaceae includes a diverse group of flowering plants, comprising >600 recognized species that are primarily found in the tropical and subtropical regions[1,2]. This family comprises several ecologically and economically significant genera, including *Aceratium*, *Dubouzetia*, *Elaeocarpus*, *Tetratheca*, and *Sloanea*. Members of the Elaeocarpaceae family are commonly found in forest ecosystems, which contribute to biodiversity and provide valuable hardwood, ornamental, and medicinal resources, thereby rendering them ecologically and economically significant[3–8].

Elaeocarpus petiolatus is an important species within the Elaeocarpaceae family, recognized for its durable hardwood widely utilized in artisanal crafts and timber production. Traditionally, extracts from its leaves and bark have been employed for their medicinal properties, exhibiting notable anti-inflammatory, antioxidant, and cytotoxic activities, relevant in the treatment of various diseases including inflammation and cancer[9–11]. Widely distributed across southern China, northeast India, and Southeast Asia, *E. petiolatus* occupies evergreen broad-leaved forests at elevations between 400 and 1,500 m. Despite its current conservation status of Least Concern according to the IUCN Red List, continued habitat monitoring is essential due to increasing anthropogenic pressures[3]. The species' ecological versatility and rich bioactive compounds significantly enhance its value for ecological, medicinal, and economic purposes[3,9–11].

However, despite its significance, research on *E. petiolatus* remains limited. To date, only the plastome of *E. petiolatus* is publicly available[12]. Most genomic studies within the Elaeocarpaceae family have focused on a few model species[2,13–15], resulting in a significant gap in our understanding of *E. petiolatus*. The absence of high-quality genome assemblies has impeded investigations into the genetic basis of its adaptive traits, ecological interactions, and evolutionary relationships with other species within this family.

This study aimed to fill the existing research gap by presenting a comprehensive chromosome-level genome assembly of *E. petiolatus*. Using advanced sequencing technologies, specifically PacBio high-fidelity (HiFi) long-read sequencing combined with Hi-C data, we generated a high-quality genome that significantly

[1]Center for Gardening and Horticulture, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Xishuangbanna, 666303, Yunnan, China. [2]Xishuangbanna Customs Comprehensive Technology Center, No. 65 Mengle Road, Xishuangbanna, 666100, Yunnan, China. [3]Institute for Advanced Study, Chengdu University, No. 2025 Chengluo Road, Chengdu, 610106, China. [4]Kunming Novo Medical Laboratory Co., Ltd., Kunming, 650032, China. ✉e-mail: wangmingcheng@cdu.edu.cn; 654699191@qq.com

Data type	Tissue	Number of reads	Total data (Gb)	Sequence coverage (×)
PacBio HiFi	Leaf	651,650	11.23	37.86
Hi-C	Leaf	282,986,532	42.45	143.09
RNA-seq (Total)	—	117,921,014	17.59	—
RNA-seq	Leaf	38,977,610	5.82	—
RNA-seq	Stem	39,414,576	5.88	—
RNA-seq	Fruit	39,528,828	5.89	—

Table 1. Statistics of sequencing data of *E. petiolatus*.

improves existing genomic resources for this species. This genome represents the first chromosome-level assembly reported within Elaeocarpaceae. To date, the only publicly available genome sequences in this family are two low-contiguity scaffold-level assemblies of *Aristotelia chilensis*[16,17], which are highly fragmented (contig N50 ≤ 2.6 kb). In contrast, the *E. petiolatus* genome exhibits greatly improved continuity (contig N50: 19.59 Mb), and shows excellent completeness based on BUSCO analysis (99.07%). However, due to the absence of other high-quality genomes in Elaeocarpaceae, family-wide comparative analysis (e.g., synteny or gene family evolution) are currently not feasible. Nonetheless, this assembly provides a critical genomic reference for future studies on the genetic basis of environmental adaptability and bioactive compound biosynthesis. It also lays a foundation for advancing ecological and evolutionary research on *E. petiolatus* and related Elaeocarpaceae species.

Methods

Sample collection and sequencing. For genome sequencing, young leaves were collected from a 4-year-old *E. petiolatus* plant located in Menglun, Xishuangbanna, China. High-molecular-weight DNA was extracted using a modified CTAB method[18] and SMRTbell libraries were prepared following the PacBio 15 kb protocol. Sequencing was performed using circular consensus sequencing (CCS) on a PacBio Revio platform, generating HiFi long reads for genome survey analysis and long-read assembly. Hi-C libraries were prepared from more than 2 g of young leaves using chromatin extraction, digestion, ligation, and DNA purification[19]. Libraries were sequenced on an Illumina NovaSeq. 6000 platform to generate paired-end reads for scaffolding. For transcriptomic analysis, total RNA was isolated from the fruit, stem, and leaves of the same plant, and RNA sequencing (RNA-seq) libraries were prepared using the TruSeq Stranded mRNA-Seq protocol. Sequencing was performed on the Illumina NovaSeq6000 platform. RNA-seq reads were used for genome quality assessment and gene annotation based on the transcriptome. In total, the sequencing effort produced 11.23 Gb of HiFi reads, 42.45 Gb of Hi-C reads, and 17.59 Gb of RNA-seq reads (Table 1).

Genome assembly. Before assembling the *E. petiolatus* genome, genome survey analysis was performed using a 19-mer frequency distribution of HiFi reads generated using Jellyfish v1.1.12[20]. This analysis identified 10,416,542,470 *k*-mers, with a primary peak observed at a *k*-depth of 32 (Fig. 1a). From the *k*-mer distribution, the genome size was estimated to be 325.52 Mb, with a repeat content of 47.84% and a heterozygosity rate of 1.68%. HiFi reads were processed using the CCS workflow in SMRT Link v8.0 (PacBio) and assembled into contigs with hifiasm version 0.19.9[21], using the default settings. Potential redundant haplotypes were removed using Purge Haplotigs v1.1.1[22]. After mapping the HiFi reads to the polished contigs using Minimap2 v2.28[23], contigs with low coverage (<5) or abnormal GC content (>45%) were excluded. This resulted in 33 contigs, totaling 322.44 Mb, with an N50 of 19.59 Mb. Hi-C reads were subsequently mapped to these contigs using the Juicer v1.8.8[24]. The uniquely mapped reads were then used to anchor the contigs onto the pseudochromosomes using 3D-DNA software[25]. Manual curation of the Hi-C contact maps resulted in a final chromosome-level assembly measuring 322.45 Mb (Fig. 2). The N50 for the contigs and scaffolds were 19.59 Mb and 20.58 Mb, respectively (Table 2). In total, 98.73% (318.37 Mb) of the sequences were anchored to 15 pseudochromosomes, with sizes ranging from 16.38 Mb to 29.33 Mb (Fig. 1b; Table 3). Of the 15 pseudochromosomes, seven were gapless, five contained one gap, and three had three gaps (Table 3).

Genome annotation. Repetitive elements were annotated using a combination of *de novo* prediction and homology-based searches with RepeatMasker v4.0.9[26] and RepeatModeler v2.0.5[27], following the methodology outlined by Wang *et al.*[28]. This analysis identified 119.67 Mb of repetitive sequences, accounting for 37.11% of the total genome (Fig. 2; Table 4). A significant portion of these repeats (70.25 Mb, or 21.76% of the genome) could not be assigned to known transposable element families and were therefore designated as "unknown." These sequences may represent lineage-specific or uncharacterized repeats in *E. petiolatus* or the *Elaeocarpus* genus. This unclassified fraction was significantly larger than the portion of identified transposable elements, which accounted for 49.48 Mb (15.35%) of the genome. Among these, long-terminal repeat (LTR) retrotransposons were the most prevalent, comprising 8.27% of the genome (Table 4).

After masking all repetitive elements in the *E. petiolatus* genome, we used three complementary approaches to predict the protein-coding genes. First, we aligned protein sequences from five Rosanae species—*Arabidopsis thaliana* (TAIR10)[29], *Malus domestica* (v1.1)[30], *Vitis vinifera* (Genoscope.12X)[31], *Citrus sinensis* (v1.1)[32], and *Cucumis sativus* (v1.0)[33]—to the *E. petiolatus* genome using BLASTN v2.2.31+[34] to identify homologous regions, followed by gene structure prediction using GeneWise v2.4.1[35]. Second, RNA-seq data from the three tissues were assembled with Trinity 2.8.4[36]. The assembled transcripts were aligned to the genome, and gene-coding sequences were predicted using PASA v2.3.3[37]. Third, *de novo* predictions were made using

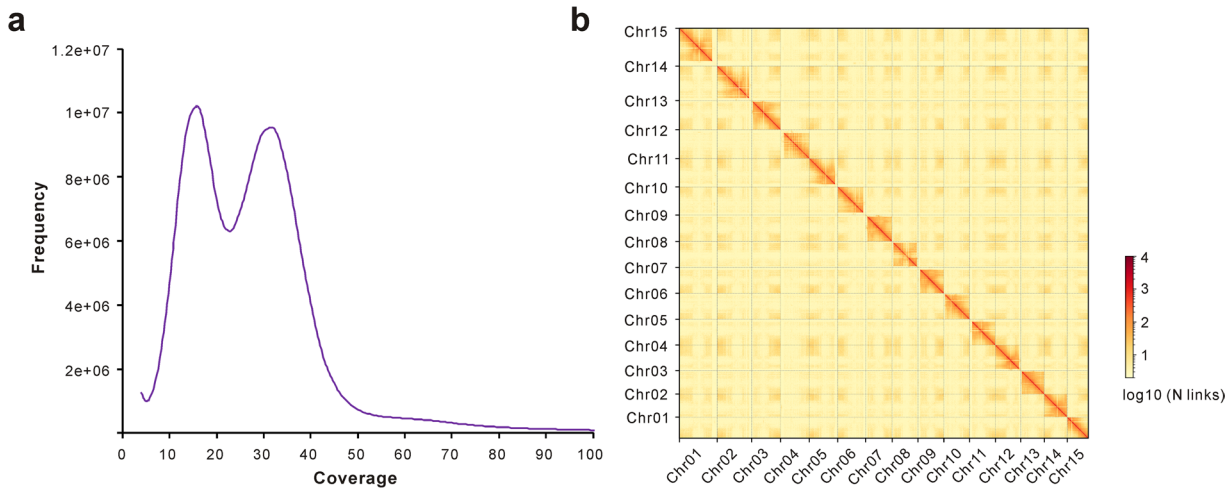

Fig. 1 Chromosome-level genome assembly of *E. petiolatus*. (**a**) *K*-mer frequency distribution derived from PacBio HiFi reads. (**b**) Hi-C interaction heatmap for 15 pseudochromosomes of the *E. petiolatus* genome.

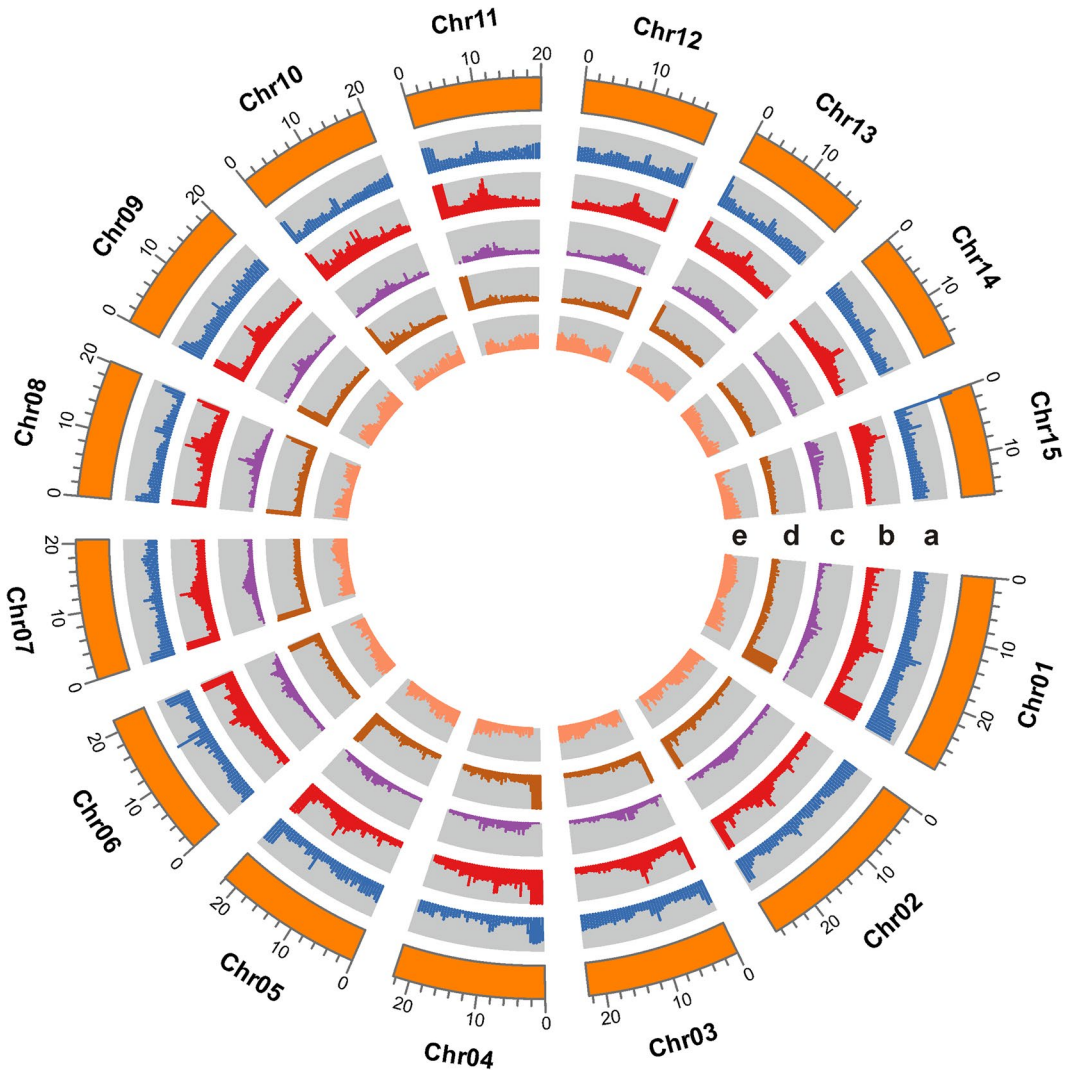

Fig. 2 Genomic features across non-overlapping 500 kb windows of the *E. petiolatus* genome. The tracks from the outer to the inner rings represent (**a**) GC content, (**b**) repeat density, (**c**) transposable element density, (**d**) density of unknown repeats, and (**e**) gene density.

Genome feature	Value
Assembly	
Total size (Mb)	322.45
Number of contigs	35
Longest contig (Mb)	26.35
Contig N50 (Mb)	19.59
Number of scaffolds	21
Longest scaffold (Mb)	29.33
Scaffold N50 (Mb)	20.58
Number of pseudochromosomes	15
Total size of pseudochromosomes (Mb)	318.37
BUSCO completeness (%)	99.07
LTR assembly index	18.56
RNA-seq reads mapping rate (%)	96.11
GC content (%)	33.38
Annotation	
Repeat content (%)	37.11
Number of protein-coding genes	25,295
Average transcript length (bp)	3,424.28
Average coding sequence length (bp)	1,211.30
Average exon length (bp)	214.65
Average exon number per gene	5.64
Average intron length (bp)	476.63
BUSCO completeness (%)	92.56

Table 2. Statistics of the assembly and annotation of the *E. petiolatus* genome.

Pseudochromosome	Length (bp)	Number of contigs	Number of genes
Chr01	29,326,508	4	2,302
Chr02	26,823,607	2	2,247
Chr03	22,546,077	4	2,032
Chr04	22,420,272	2	1,289
Chr05	22,096,012	1	1,750
Chr06	21,768,957	2	1,613
Chr07	20,575,603	2	1,613
Chr08	20,180,515	4	1,373
Chr09	20,134,892	1	1,796
Chr10	20,086,218	1	1,405
Chr11	20,053,831	1	1,459
Chr12	19,747,158	1	1,725
Chr13	18,365,513	1	1,419
Chr14	17,864,998	2	1,582
Chr15	16,380,615	1	1,510
Total	**318,370,776**	**29**	**25,115**

Table 3. Summary of 15 pseudochromosomes of the *E. petiolatus* genome.

AUGUSTUS v3.2.3[38], with parameters trained on gene models derived from PASA results. This analysis focused on genes containing at least three exons and coding sequences longer than 1500 bp. These gene models were integrated into a final consensus set using EVidenceModeler v1.1.1[39].

We identified 25,295 protein-coding genes, with average lengths of transcripts, coding sequences, exons, and introns of 3,424.28 bp, 1,211.30 bp, 214.65 bp, and 476.63 bp, respectively (Table 2). Of these, 25,115 genes were anchored to the 15 pseudochromosomes, while the remaining 180 genes were located on unanchored scaffolds (Table 3). In addition, 1,431 genes (5.66%) were identified as transcription factors by comparison with the Plant Transcription Factor Database v5.0[40] (Fig. 3). Functional annotation was performed by aligning the protein sequences of these genes with several publicly available databases, including SwissProt, TrEMBL[41], InterPro[42], Gene Ontology[43], and the Kyoto Encyclopedia of Genes and Genomes[44]. Consequently, 24,470 protein-coding genes, representing 96.74% of all the genes, were functionally annotated in at least one database (Table 5).

Class	Count	Total length (bp)	% of genome
DNA	46,338	13,356,241	4.14
LINE	7,578	2,960,306	0.92
SINE	281	32,326	0.01
LTR	38,335	26,663,526	8.27
Satellite	2,601	6,468,208	2.01
Simple repeat	867	159,400	0.05
Low complexity	12	1,877	0.00
Unknown	178,017	70,253,836	21.79
Total	**274,029**	**119,670,534**	**37.11**

Table 4. Summary of repetitive elements within the *E. petiolatus* genome. "Unknown" refers to repetitive sequences that could not be classified into known transposable element families.

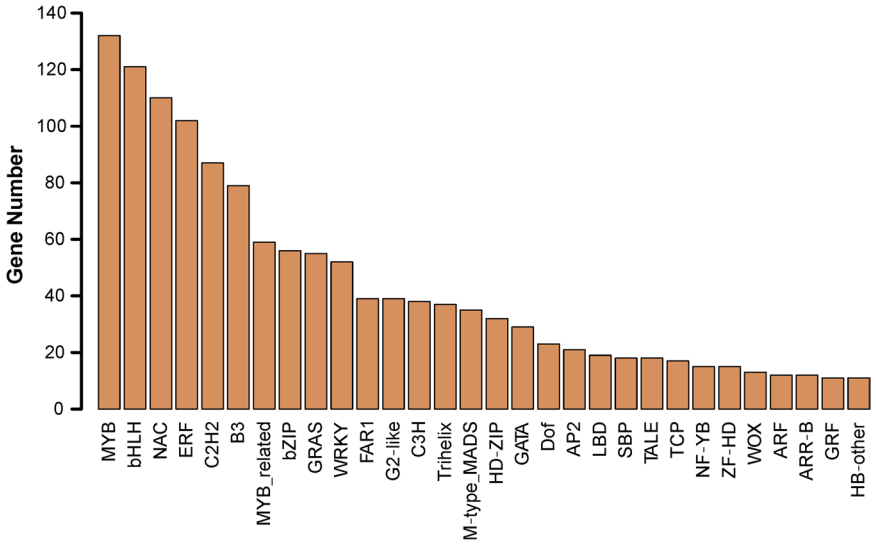

Fig. 3 Distribution of the top 30 transcription factor families identified in the *E. petiolatus* genome.

	Number of genes	Percentage (%)
Total	25,295	—
Annotated	24,470	96.74
InterPro	23,211	91.76
KEGG	8,567	33.87
SwissProt	17,560	69.42
TrEMBL	23,553	93.11
GO	21,987	86.92
Unannotated	825	3.26

Table 5. Functional annotation of protein-coding genes in the *E. petiolatus* genome.

Data Records

The genome assembly of *E. petiolatus* and the associated sequence data have been made publicly available through the NCBI database under BioProject PRJNA1173526[45]. The genome assembly is available in GenBank under accession number JBINKO000000000[46]. Raw sequencing data, including PacBio HiFi and Hi-C reads, are available in the Sequence Read Archive (SRA) under accession numbers SRR31035705[47] and SRR31024753[48], respectively. RNA-seq reads were deposited under the SRA accession numbers SRR31013082–SRR31013084[49–51]. Additionally, the genome assembly, along with annotations of repetitive elements, gene structures, and functional features, has been archived in Figshare[52].

Technical Validation

The quality of the *E. petiolatus* genome was evaluated using three distinct approaches. First, RNA-seq reads were aligned to the final assembly using HISAT2 v2.1.0[53], achieving an overall alignment rate of 96.11% (Table 2), reflecting the high completeness of the assembly. Second, BUSCO analysis (v4.1.2)[54] was performed for both genome assembly and annotation using the "embryophyte_odb10" database. Results indicated that 99.07% of complete BUSCO genes were identified at the assembly level, whereas 92.56% were detected at the protein level (Table 2). Finally, the LTR Assembly Index, calculated using LTR_retriever v3.0.1[55], yielded a score of 18.56, indicating the reference genome quality. Together, these metrics confirmed the exceptional completeness and integrity of the *E. petiolatus* genome in this study.

Code availability

No specific script was generated in this study. All commands and pipelines for data analysis followed the user manuals and standard protocols of the relevant bioinformatics software.

Received: 3 March 2025; Accepted: 8 July 2025;
Published online: 16 July 2025

References

1. Coode, M. J. E. in *Flowering Plants·Dicotyledons: Celastrales*, *Oxalidales*, *Rosales*, *Cornales*: Elaeocarpaceae. (ed. Kubitzki, K.) (Heidelberg, Berlin, 2004).
2. Wang, Y. *et al*. Comparison of the chloroplast genomes and phylogenomic analysis of Elaeocarpaceae. *PeerJ* **11**, e15322 (2023).
3. Sharma, S., Hussain, S., Rai, D. V. & Singh, A. N. A comprehensive analysis on the ecosystem services of *Elaeocarpus* L.(Elaeocarpaceae): a review. *J. Phytol.* **15**, 12–37 (2023).
4. Debski, I., Burslem, D. F. & Lamb, D. Ecological processes maintaining differential tree species distributions in an Australian subtropical rain forest: implications for models of species coexistence. *J. Trop. Ecol.* **16**, 387–415 (2000).
5. Prasannan, P., Jeyaram, Y., Pandian, A., Raju, R. & Sekar, S. A review on taxonomy, phytochemistry, pharmacology, threats and conservation of *Elaeocarpus* L.(Elaeocarpaceae). *Bot. Rev.* **86**, 298–328 (2020).
6. El-Dien, R. T. M. *et al*. Antiulcer secondary metabolites from *Elaeocarpus grandis*, family Elaeocarpaceae, supported by in silico studies. *RSC Adv.* **10**, 34788–34799 (2020).
7. Muñoz, O. *et al*. Chemical study and anti-inflammatory, analgesic and antioxidant activities of the leaves of *Aristotelia chilensis* (Mol.) Stuntz, Elaeocarpaceae. *J. Pharm. Pharmacol.* **63**, 849–859 (2011).
8. Quintero-Rincón, P. *et al*. *Sloanea chocoana* and *S. pittieriana* (Elaeocarpaceae): Chemical and biological studies of ethanolic extracts and skincare properties. *Plants* **12**, 3953 (2023).
9. Kwon, O. K. *et al*. Ethanol extract of *Elaeocarpus petiolatus* inhibits lipopolysaccharide-induced inflammation in macrophage cells. *Inflammation* **35**, 535–544 (2012).
10. Cho, E. S. *et al*. Pentacyclic and hexacyclic cucurbitacins from *Elaeocarpus petiolatus*. *Phytochemistry* **193**, 112988 (2022).
11. Sudradjat, S. E. & Timotius, K. H. Pharmacological properties and phytochemical components of *Elaeocarpus*: A comparative study. *Phytomedicine Plus* **2**, 100365 (2022).
12. Jin, L. *et al*. Stronger latitudinal phylogenetic patterns in woody angiosperm assemblages with higher dispersal abilities in China. *J. Biogeogr.* **51**, 269–279 (2024).
13. Bastías, A. *et al*. Draft genome sequence data of maqui (*Aristotelia chilensis*) and identification of SSR markers. *Data Brief* **27**, 104545 (2019).
14. Crayn, D. M., Rossetto, M. & Maynard, D. J. Molecular phylogeny and dating reveals an Oligo-Miocene radiation of dry-adapted shrubs (former Tremandraceae) from rainforest tree progenitors (Elaeocarpaceae) in Australia. *Am. J. Bot.* **93**, 1328–1342 (2006).
15. Anthony, J. M., Allcock, R. J. & Krauss, S. L. Isolation and characterization of 13 microsatellites for the rare endemic shrub *Tetratheca erubescens* (Elaeocarpaceae). *Appl. Plant Sci.* **4**, 1500102 (2016).
16. INIA. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_008921755.1 (2019).
17. Universidad Mayor. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_015342455.1 (2020).
18. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytoch. Bull.* **19**, 11–15 (1987).
19. Louwers, M., Splinter, E., Van Driel, R., De Laat, W. & Stam, M. Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). *Nat. Protoc.* **4**, 1216–1229 (2009).
20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
22. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10 (2018).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
25. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2016).
26. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics.* **25**, 4–10 (2009).
27. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
28. Wang, M. *et al*. Improved assembly and annotation of the sesame genome. *DNA Res.* **29**, dsac041 (2022).
29. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).
30. Daccord, N. *et al*. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
31. Jaillon, O. *et al*. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
32. Wu, G. A. *et al*. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662 (2014).
33. Huang, S. *et al*. The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
34. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 1–9 (2009).
35. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).

36. Haas, B. J. *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
37. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
38. Stanke, M. *et al*. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
39. Haas, B. J. *et al*. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
40. Jin, J. *et al*. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
41. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
42. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
43. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
44. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
45. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1173526 (2024).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBINKO000000000 (2024).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31035705 (2024).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31024753 (2024).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31013082 (2024).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31013083 (2024).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31013084 (2024).
52. Wang, M. Chromosome-level genome assembly and annotation of *Elaeocarpus petiolatus*. *Figshare* https://doi.org/10.6084/m9.figshare.29261864.v1 (2025).
53. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
54. 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 (2015).
55. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

Acknowledgements

This work was supported by the start-up funds provided by Chengdu University (Grant No. 2081921039).

Author contributions

L.W. and M.W. designed the study. H.X. and B.L. analyzed the data and drafted the manuscript. L.W. and M.W. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.W. or L.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.