scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A high-quality chromosome-level genome assembly of *Eucalyptus globulus* Labill.

Xuexian Li[1], Hua Li[1], Zhuangyue Lu[1], Guanben Du[2] ✉ & Xiaoli Wang[1] ✉

Eucalyptus globulus **Labill. is the primary plant material used for the extraction of eucalyptus oil. For this study, we generated a chromosome-level genome assembly of *E. globulus* using a combination of HiFi long-reads, Illumina short-reads and Hi-C data. The assembled genome size was 556.98 Mb, with a contig N50 of 37.93 Mb and scaffold N50 of 54.03 Mb. The completeness of the genome assembly was evaluated at 98.30% by BUSCO, and 99.05% of the genome sequence was anchored to 11 chromosomes. Additionally, 52.55% of repetitive sequences were identified in the genome, and 36,387 protein-coding genes were predicted, of which 96.80% were functionally annotated. Notably, a comparison of the genome of *E. globulus* with six other congeneric species revealed strong conservation, in terms of gene numbers and structures. Overall, we assembled the most complete *E. globulus* genome sequence to date. This will serve as a valuable resource toward the elucidation of the molecular kinetics behind the accumulation of eucalyptus oil in *E. globulus* leaves to facilitate further genetic improvements.**

Background & Summary

Eucalyptus globulus Labill. is a towering evergreen arbor species that belongs to the Myrtaceae family, which is indigenous to Victoria and Tasmanian, Australia. It was originally introduced to China in 1890[1] and is now primarily distributed across Yunnan, Guangxi, Sichuan, and Fujian Provinces. The leaves of *E. globulus* contain an exceptionally high concentration of eucalyptus oil, which is evident when sunlight filters through the foliage to reveal abundant oil glands. Subsequently, these naturally oil-rich leaves and young branches can undergo steam distillation to extract the eucalyptus oil[2], which exhibits potent antibacterial, insecticidal, and antioxidant properties. Consequently, it is used extensively in pharmaceutical products, and due to its sweet and strong taste, has become a popular raw material for daily necessities and cosmetics[3]. *E. globulus* grows rapidly with a strong coppice ability, which can provide valuable renewable resources (e.g., eucalyptus oil, pulp wood, slab timber) within a short rotation period[4,5]. Thus, *E. globulus*, as a tree species that generates both oil and timber in eucalyptus, has a very high economic, social, and ecological value.

In recent years, reports on *E. globulus* have focused mainly on eucalyptus oil extraction techniques from its leaves and tender branches, as well as the analysis of its key components, functions, and applications. Studies have verified that the main components of eucalyptus oil in the leaves of *E. globulus* include 1,8-cineole (commonly also known as cineole), α-pinene, (-)-eucalyptol, α-terpineol, and aromadendrene. Further, the relative content of cineole (50%–85.8%) was the highest in eucalyptus oil[6,7]. Cineole is a cyclic monoterpene compound whose biosynthesis in plants is jointly facilitated through the methylerythritol phosphate (MEP) and mevalonate (MVA) pathways. Additionally, cineole synthase is the direct catalyst for the biosynthesis of cineole[8,9], where numerous studies have found that the expression levels of cineole synthase genes are intimately correlated with the cineole content of plants. Thus, cineole synthase (CinS) genes from different plant sources (e.g., *Lavandula* and *Salvia guaranitica*) have been cloned[10,11] to demonstrate that the enzymes encoded by these genes can convert geranyl diphosphate (GPP) into cineole. Furthermore, by way of the terpene synthesis pathway, genes that encode certain key enzymes (e.g., 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), farnesyl diphosphate synthase (FPS), and geranylgeranyl diphosphate synthase (GGPPS)), have been isolated

[1]The Key Laboratory of Forest Resources Conservation and Utilization in the Southwest Mountains of China Ministry of Education, Southwest Forestry University, 650224, Kunming, China. [2]College of Materials and Chemical Engineering, Southwest Forestry University, 650224, Kunming, China. ✉e-mail: 15125572466@139.com; ynwangxiaoli@swfu.edu.cn

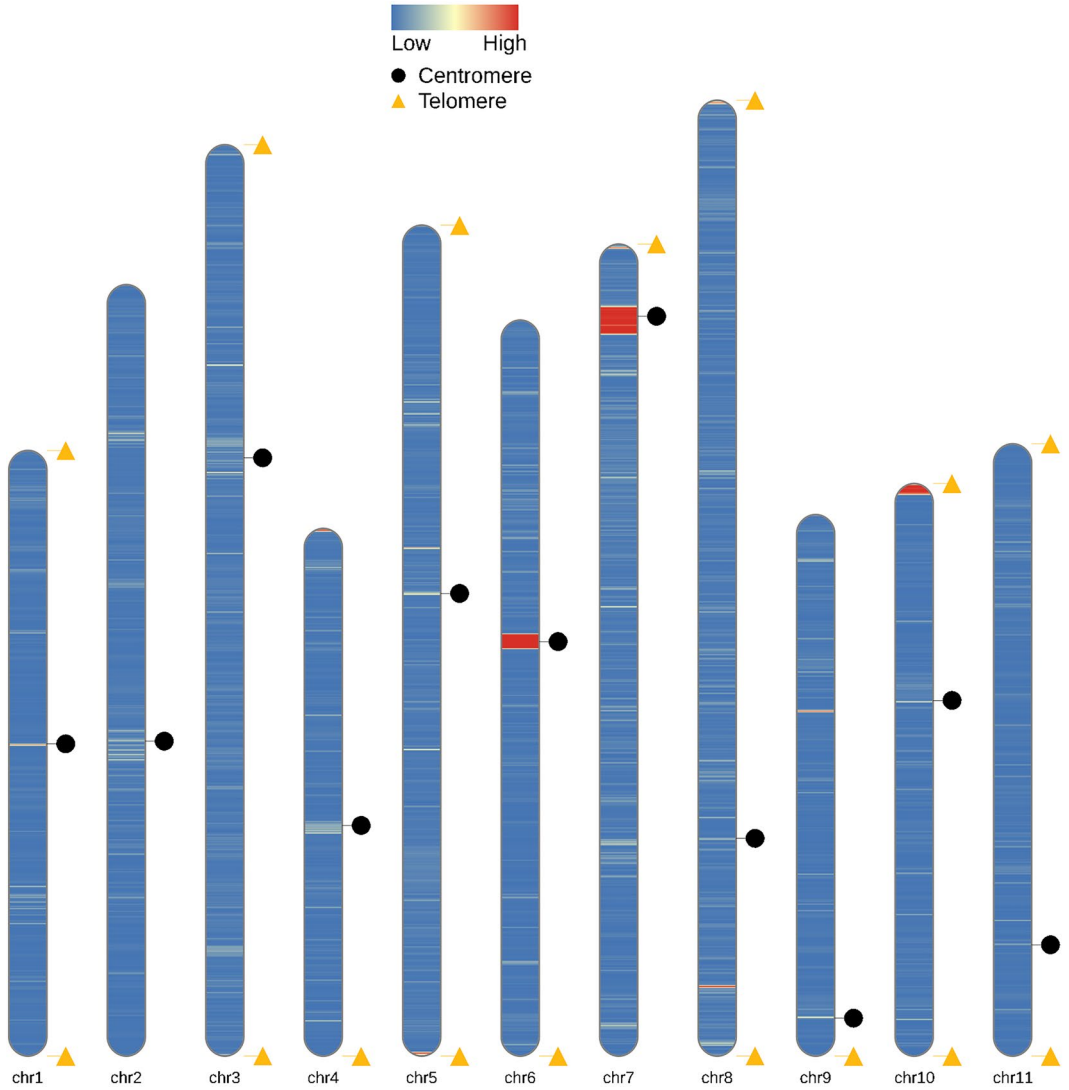

Fig. 1 Schematic diagram of *E. globulus* chromosome structure. Each blue bar represents a chromosome (numbered from chr1 to chr11). Black dots indicate centromeres and yellow triangles indicate telomeres.

from different plants. Their primary functions in expediting the accumulation of terpenoid substances has been verified through heterologous expression[12–16].

However, as the main functional component with high content in eucalyptus oil, research into candidate genes related to cineole biosynthesis in *E. globulus* leaves is lacking due to the absence of genomic investigations. To date, the genomes of several eucalyptus species (e.g., *Corymbia citriodora*, *Eucalyptus urophylla* × *Eucalyptus grandis*, *Eucalyptus regnans*, and *Eucalyptus grandis*)[17–20] have been published. However, as the primary source plant for eucalyptus oil, further genomic research into *E. globulus* is still required. This will not only expand the research scope of eucalyptus species in this field, but also provide an improved elucidation of the molecular mechanisms that underly the production of eucalyptus oil and cineole in this important plant species.

Genome sequences are a valuable resource for the study of species evolution, identification of genetic variations, and selection of traits for tree improvement[21]. Therefore, we generated a chromosome-level genome assembly of *E. globulus* by integrating PacBio Sequel II HiFi long-reads, Illumina NovaSeq X Plus short-reads, and Hi-C data. The assembled genome spanned 556.98 Mb with a scaffold N50 of 54.03 Mb, with 99.05% of the sequences being anchored to 11 chromosomes. Further, 16 of the expected 22 telomeres were assembled, where the complete sequence assembly of telomeres from one to the other was completed on a total of six chromosomes (1, 3, 5, 8, 10, and 11, respectively), which indicated that this was a complete genome assembly, which was close to telomere-to-telomere (T2T) (Fig. 1). Simultaneously, 52.55% of the repetitive sequences were identified in the genome, 36,387 protein-coding genes were predicted, and their functions were annotated. We assembled the most complete genome sequence of *E. globulus* thus far, which provides a valuable resource for exploring the molecular kinetics behind the accumulation of cineole in *E. globulus* leaves and facilitating further genetic improvements. Meanwhile, this chromosome-level genome assembly established the foundation for future evolutionary studies into Myrtaceae and the genomic research of this species.

Type	Platform	Number of Reads	Number of Bases	Mean Read Length	N50 Read Length	GC(%)	Q20 (%)	Q30 (%)
HIFI	PacBio SequeII	1,311,106	21,735,092,245	16,577	16,927	—	—	—
DNA	Illumina Novaseq X Plus	116,625,236	34,987,570,800	—	—	39.47	97.80	93.88
Hi-C	Illumina Novaseq X Plus	331,623,974	99,128,659,338	—	—	40.13	98.25	95.12
RNA	Oxford Nanopore PromethION	9,305,591	12,585,688,622	1,352	1,557	—	—	—

Table 1. Statistics of genomic and transcriptomic sequencing.

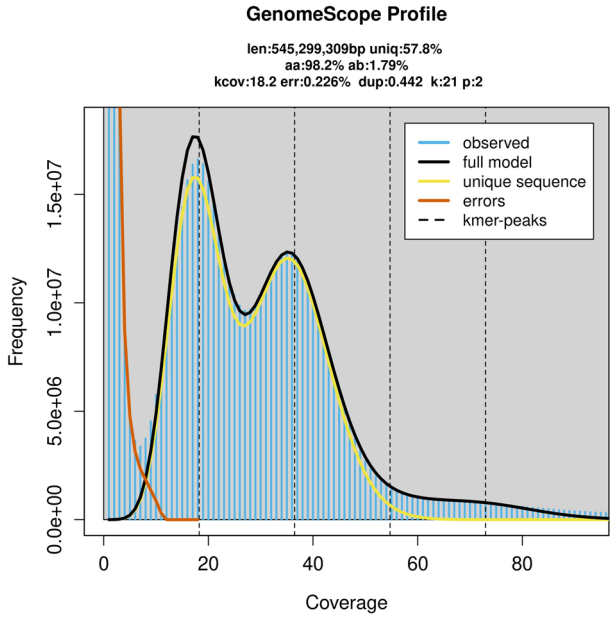

Fig. 2 K-mer analysis (K = 21) of *E. globulus* genome. A GenomeScope Profile plot (K = 21) showing the distribution of observed data (blue), model fit (black), unique sequences (yellow), and error sequences (orange).

Although the pan-genome of *E. globulus* has been published, the integrity of the assembly needs to be improved[22]. The size of the *E. globulus* genome we assembled was 556.98 Mb in contrast to 545.02 Mb when assembled by Ferguson S *et al.* for a study of the evolutionary mechanisms of the genomes of 31 eucalyptus species[22]. Concurrently, the contig N50 assembled in this study was 37.93 Mb, which was much larger than the contig N50 (0.64 Mb) of the *E. globulus* genome in the Ferguson S study. This indicated that we had lower genome fragmentation, better continuity, and higher quality, as the HiFi mode of PacBio used in this study ensured the long read length (10 kb~20 kb) simultaneously, and obtained highly accurate reads (>99%) via the cyclic sequencing of the same DNA sequence.

Methods

Sample collection. The young leaves of a robust 7-year-old *E. globulus* plant devoid of pests and disease were extracted as test materials for genome and transcriptome sequencing at the Southwest Forestry University nursery in Kunming, of Yunnan Province (106.71°E, 29.04°N, 1949.58 m). Once the leaves were removed, they were wrapped in tin foil, labeled, immediately irrigated with liquid nitrogen, and transferred to Genepioneer Biotechnologies (Nanjing, China) for the sequencing and assembly of the *E. globulus* genome.

DNA extraction and sequencing. High-quality genomic DNA was isolated from the young *E. globulus* leaves using an optimized CTAB[23] protocol. To ensure high-quality genome sequencing and assembly, a multi-platform sequencing strategy was employed (Genepioneer Biotechnologies, Nanjing, China). Prior to HiFi sequencing using the PacBio Sequel II platform, genomic DNA samples that passed stringent quality control screening were initially fragmented and then sequentially processed through damage repair, end-repair, adapter ligation, exonuclease digestion, and purification. The final sequencing depth reached ~39×, which yielded 21.74 Gb of high-quality data comprising 1,311,106 reads with an N50 read length of 16.93 kb and mean read length of 16.58 kb (Table 1).

The same DNA samples used for HiFi sequencing were mechanically sheared (via ultrasonication), followed by fragment purification, end repair, A-tailing, adapter ligation, size selection through agarose gel electrophoresis, and PCR amplification for library construction. The qualified library was sequenced on an Illumina NovaSeq X Plus platform (Illumina, CA, USA). Post-filtering, the sequencing yielded 34.99 Gb of high-quality data with a total depth of 63×, where Q20 and Q30 reached 97.80% and 93.88%, respectively, and the GC content was 39.47% (Table 1). These clean reads were processed using Jellyfish (v2.0) for K-mer counting (K = 21), followed by the generation of a K-mer frequency histogram. Subsequent analysis with GenomeScope (v3.3.1) estimated the *E. globulus* genome size to be ~545 Mb, which exhibited 1.79% heterozygosity, containing 44.20% repetitive sequences (Fig. 2).

Parameter	Scaffold	Contig
Length (bp)	556,979,041	556,978,241
GC (%)	39.5	39.5
N50	54,030,587	37,925,749
L50	5	6
N75	42,418,742	22,650,617
L75	8	10
N90	37,925,749	15,591,165
L90	10	14
Longest	66,945,328	63,830,519
Gap number	8	—
N number	800	—

Table 2. Statistics of *E. globulus* genome assembly.

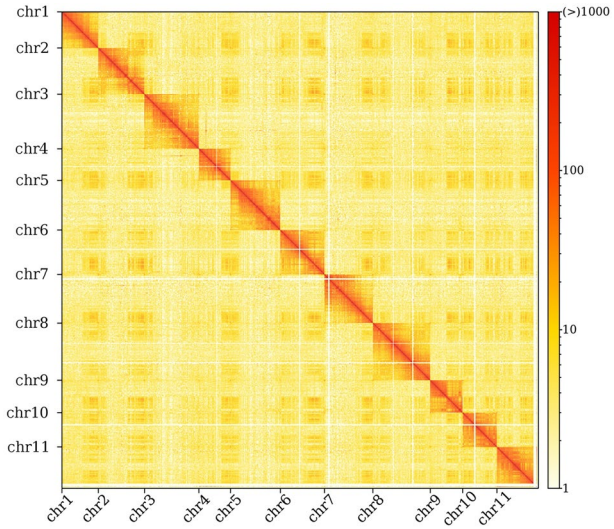

Fig. 3 Hi–C heat map of chromosome interactions. The x-axis and y-axis represent chromosome numbers, respectively. The color intensity indicates the frequency or strength of interactions, with darker colors signifying stronger interactions.

Hi-C libraries were constructed using fresh tissue samples. The samples were fixed with formaldehyde to crosslink proteins to DNA and DNA to DNA within cells. The DNA was then digested with the DpnII restriction enzyme, which created sticky ends on either side of the crosslinked sites. Subsequently, the DNA ends were repaired to incorporate biotin-labeled bases. Next, DNA circularization was performed between the DNA fragments that contained interactions. Once the crosslinks were reversed, the DNA was purified, fragmented into 300 base pairs (bp) to 700 bp segments and captured using streptavidin coated magnetic beads to enrich for the DNA fragments involved in interactions for library construction. Once the library construction was completed, the concentration and insert size of the library were determined using Qubit 2.0 and Agilent 2100, respectively. Quantitative PCR (Q-PCR) was used to accurately quantify the effective concentration of the library to ensure its quality. Once the library passed the quality check, high-throughput sequencing was performed using an Illumina Novaseq 6000 platform, and the raw data sequencing was filtered using fastp (v0.20.0)[24] software to obtain clean data. A total of 99.13 Gb of clean data was obtained from Hi-C processed data, which comprised 331,623,974 clean pair-end reads. The GC content was 40.13%, with Q20 and Q30 scores reaching 98.25% and 95.12%, respectively (Table 1).

RNA extraction and sequencing. The total RNA was extracted from the young *E. globulus* leaves using the RNA prep Pure Plant Plus Kit (Tiangen Biotech (Beijing) Co., Ltd., China) following the manufacturer's instructions. The RNA was reverse-transcribed into cDNA, and a sequencing library was prepared using the cDNA-PCR Sequencing Kit (SQK-PCS109). Single-molecule real-time sequencing was subsequently performed using the Oxford Nanopore PromethION platform, which yielded a final dataset of 12.59 Gb. The dataset consisted of 9,305,591 reads with an N50 read length of 1.6 kb, and a mean read length of 1.4 kb, which were employed for gene prediction (Table 1).

Chromosome	Length (bp)	Contig number	Number of Ns	GC(%)	Number of genes
chr1	42,418,742	1	0	39.50	2,989
chr2	54,030,587	3	200	39.67	3,977
chr3	63,830,519	1	0	39.34	3,533
chr4	36,955,289	1	0	39.23	2,340
chr5	58,195,104	1	0	39.17	3,216
chr6	51,545,551	1	0	39.77	3,982
chr7	56,868,906	3	200	39.24	3,127
chr8	66,945,328	2	100	39.42	4,299
chr9	37,925,749	1	0	39.63	2,566
chr10	40,100,083	3	200	39.82	2,991
chr11	42,883,033	2	100	39.51	3,258
chrUn	5,280,150	12	0	42.92	109
Total	556,979,041	31	800	39.50	36,387

Table 3. Statistics of chromosome assembly.

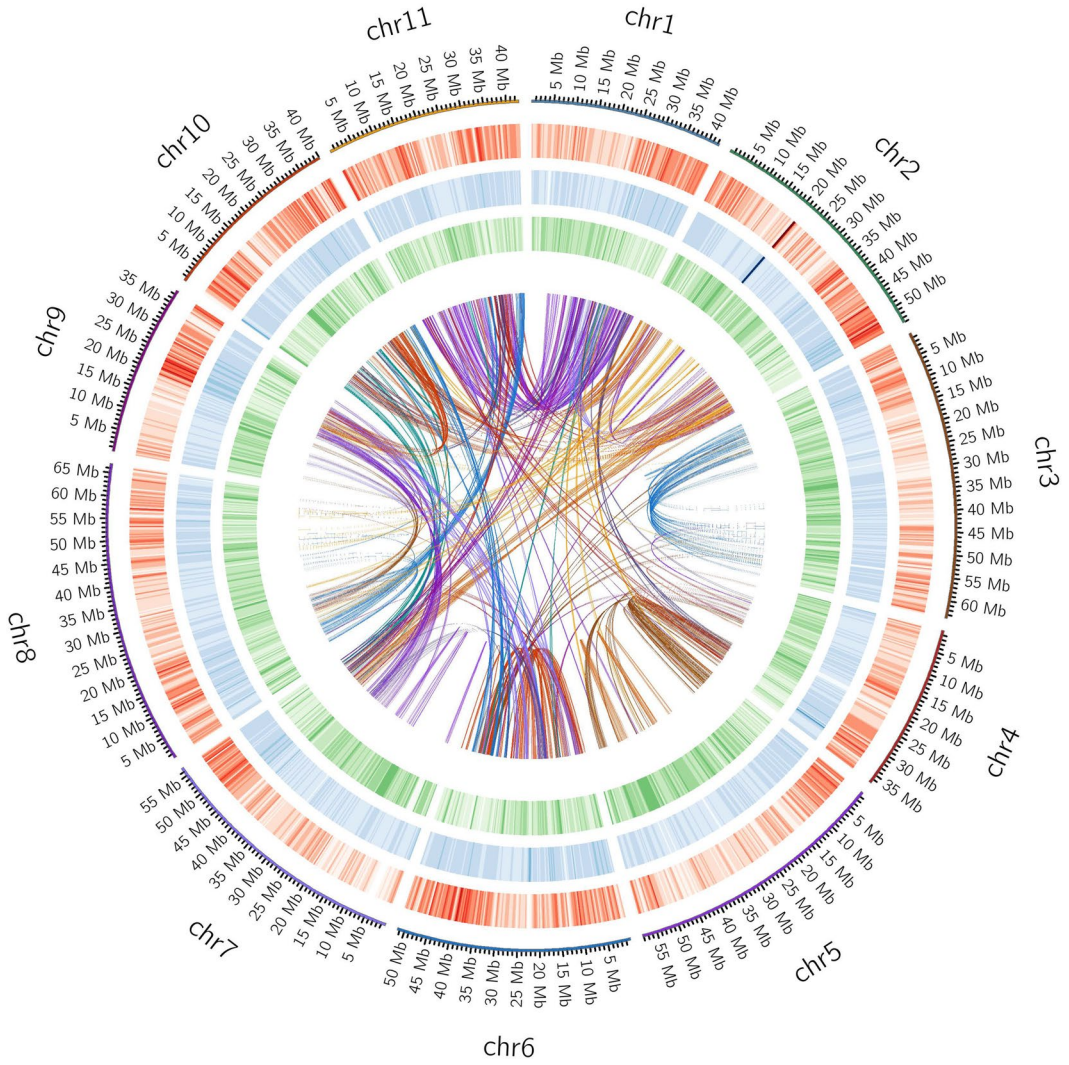

Fig. 4 *E. globulus* genome characteristics. The first circle (from outside to inside) represents different chromosomal sequences, second circle shows gene density, third circle indicates GC content, and fourth circle displays repetitive sequence content. The connecting lines in the center represent genes with collinearity.

Type	De novo + repbase	(%) in genome
DNA	47,962,230	8.61
LINE	16,017,879	2.88
SINE	182,028	0.03
LTR	142,644,625	25.61
Satellite	673,300	0.12
Other	5,287,781	0.95
Unknown	79,912,127	14.35
Total	292,679,970	52.55

Table 4. Repeat sequence statistics.

Method	Software	Species	Gene number	Average gene length(bp)	Average CDS length(bp)	Average exon number per gene	Average exon length(bp)	Average intron length(bp)
Ab initio	AUGUSTUS		35,610	4,096	1,082	4.63	233	830
Ab initio	GlimmerHMM		64,464	7,440	579	2.96	195	3,500
Homology-based	Gemoma	Corymbia citriodora	38,619	3,223	1,231	4.79	257	525
Homology-based	Gemoma	Eucalyptus grandis	38,969	3,463	1,309	5.02	260	535
Homology-based	Gemoma	Psidium guajava	26,336	2,753	1,122	4.43	253	475
Homology-based	Gemoma	Rhodamnia argentea	32,589	3,683	1,382	5.31	260	533
Homology-based	Gemoma	Syzygium grande	37,855	3,341	1,096	4.83	227	586
Homology-based	Gemoma	Syzygium oleosum	34,802	3,608	1,369	5.16	265	538
RNAseq	TransDecoder	TGS	21,882	7,052	1,601	7.13	387	807
Integration	EVM		36,484	3,438	1,269	4.77	266	575
Final set	PASA		36,387	3,950	1,291	5.06	330	581

Table 5. General statistics of predicted protein-coding genes of the *E. globulus* genome.

Genome assembly. Hifiasm software (v0.16.1, https://github.com/chhylp123/hifiasm) with default parameters was used to assemble the HiFi data generated by the PacBio Sequel II platform, while Hi-C clean data was also inputted to assist with improving the assembly contiguity. The primary contigs obtained were then adjusted for subsequent analysis. The final assembly comprised 31 contigs with an N50 of 37.93 Mb, a GC content of 39.5%, and maximum contig length of 63.83 Mb (Table 2). Juicer (v2.0) was utilized to generate a Hi-C interaction matrix from the contigs and Hi-C clean data. This was followed by processing the Hi-C interaction matrix and contigs through 3D-DNA[25] to cluster the contigs and ascertain the degrees of association between them. Subsequently, juciertools (v3.0, https://github.com/aidenlab/JuicerTools) software was used to convert the interactions between the contigs into a specified binary file (hic file), after which the sequenced and oriented contig were manually corrected by Juciebox[26] (v2.15.07) software. Thus, the final chromosome-level (Scaffold) assembly results were obtained. Ultimately, 551.70 Mb of the genome sequence were anchored on 11 chromosomes, which accounted for 99.05% of the total genome sequence (Fig. 3, Table 3). The genome assembly size of *E. globulus* was 556.98 Mb with a scaffold N50 of 54.03 Mb (Table 2). An overview of the genome assembly is shown in Fig. 4.

Gene prediction and functional annotation. To identify repetitive sequences, we initially employed Repeat Modeler (v2.0.1) to conduct the *de novo* prediction of repetitive sequences within the genome. The outcomes of this prediction were then integrated with the RepBase database (http://www.girinst.org/repbase). Subsequently, Repeat Masker (v4.1.0, http://www.repeatmasker.org) was utilized to predict repetitive sequences in the *E. globulus* genome. A total of 292.68 Mb repetitive sequences were identified, which covered 52.55% of the assembled genome, as indicated by *de novo* and homology-based predictions. Among these, the long terminal repeats (LTRs) had the highest abundance (142.64 Mb, 25.61%), followed by DNA transposon repeats (47.96 Mb, 8.61%). The remaining dispersed repetitive sequences exhibited relatively low proportions (0.03%–2.88%), while satellite tandem repeats constituted only 0.12% of the genome (Table 4).

The mRNA prediction was achieved using three approaches: *ab initio*, homology-based, and transcriptome-based prediction. For the *ab initio* prediction, Augustus (v3.3.3, https://github.com/Gaius-Augustus/Augustus) and GlimmerHMM (v3.0.4)[27] were employed. The homology-based prediction was conducted using GeMoMa (v1.9)[28]. For transcriptome-based prediction, RNA-seq data were assembled into transcripts using StringTie (v2.1.3)[29], followed by coding sequence prediction with TransDecoder (v5.1.0, https://github.com/TransDecoder/TransDecoder). The resulting datasets were integrated using EVidenceModeler (v1.1.1)[30]. Finally, the consolidated data were updated with PASA (v2.5.2, https://github.com/PASApipeline/PASApipeline) to add UTR regions and identify novel transcripts. The integrated *de novo*, homology-based, and transcriptome-based predictions identified 36,387 protein-coding genes in *E. globulus*. The average lengths of the exons and introns were 330 bp and 581 bp, respectively. Each gene possessed an average of 5.06 exons, while *E. globulus* had an average coding sequence (CDS) length of 1,291 bp and gene length of 3,950 bp (Table 5). The statistics of the predicted gene models of *E. globulus* exhibited similar distribution patterns in gene length, CDS length, exon length, and intron length compared with its six homologous species (Fig. 5).

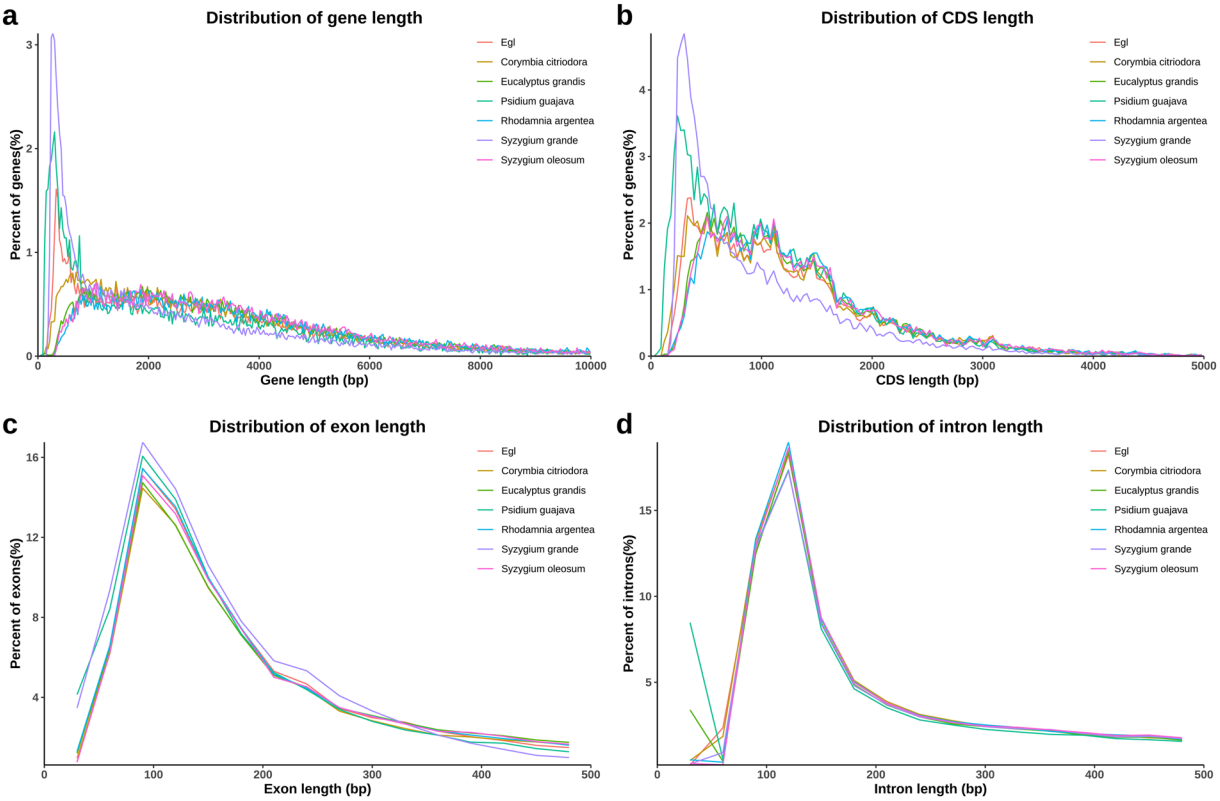

Fig. 5 Comparison of gene structures of seven relative species. (**a**) Gene length distribution. (**b**) CDS length distribution. (**c**) Exon length distribution. (**d**) Intron length distribution. The vertical axis represents the percentage of genes of a certain statistical length from the total number of genes. Various colored lines represent different species.

Type	Number	Average length(bp)	Total length(bp)	(%) in genome
18S_rRNA	41	1,806	74,063	0.0133
28S_rRNA	47	3,635	170,850	0.0307
5.8S_rRNA	45	152	6,868	0.0012
5S_rRNA	2,501	112	280,666	0.0504
rRNA(Total)	2,634	202	532,447	0.0956
tRNA(Total)	533	74	39,905	0.0072
CD-box	256	100	25,825	0.0046
HACA-box	42	127	5,345	0.001
splicing	128	141	18,162	0.0033
snRNA(Total)	426	115	49,332	0.0089
riboswitch	3	145	435	0.0001
Cis-reg:other	5	46	234	0
Cis-reg (Total)	8	83	669	0.0001
miRNA	906	95	86,693	0.0156
ribozyme	152	60	9,207	0.0017

Table 6. Statistics of the noncoding RNA in the *E. globulus* genome.

The prediction of non-coding RNAs consisted of three phases: the prediction of rRNA sequences using barrnap (v0.9, https://github.com/tseemann/barrnap); tRNA sequences prediction using Trnascan[31] (v2.0.0, http://lowelab.ucsc.edu/tRNAscan-SE/); with the remaining non-coding RNAs being predicted by searching the Rfam database (ftp://ftp.ebi.ac.uk/pub/databases/Rfam/) using infernal (v1.1.3, https://github.com/EddyRivasLab/infernal). The non-coding RNA predictions identified a total of 2,634 rRNAs, 426 snRNAs, 906 miRNAs, 533 tRNAs, 152 ribozymes, and only 8 of Cis-regs (Table 6).

The predicted gene annotations were obtained by aligning the longest transcript nucleotide sequences against the Non-Redundant Protein Database[32] (NR), Swiss-Prot Protein Sequence Database[33] (Swiss-Prot), Gene Ontology (GO)[34], Database of Clusters of Orthologous Genes[35] (COG), Eukaryotic Orthologous Groups[36]

Annotation_Database	Annotated_Number	300 <= length <1000	length >= 1000
COG_Annotation	13,424	3,707	9,604
GO_Annotation	25,001	8,831	15,590
KEGG_Annotation	34,800	14,228	19,137
KOG_Annotation	17,939	6,268	11,300
Pfam_Annotation	29,225	10,598	17,994
Swissprot_Annotation	26,938	9,624	16,696
nr_Annotation	35,199	14,444	19,142
All_Annotated	35,223	14,454	19,143

Table 7. Statistics of functional annotation for predicted genes.

	Assembly		Annotation	
Type	Gene number	Percentage(%)	Gene number	Percentage (%)
Complete BUSCOs	1,587	98.3	1,610	99.8
Complete and single-copy BUSCOs	1,540	95.4	1,324	82
Complete and duplicated BUSCOs	47	2.9	286	17.7
Fragmented BUSCOs	17	1.1	3	0.2
Missing BUSCOs	10	0.6	1	0.1
Total lineage BUSCOs	1,614	100	1,614	100

Table 8. The completeness of the genome assembly and annotation was assessed using BUSCO.

(KOG), and the Kyoto Encyclopedia of Genes and Genomes[37] (KEGG) using BLAST (v2.10.1+)[38]. The amino acid sequences were analyzed against the Protein Family Database (Pfam) with HMMER (v3.2.1)[39]. A total of 35,223 genes (96.80%) were annotated, including NR (96.74%), KEGG (95.64%), Pfam (80.32%), Swiss-Prot (74.03%), GO (68.71%), KOG (49.30%), and COG (36.90%), with only 1,164 (3.20%) non-annotated genes (Table 7).

Data Records

The genomic Illumina, PacBio, and Hi-C data were deposited in the NCBI Sequence Read Archive (SRA) database under the accession numbers SRR31482885[40], SRR31482884[41], and SRR31482883[42].

Further, the transcriptomic sequencing data (ONT) were deposited in the SRA database under the accession number SRR31482882[43]. The chromosomal-level genome assembly with annotation is accessible through NCBI GenBank under accession GCA_046226855.146[44] (BioProject: PRJNA1190129; BioSample: SAMN45026323).

This Whole Genome Shotgun project was deposited at GenBank under the accession number JBJKBG000000000[45].

Technical Validation

DNA and RNA integrity. The extracted DNA and RNA were quantified using Nanodrop (Thermo Fisher Scientific, model: NanoDrop2000) and Qubit (Invitrogen, model: QubitTM 3 Fluorometer). The integrity of DNA and RNA was assessed by agarose gel electrophoresis (electrophoresis apparatus: Tanon, model: EPS600; electrophoresis tank: TIANGEN Biochemical Technology (Beijing), model: HE-120).

High-quality DNA samples (concentration $\geq$ 50 (ng/μL), OD260/280 = 1.8~2.2, OD260/230 = 1.8~2.5, fragment size $\geq$ 23 kb, Nanodrop-to-Qubit concentration ratio (N/Q) = 0.9~2.0, showing no visible color or precipitation) were used for the construction of the sequencing library. RNA samples that met the high-quality requirements (total amount $\geq$ 3 μg, sufficient for three library constructions, concentration $\geq$ 40 ng/μL, volume $\geq$ 10 μL, OD260/280 = 1.7~2.5, OD260/230 = 0.5~2.5, normal 260 nm absorption peak, RIN $\geq$ 7.5) were used to construct the sequencing libraries.

Genome assembly and annotation evaluation. We aligned the short reads to the genome to verify accuracy, which ultimately achieved a 98.82% mapping rate.

The completeness of the genome assembly and annotation was assessed using BUSCO v5.2.1 with the embryophyta_odb10 lineage dataset (1,614 conserved single-copy orthologs). The analysis revealed 98.3% completeness (95.4% single-copy, 2.9% duplicated), with 1.1% fragmented and 0.6% missing orthologs (Table 8). BUSCO completeness analysis of the annotated CDS sequences revealed a 99.8% coverage of the complete BUSCO genes (n = 1,610) in the genome annotation. The composition consisted of 82% complete single-copy genes (n = 1,324) and 17.7% duplicated genes (n = 286), with the remaining 0.3% comprising fragmented (0.2%, n = 3) and missing (0.1%, n = 1) orthologs (Table 8).

Code availability

This work did not utilize a custom script. Data processing was performed using the protocols and manuals of the relevant bioinformatics software. Some of the parameters are described in the Methods section, while those not mentioned were set as default parameters.

Received: 27 December 2024; Accepted: 18 June 2025;
Published online: 01 July 2025

References

1. Qi, S. X. *Chinese Eucalyptus* (2nd edn). (Beijing: China Forestry Publishing House, 2002).
2. Li, Y. *et al*. Study on the accumulation pattern of eucalyptus oil and cineole in *Eucalyptus globulus* Labill. leaves. *Plant Sci. J.* **41**, 79–88, https://doi.org/10.11913/PSJ.2095-0837.22081 (2023).
3. Ahmad, R. S. *et al*. *Eucalyptus essential oils*. Ch. 9, 217–239 (Academic Press, 2023).
4. Jorge, F., Quilhó, T. & Pereira, H. Variability of fibre length in wood and bark in *Eucalyptus globulus*. *IAWA J.* **21**, 41–48, https://doi.org/10.1016/B978-0-323-91740-7.00005-0 (2000).
5. Carrillo, I., Vidal, C., Elissetche, J. P. & Mendonça, R. T. Wood anatomical and chemical properties related to the pulpability of *Eucalyptus globulus*: a review. *SOUTH FORESTS*. **80**, 1–8, https://doi.org/10.2989/20702620.2016.1274859 (2018).
6. Song, A. H., Wang, Y. & Liu, Y. M. Study on the chemical components of volatile oil from *Eucalyptus globulus* leaves by Gas Chromatography-Mass Spectrometry. *Food and Drug J.* **11**, 30–32, https://doi.org/10.3969/j.issn.1672-979X.2009.01.010 (2009).
7. Boukhatem, M. N. *et al*. Quality assessment of the essential oil from *Eucalyptus globulus* Labill of Blida (Algeria) origin. *ILCPA*. **17**, 303-315, https://doi.org/10.18052/www.scipress.com/ilcpa.36.303 (2014).
8. Shaw, J. J. *et al*. Identification of a fungal 1,8-cineole synthase from *Hypoxylon* sp. with specificity determinants in common with the plant synthases. *J. Biol. Chem.* **290**, 8511–26, https://doi.org/10.1074/jbc.M114.636159 (2015).
9. Ruan, J. X. *et al*. Isolation and characterization of three new monoterpene synthases from *Artemisia annua*. *Front. Plant Sci.* **7**, 638, https://doi.org/10.3389/fpls.2016.00638 (2016).
10. Demissie, Z. A. *et al*. Cloning, functional characterization and genomic organization of 1, 8-cineole synthases from *Lavandula*. *Plant Mol. Biol.* **79**, 393–411, https://doi.org/10.1007/s11103-012-9920-3 (2012).
11. Ali, M., Alshehri, D., Alkhaibari, A. M., Elhalem, N. A. & Darwish, D. B. E. Cloning and characterization of 1, 8-cineole synthase (SgCINS) gene from the leaves of *Salvia guaranitica* plant. *Front. Plant Sci.* **13**, 869432, https://doi.org/10.3389/fpls.2022.869432 (2022).
12. Yan, X. *et al*. Molecular characterization and expression of 1-deoxy-d-xylulose 5-phosphate reductoisomerase (DXR) gene from *Salvia miltiorrhiza*. *Acta Physiol. Plant*. **31**, 1015–1022, https://doi.org/10.1007/s11738-009-0320-5 (2009).
13. Carretero-Paulet, L. *et al*. Enhanced flux through the methylerythritol 4-phosphate pathway in *Arabidopsis* plants overexpressing deoxyxylulose 5-phosphate reductoisomerase. *Plant Mol. Biol.* **62**, 683–695, https://doi.org/10.1007/s11103-006-9051-9 (2006).
14. Ma, D. *et al*. Overexpression and suppression of *Artemisia annua* 4-hydroxy-3-methylbut-2-enyl diphosphate reductase 1 gene (AaHDR1) differentially regulate artemisinin and terpenoid biosynthesis. *Front. Plant Sci.* **8**, 77, https://doi.org/10.3389/fpls.2017.00077 (2017).
15. Liu, M. *et al*. Cloning, expression characteristics of farnesyl pyrophosphate synthase gene from *Platycodon grandiflorus* and functional identification in triterpenoid synthesis. *Agric. Food. Chem*. https://doi.org/10.1021/acs.jafc.3c09293 (2024).
16. Kai, G. *et al*. Characterization, expression profiling, and functional identification of a gene encoding geranylgeranyl diphosphate synthase from *Salvia miltiorrhiza*. *Biotechnol. Bioprocess Eng.* **15**, 236–245, https://doi.org/10.1007/s12257-009-0123-y (2010).
17. Healey, A. L. *et al*. Pests, diseases, and aridity have shaped the genome of *Corymbia citriodora*. *Commun. Biol.* **4**, 537, https://doi.org/10.1038/s42003-021-02009-0 (2021).
18. Shen, C., Li, L., Ouyang, L., Su, M. & Guo, K. *E. urophylla* × *E. grandis* high-quality genome and comparative genomics provide insights on evolution and diversification of eucalyptus. *BMC Genomics*. **24**, 223, https://doi.org/10.1186/s12864-023-09318-0 (2023).
19. Ferguson, S., Bar-Ness, Y. D., Borevitz, J. & Jones, A. A diploid chromosome-level genome of *Eucalyptus regnans*: unveiling haplotype variance in structure and genes within one of the world's tallest trees. *bioRxiv*. https://doi.org/10.1101/2024.06.29.600429 (2024).
20. Myburg, A. A. *et al*. The genome of *Eucalyptus grandis*. *Nature*. **510**, 356–362, https://doi.org/10.1038/nature13308 (2014).
21. Chen, F. *et al*. The sequenced angiosperm genomes and genome databases. *Front Plant Sci*. **9**, 418, https://doi.org/10.3389/fpls.2018.00418 (2018).
22. Ferguson, S. *et al*. Plant genome evolution in the genus Eucalyptus is driven by structural rearrangements that promote sequence divergence. *Genome Res.* **34**, 606–619, https://www.genome.org/cgi/doi/10.1101/gr.277999.123 (2024).
23. Li, J., Wang, S., Yu, J., Wang, L. & Zhou, S. A modified CTAB protocol for plant DNA extraction. *CBB*. **48**, 72, https://doi.org/10.3724/SP.J.1259.2013.00072 (2013).
24. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
25. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
26. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst*. **6**, 256–258, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
27. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
28. Keilwagen, J., Hartung, F. & Grau, J. *Gene prediction: Methods and protocols*. (Humana Press, 2019).
29. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
30. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
31. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096, https://doi.org/10.1101/614032 (2021).
32. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
33. Amos, B., Rolf, A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **45**, https://doi.org/10.1093/nar/21.13.3093 (2000).
34. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet.* **25**, 25–29 (2000).
35. Tatusov, R. L. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36, https://doi.org/10.1093/nar/28.1.33 (2000).
36. Korf, I. Gene finding in novel genomes. *BMC Bioinf.* **5**, 59 (2004).
37. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462, https://doi.org/10.1093/nar/gkv1070 (2016).
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
39. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome. Inform.* **23**, 205–211, https://doi.org/10.1142/9781848165632_0019 (2009).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31482885 (2024).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31482884 (2024).

42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31482883 (2024).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31482882 (2024).
44. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046226855.1 (2024).
45. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBJKBG000000000 (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32260385); Seed Industry Joint Laboratory Project of Yunnan Province (202205AR070001-22); Scientific Research Fund of Yunnan Provincial Department of Education (2023Y0737); and the Scientific Research Fund of Yunnan Provincial Department of Education (2025Y0862).

Author contributions

The authors confirm contributions to the paper as follows: conceptualization, writing-original draft: Xuexian Li, Xiaoli Wang; methodology: Xuexian Li, Hua Li, Zhuangyue Lu; software, investigation: Hua Li, Xuexian Li, Zhuangyue Lu; validation, formal analysis, data curation, visualization: Xiaoli Wang, Xuexian Li, Hua Li; writing-review & editing: Xuexian Li, Xiaoli Wang; project administration: Xiaoli Wang, Xuexian Li, Hua Li; supervision, funding acquisition: Xiaoli Wang, Guanben Du. All authors reviewed the results 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 G.D. or X.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.