scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Near telomere-to-telomere genome assembly of *Camellia pitardii*

Zhiyuan Lv[1,2,3], Jiao Wang[1,2,3], Ling Zhou[1,2,3], Shihui Zou[1,2,3] & Lijiao Ai[1,2,3] ✉

***Camellia pitardii*** **Cohen-Stuart, is an endemic winter-spring flowering wild *Camellia* species in southwestern China, with important ecological and horticultural value. Here, we report the near telomere-to-telomere (T2T) genome of *C. pitardii* by combining PacBio HiFi, ONT ultra-long, Hi-C and Illumina sequencing platforms. The genome size was 2.63 Gb, with repetitive content of 78.91% and contig N50 of 146.18 Mb. 95.5% of genomic sequences were anchored onto 15 chromosomes, including eight gap-free chromosomes, with 25 telomeres and 15 centromeres. A total of 52,848 protein-coding genes were predicted, of which 98.27% were functionally annotated. The genomic quality was further confirmed by the quality value (QV) of 44.87 and the Benchmarking Universal Single-Copy Orthologs (BUSCO) value of 96.34%. Comparative genomic analysis showed that *C. pitardii* had a close genetic relationship with *C. oleifera* and *C. chekiangoleosa*. The availability of *C. pitardii* near T2T genome will provide a foundation for identifying key genes related to stress resistance, flowering time regulation, and flower color formation, and promotes the breeding, utilization and conservation of the germplasm resources.**

## Background & Summary

*C. pitardii* Cohen-Stuart, a member of the genus *Camellia* in the Theaceae family, is an evergreen shrub or small tree endemic to southwestern China[1]. This species primarily inhabits understory, forest edges, and shrubland ecosystems at altitude of 800–2200 m, characterized by a flowering period spanning late winter to spring (Fig. 1a). *C. pitardii* has dark green leathery foliage and brightly colored terminal flowers with white to rose-pink (Fig. 1b,c). These distinctive phenological patterns and morphological adaptations synergistically confer remarkable ecological plasticity, positioning it as a valuable germplasm resource for ornamental cultivation in landscape horticulture. Despite *C. pitardii* has dual ecological and horticultural value as a winter-spring flowering ornamental plant, its development and utilization in production and progress in scientific research are relatively scarce. Particularly in the present era of rapid development in molecular biology, the deficiency of genomic information of *C. pitardii* impedes the molecular-level research on its adaptive evolution and biosynthetic potential.

Chromosome-level genome assemblies have become indispensable for deciphering plant evolution, adaptive traits, and biosynthetic pathways. Within the genus *Camellia*, reference genomes of tea trees (*C. sinensis*)[2–5], oil-Camellias (*C. oleifera*)[6–8], and the ornamental Camellias (*C. japonica*)[9,10] provide valuable information for genomic diversity and architecture, thus promoting the analysis of population and functional genes. However, there are more than 200 species in the genus *Camellia*, and among the species that have obtained genome assembly, the genome of ornamental Camellia is far behind that of tea trees and oil-Camellias in terms of quantity and quality. Ornamental Camellias remain genomically underrepresented, impeding comparative analyses of their evolutionary trajectories and molecular mechanisms underlying adaptive traits. Recent advances in long-read sequencing enable precise resolution of complex genome, and high-quality genome assembly lays the foundation for the in-depth study of complex regions such as centromere and telomeres, and is of great significance for the mining of functional genes and the study of important biological mechanisms[11].

In this study, we report the first near telomere-to-telomere (T2T) genome assembly of a wild Camellia species, *C. pitardii*, generated by integrated PacBio HiFi sequencing, ONT ultra-long sequencing, Hi-C and

[1]Chongqing Landscape and Gardening Research Institute, Chongqing, 401329, China. [2]Chongqing Key Laboratory of Germplasm Innovation and Utilization of Native Plants, Chongqing, 401329, China. [3]Chongqing Urban Landscaping Engineering Technology Research Center, Chongqing, 401329, China. ✉e-mail: alj01461@foxmail.com


**Fig. 1** The overview features of the *C. pitardii*. (**a**) Wild habitat of *C. pitardii*. (**b**) Bloomed flower, showing yellow anthers and pink petals. (**c**) Floral bud.

Illumina sequencing technology. The final assembled genome was 2.63 Gb, with repetitive content of 78.91% and contig N50 of 146.18 Mb. Based on the karyotype of the species (2n = 30), The genomic sequences were anchored onto 15 chromosomes, and 25 telomeres and 15 centromeres were identified. This study constructed the high-quality reference genome of *C. pitardii*, with eight gap-free chromosomes and remaining seven chromosomes containing nine gaps. Genome annotation identified 52,848 protein-coding genes, and 98.27% were functionally annotated. The near T2T genome assembly contributes to the population genetic and evolutionary analysis of *C. pitardii*, while promoting germplasm utilization for breeding more ornamental Camellia varieties.

## Methods

**Plant materials, sample collection, and sequencing.** Young floral buds were collected from wild *C. pitardii* on Jinfo Mountain, Nanchuan District, Chongqing, China (Fig. 1c). The collected samples were immediately flash-frozen in liquid nitrogen and stored at −80 °C prior to downstream analysis. High molecular weight genomic DNA (gDNA) was extracted from young floral buds using an improved CTAB method[12]. Quantitative measurement of gDNA was ultimately achieved using the Qubit®3.0 Fluorometric system (Invitrogen, USA). Then, gDNA was sheared into ~15 kb fragments by Megaruptor®2 (Diagenode). The SMRTbell library was constructed using the SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, California, USA). Library size and quantity were assessed using the FEMTO Pulse and the Qubit dsDNA HS reagents Assay kit. Subsequent circular consensus sequencing (CCS) was then performed on the PacBio Sequel II platform using the Sequel II Sequencing Kit. We obtained 148.72 Gb HiFi reads with an average read length of 14.52 kb, which covered about 58 × of the *C. pitardii* genome (Table S1). Ultra-high molecular weight gDNA was extracted from young floral buds of *C. pitardii*, purified and assessed using a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). For ONT ultra-long sequencing, the standard library was prepared using the SQK-ULK114 kit following the standard protocol. The purified library was sequenced using a PromethION48 sequencer (Oxford Nanopore Technologies, Oxford, UK). A clean read of 150.7 Gb was obtained, with an average read length of 98.15 kb, which covering approximately 50 × of the *C. pitardii* genome (Table S1). Clean reads were preserved for downstream genome assembly. For Hi-C sequencing, formaldehyde was used for crosslinking the fresh young floral buds. Genomic DNA was digested with the restriction endonuclease *Hin*d III and then labeled with biotin through end repair. The purified DNA fragments were circularized and broken into a length of 300–700 bp. Finally, the DNA fragments captured by streptavidin protein beads were used for Hi-C library construction. Hi-C library sequencing was performed on the Illumina platform (PE 150 bp), and 332.81 Gb clean reads were generated. The quality assessment of Hi-C library was conducted using HiC-Pro (v2.10.0)[13]. The valid interaction pairs were retained for subsequent genome assembly. To assist in the annotation of the *C. pitardii* genome, total RNA was isolated from five different tissues (leaf, root, floral bud, stem, flower) with TRIzol reagent (Invitrogen, USA). After assessment and measurement, the library sequencing was performed on the Illumina platform (PE 150 bp).

**Genome assembly and quality assessment.** Initially, the CCS data were used to estimate genome size and heterozygosity rate of *C. pitardii* based on *K*-mer analysis with Jellyfish (v2.1.4; parameter: -h 10,000,000,000)[14] and GenomeScope (v2.0; parameters: -k 21 -p 2 -m 100,000)[15]. A 21-mer frequency distribution was generated. Consequently, genome survey indicated the genome of *C. pitardii* was about 2.53 Gb with 1.92% heterozygosity (Fig. 2). High-quality genome assembly was performed on *C. pitardii* using multiple sequencing platforms. The high accuracy PacBio HiFi reads and ONT ultra-long reads were assembled using hifiasm (v0.16)[16] with default parameters to generate a primary contig genome. The 332.8 Gb Hi-C clean reads were mapped to the assembled genome by BWA (v0.7.17-r1188)[17] with the default parameters. Assembly contamination, self-ligation, non-ligation and other invalid reads were filtered. The Lachesis software (v2.0.1; parameters: CLUSTER_MIN_RE_SITES = 358; CLUSTER_MAX_LINK_DENSITY = 2; ORDER_MIN_N_RES_IN_TRUNK = 25; ORDER_MIN_N_RES_IN_SHREDS = 45)[18] was used to correct, cluster, and orient the contigs. The genome generated

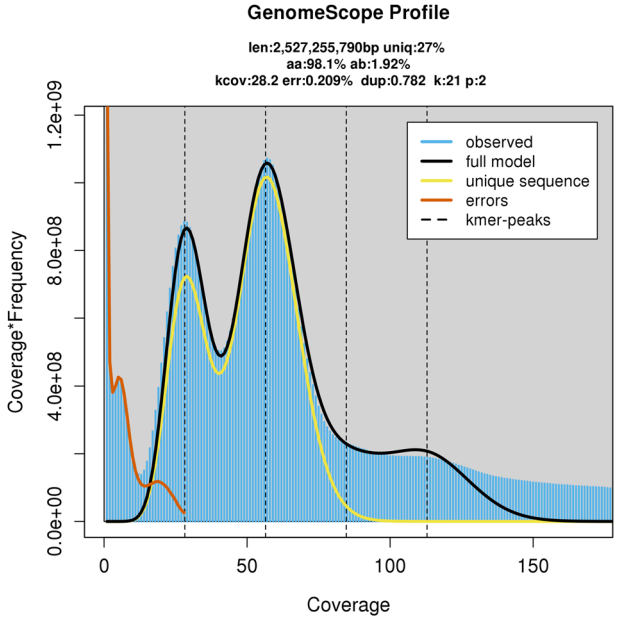


**Fig. 2** Estimations of genome size and heterozygosity rate based on *K*-mer counting. The following formula is used for genome size assessment: Genome size = (total number of *K*-mers)/(the volume peak).

with HiFi and ONT data formed 15 large contigs representing 15 chromosomes. To further polish the genome, the HiFi-based reads and ultra-long reads were implemented to fill and correct the gaps through quarTeT (v1.1.6)[19]. Finally, a near T2T genome was obtained for *C. pitardii* with nine gaps, showing a genome size of 2.63 Gb, contig N50 length of 146.18 Mb, and GC content of 38.67%; 95.5% of genome sequence was successfully anchored onto 15 large contigs representing 15 chromosomes (Fig. 3, Fig. S1, Table 1, Table S2).

In order to verify the integrity and accuracy of the assembled near T2T genome, BUSCO (Benchmarking Universal Single-Copy Orthologs, v5.2.1)[20] assessments of genome assembly were performed based on embryophyta_odb10[20]. The analysis revealed that 96.34% of the core conserved genes were detected, including 1414 single-copy and 141 duplicated genes (Table 1). A high mapping rate of Illumina short reads (98.60%), HiFi long reads (99.85%), and ultra-long data (99.27%) were aligned against the genome (Table S3). The assembly quality of the near T2T genome was also assessed using the *K*-mer method implemented in Merqury[21] based on Illumina data. The quality value (QV) score of the genome was 44.87 (Table 1). The completeness test of long terminal repeat (LTR) showed a 14.08 LTR assembly index (LAI) value for the assembly. These data suggested a high-quality *C. pitardii* genome assembly.

**Telomere and centromere identification.** Chromosome ideograms, transposable element (TE) density, SSR density, gene density and GC content were represented in the circos plot (Fig. 3a). To investigate the genome structure, telomeres and centromeres were identified in the current *C. pitardii* genome. The telomere repeat units were explored by using the TIDK (v0.1.5)[22] with default parameters. A total of 25 telomeres were detected using repetitive sequences (5′-AAACCCT -3′ and 5′-AGGGTTT -3′) as queries (Fig. 3c). Five of the 15 chromosomes (chr 1, 7, 9, 11 and 14) have one telomere, and all other chromosomes have two telomeres at each chromosome end (Fig. 3c). Centromics software (v0.3; https://github.com/zhangrengang/Centromics) was used to identify centromeres. The assembled genome was analyzed using HiFi reads and ONT ultra-long reads to identify identify potential centromeric repeat sequences. A total of 15 centromeres were identified (Fig. 3c).

**Repetitive sequences annotation.** TEs were identified by combination of homology-based and *de novo* approaches. Firstly, a *de novo* repeat library of the genome was customized using RepeatModeler (v2.0.1)[23], which automatically executed two *de novo* repeat finding programs, including RECON (v1.0.8)[24] and RepeatScout (v1.0.6)[25]. Then full-length long terminal repeat retrotransposons (LTR-RTs) were identified using both LTR harvest (v1.5.10)[26] and LTR_FINDER (v1.07)[27]. The intact LTR-RTs and non-redundant LTR library were then produced by LTR_retriever (2.9.0)[28]. Final TE sequences in the *C. pitardii* genome were identified and classified by homology search against the library using RepeatMasker (v4.1.2; parameters: repeatmasker -nolow -no_is -norna -engine wublast -parallel 8 -qq)[29]. Tandem repeats were annotated by MIcroSAtellite identification tool (MISA v2.1)[30] and Tandem Repeat Finder (TRF 409)[31]. The annotation results showed that 2,076.51 Mb repeat sequences were identified, which accounted for 78.91% of the *C. pitardii* genome. The content of tandem repeat sequences and TEs was 10.63% and 68.28%, respectively. Among TE sequences, long terminal repeat retrotransposons (LTR-RTs) accounted for 53.84%, SINEs accounted for 0.12% and LINEs accounted for 0.89% (Table S4).

**Genome annotation and functional prediction.** The annotation of protein-coding genes in *C. pitardii* was performed through a comprehensive framework utilizing three complementary methods: *ab initio* prediction,

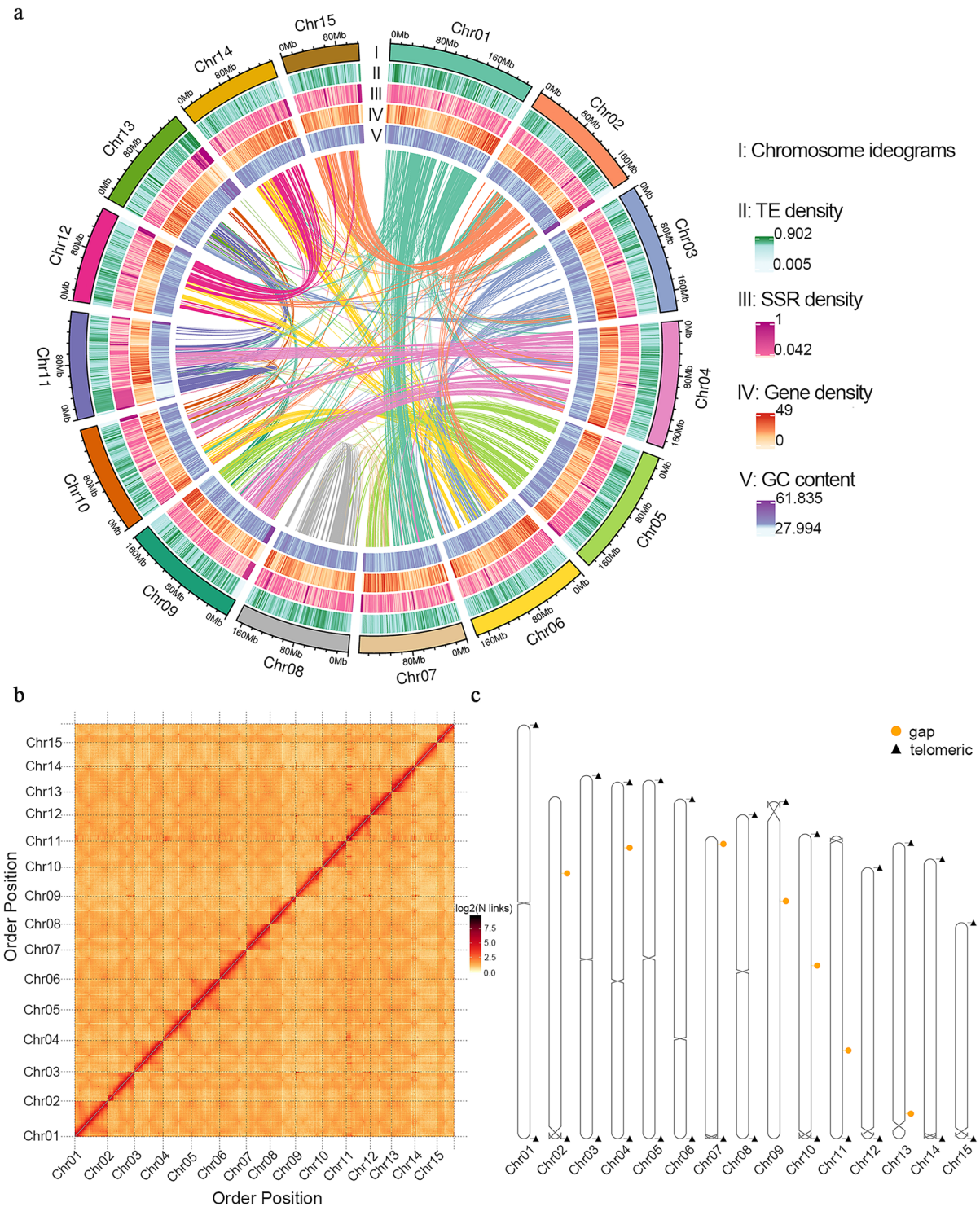


**Fig. 3** Features of the *C. pitardii* genome and Hi-C-assisted genome assembly. (**a**) Overview of features of the *C. pitardii* genome. From outside to inside, (I) chromosome ideograms, (II) TE density, (III) SSR density, (IV) Gene density, (V) GC content. (**b**) Hi-C contact map. The horizontal and vertical axes represent the order of each bin in the corresponding chromosome group. (**c**) Distribution of telomeres and centromeres on *C. pitardii* chromosome. Triangles and circles represent telomeres and centromeres, respectively.

homology search, and transcriptome-based assembly. The *ab initio* gene prediction was performed using two *ab initio* gene-prediction software tools, Augustus (v3.1.0)[32] and SNAP (v2006-07-28)[33] with default parameters. Gene annotation of homologous species and transcriptome was used as the training set for Augustus. For homology-based prediction, GeMoMa (v1.7)[34] was employed to perform reference-guided homology inference

| Genomics feature | Value |
|---|---|
| Assembled genome size (bp) | 2,631,421,036 |
| Number of Chromosome | 15 |
| Number of Scaffold | 363 |
| Number of gap-free chromosomes | 8 |
| Chromosome length (bp) | 2,513,055,549 |
| Complete (%) | 95.50 |
| Number of genes | 52,848 |
| Scaffold N50 (bp) | 169,390,313 |
| Contig N50 (bp) | 146,181,769 |
| Repetitive sequence (%) | 78.91 |
| LTR assembly index (LAI) | 14.08 |
| GC content (%) | 38.67 |
| Quality value (QV) | 44.87 |
| Genome BUSCOs (%) | 96.34 |
| BUSCO genes (%) | 96.96 |
| Pseudogene prediction | 968 |

**Table 1.** Summary statistics of *C. pitardii* genome assembly.

using gene model from four phylogenetically representative species: *Arabidopsis thaliana*, *C. chekiangoleosa, C. lanceoleosa*, and *C. sinensis*. For the transcriptome-based prediction, RNA-sequencing data were mapped to the reference genome using Hisat (v2.1.0; parameters: hisat2--dta -p 10)[35] and assembled by Stringtie (v2.1.4; parameter: stringtie -p 2)[36]. GeneMarkS-T (v5.1)[37] with default parameters was used to predict genes based on the assembled transcripts. Transcriptome-based prediction involves two approaches to assemble transcripts for gene prediction. In addition, PASA (v2.4.1)[38] with default parameters was used to predict genes based on the transcripts assembled by Trinity (v2.11)[39]. Gene models from these different approaches were combined using the EVidenceModeler (EVM v1.1.1)[40] and updated by PASA (v2.4.1). In total, 52,848 protein-coding genes were predicted in the assembled *C. pitardii* genome. The average coding sequence length was 1,664.52 bp, and the average number of exons was 4.64 (Fig. 4). The protein domains were annotated using InterProScan (v5.34–73.0; parameters: interproscan.sh -iprlookup -pa -f xml -dp -t p -cpu 10)[41] based on InterPro protein database. The motifs and domains within gene models were identified by PFAM database. Gene Ontology (GO) IDs for each gene were obtained from TrEMBL[42], InterPro[43] and EggNOG[44]. In total, approximately 51,936 (~98.27%) of the predicted protein-coding genes could be functionally annotated (Table 2).

**Annotation of non-coding RNAs.** The tRNAscan-SE (v1.3.1)[45] algorithms with default parameters was used to identify the transfer RNA (tRNA) sequences within the *C. pitardii* genome. Barrnap (v0.9)[46] with default parameters was used to identify the genes associated with ribosomal RNA (rRNA). Small nucleolar RNAs (snRNAs) are a class of small RNA molecules that guide chemical modifications of other RNAs, mainly rRNAs, tRNAs and snRNAs. MicroRNAs (miRNAs) and snRNAs were identified by Infernal (v1.1)[47] software against the Rfam (v14.5)[48] database with default parameters. As a result, a total of 1,691 tRNAs, 5,772 rRNAs, 64 miRNAs and 119 snRNAs were predicted (Table 3).

**Pseudogene prediction.** The GenBlastA (v1.0.4)[49] program was used to scan the whole genomes after masking predicted functional genes. Putative candidates were then analyzed by searching for premature stop mutations and frame-shift mutations using GeneWise (v2.4.1)[50]. Finally, a total of 968 pseudogenes were predicted (Table 1).

## Data Records

All raw data of the whole genome have been deposited in National Genomics Data Center (NGDC)[51], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation under BioProject accession numbers PRJCA039342. The genomic PacBio sequencing data, the ultra-long ONT sequencing data, the Hi-C sequencing data, and the RNA sequencing data were deposited in Genome Sequence Archive (GSA) of NGDC with accession number CRA025251[52], CRA025199[53], CRA025200[54] and CRA025201[55]. The genome assembly has been deposited at GenBank under the accession JBPSJW000000000[56]. The genome assembly has also been deposited in Genome Warehouse (GWH) of NGDC under accession number GWHGDHR00000000.1[57] and figshare[58]. The genome annotation data are available from Zenodo https://zenodo.org/records/15744655[59].

## Technical Validation

To evaluate the quality of genome assembly, multiple parameters were employed. Firstly, the Hi-C contact map of the *C. pitardii* genome demonstrated high concordance among all chromosomes, reflecting the accuracy of sequencing, ordering, and orientation of assembled contigs (Fig. 3). Secondly, all 15 centromeres were identified and mapped on 15 chromosomes. Eight chromosomes assembled as gap-free T2T level. Subsequently,

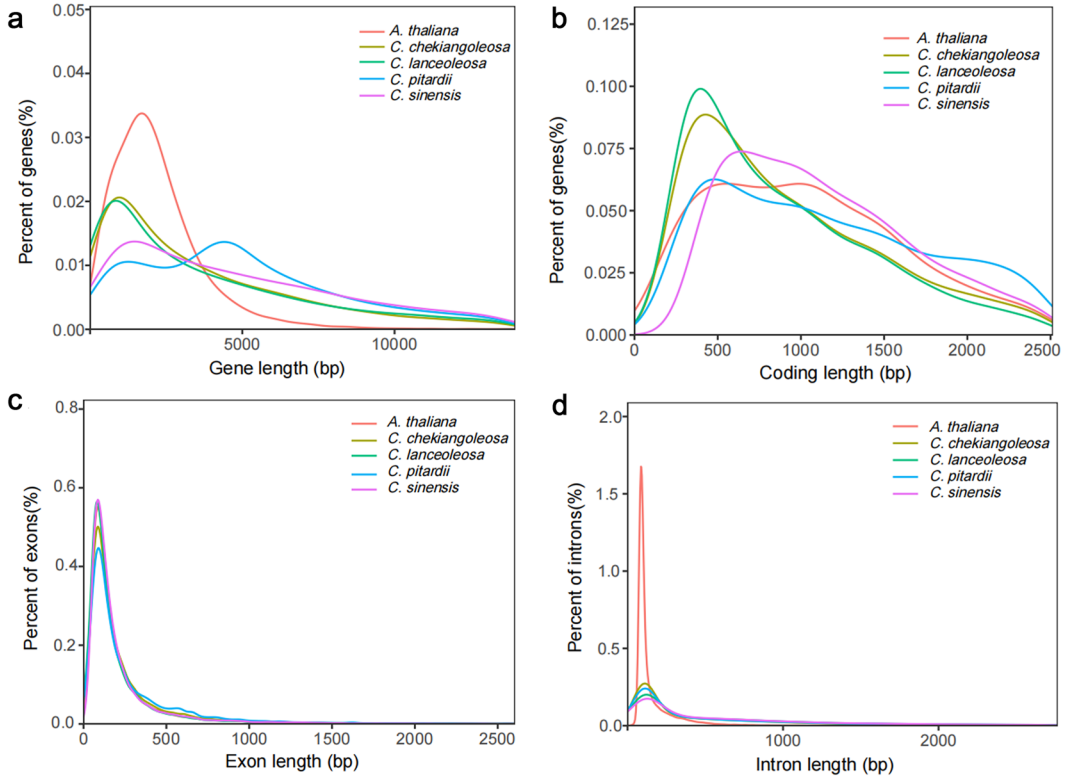


**Fig. 4** The distribution map of genetic characteristics of *C. pitardii* and four other plants. (**a**) gene length distribution. (**b**) Coding sequence length distribution. (**c**) exon length distribution. (**d**) intron length distribution.

| Categories | Annotated gene number | Percent (%) |
|---|---|---|
| GO | 40,662 | 76.94 |
| KEGG | 38,940 | 73.68 |
| KOG | 27,076 | 51.23 |
| Pfam | 44,221 | 83.68 |
| Swissprot | 38,702 | 73.23 |
| TrEMBL | 51,893 | 98.19 |
| eggNOG | 40,743 | 77.09 |
| Nr | 49,791 | 94.22 |
| All | 51,936 | 98.27 |

**Table 2.** Summary of gene function annotations.

| rRNA_num | tRNA_num | miRNA_num | snRNA_num | snoRNA_num |
|---|---|---|---|---|
| 5,772 | 1,691 | 64 | 61 | 119 |

**Table 3.** Statistics of noncoding genes.

the OrthoDB 10 embryophyta database was selected, and the BUSCO assessment of completeness was 96.34% (Fig. 5). Using the BWA[17] to align the short sequences obtained from Illumina sequencing with the assembled genome, the mapped rate was 98.60%. Using the Minimap2[60] to align the third-generation clean reads with the assembled genome, 99.85% of HiFi reads and 99.27% of ONT ultra-long reads could be aligned (Table S3). Additional validations: the QV of assembly was quantified using Merqury[21], resulting in QV of 44.87 (Table 1). LTR retrotransposon annotation using LTR_FINDER (v1.0.7)[40], LTRharvest (v1.5.9)[39], and LTR_retriever (v2.8)[41] demonstrated an LAI of 14.08 (Table 1). In summary, these results confirm that the near T2T genome assembly of *C. pitardii* exhibits high completeness and accuracy.

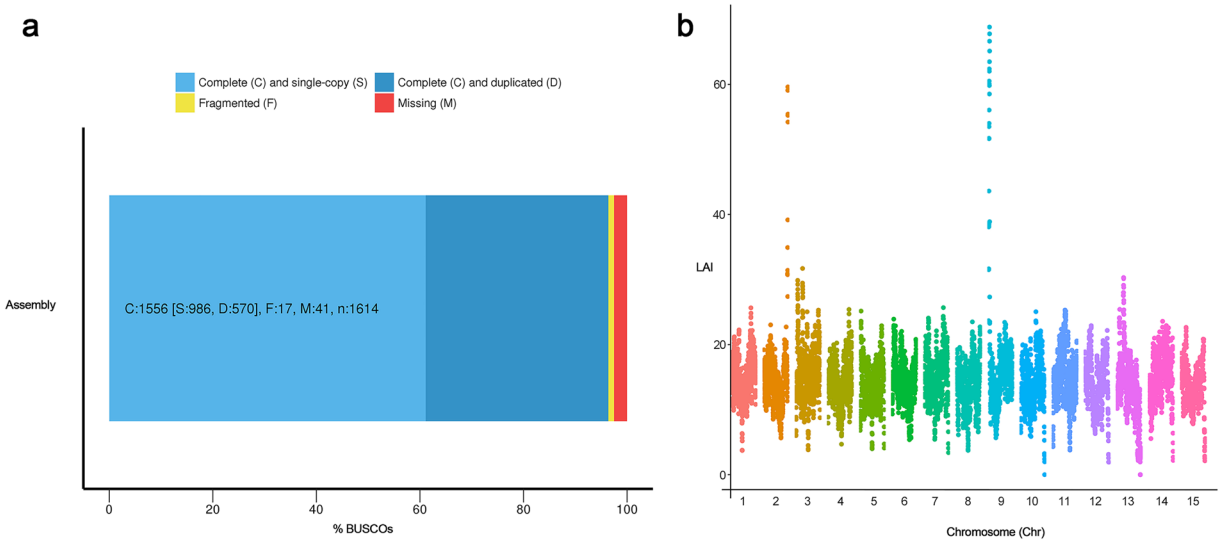


**Fig. 5** Completeness of the *C.pitardii* genome assembly. (**a**) The BUSCO assessment results of the genome. (**b**) LAI scores of the *C.pitardii* genome.

## Code availability

All software and pipelines were executed in accordance with the manuals and programs of the published bioinformatic tools. All programs used in this study are publicly available, with version and parameters explicitly described in the Methods section.

Received: 5 May 2025; Accepted: 1 August 2025;
Published online: 14 August 2025

## References

1. Editorial Committee of the Flora of China of Chinese Academy of Science. Flora of China. Beijing: Beijing Science Press (1998).
2. Wang, X. *et al.* Population sequencing enhances understanding of tea plant evolution. *Nature Communications* **11**, 4447, https://doi.org/10.1038/s41467-020-18228-8 (2020).
3. Zhang, X. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nature Genetics* **53**, 1250–1259, https://doi.org/10.1038/s41588-021-00895-y (2021).
4. Zhang, W. *et al.* Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications* **11**, 3719, https://doi.org/10.1038/s41467-020-17498-6 (2020).
5. Kong, W. *et al.* Genomic analysis of 1,325 *Camellia* accessions sheds light on agronomic and metabolic traits for tea plant improvement. *Nature Genetics* **57**, 997–1007, https://doi.org/10.1038/s41588-025-02135-z (2025).
6. Lin, P. *et al.* The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. *Genome Biology* **23**, 14, https://doi.org/10.1186/s13059-021-02599-2 (2022).
7. Zhang, L. *et al.* The tetraploid *Camellia oleifera* genome provides insights into evolution, agronomic traits, and genetic architecture of oil Camellia plants. *Cell Reports* **43**, 114902, https://doi.org/10.1016/j.celrep.2024.114902 (2024).
8. Zhu, H. *et al.* The complex hexaploid oil-Camellia genome traces back its phylogenomic history and multi-omics analysis of Camellia oil biosynthesis. *Plant Biotechnology Journal* **22**, 2890–2906, https://doi.org/10.1111/pbi.14412 (2024).
9. Hu, Z. *et al.* Genomics insights into flowering and floral pattern formation: regional duplication and seasonal pattern of gene expression in *Camellia*. *BMC Biology* **22**, 50, https://doi.org/10.1186/s12915-024-01851-y (2024).
10. Cai, Y. *et al.* Nine complete chloroplast genomes of the *Camellia* genus provide insights into evolutionary relationships and species differentiation. *Scientific Reports* **15**, 8783, https://doi.org/10.1038/s41598-025-87764-4 (2025).
11. Hou, X. *et al.* A near-complete assembly of an *Arabidopsis thaliana* genome. *Molecular Plant* **15**, 1247–1250, https://doi.org/10.1016/j.molp.2022.05.014 (2022).
12. Abu Almakarem, A. S. *et al.* Extraction of DNA from plant and fungus tissues *in situ*. *BMC Research Notes* **5**, 266, https://doi.org/10.1186/1756-0500-5-266 (2012).
13. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259, https://doi.org/10.1186/s13059-015-0831-x (2015).
14. Marçais, G. *et al.* A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
15. Ranallo-Benavidez, T. R. *et al.* GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
16. Cheng, H. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
17. Li, H. *et al.* Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
18. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119–1125, https://doi.org/10.1038/nbt.2727 (2013).
19. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, https://doi.org/10.1093/hr/uhad127 (2023).
20. Simão, F. A. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).

21. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
22. Brown, M. R. *et al*. Tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, https://doi.org/10.1093/bioinformatics/btaf049 (2025).
23. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
24. Bao, Z. *et al*. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research* **12**, 1269–1276, https://doi.org/10.1101/gr.88502 (2002).
25. Price, A. L. *et al*. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
26. Ellinghaus, D. *et al*. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, https://doi.org/10.1186/1471-2105-9-18 (2008).
27. Xu, Z. *et al*. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
28. Ou, S. *et al*. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
29. Tarailo-Graovac, M. *et al*. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
30. Beier, S. *et al*. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585, https://doi.org/10.1093/bioinformatics/btx198 (2017).
31. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
32. Stanke, M. *et al*. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research* **32**, W309–W312, https://doi.org/10.1093/nar/gkh379 (2004).
33. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
34. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **44**, e89–e89, https://doi.org/10.1093/nar/gkw092 (2016).
35. Kim, D. *et al*. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
36. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
37. Tang, S. *et al*. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
38. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
39. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
40. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
41. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
42. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
43. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research* **43**, D213–221, https://doi.org/10.1093/nar/gku1243 (2015).
44. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–d314, https://doi.org/10.1093/nar/gky1085 (2019).
45. Lowe, T. M. *et al*. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
46. Loman, T. A novel method for predicting ribosomal RNA genes in prokaryotic genomes (2017).
47. Nawrocki, E. P. *et al*. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
48. Punta, M. *et al*. The Pfam protein families database. *Nucleic Acids Research* **40**, D290–301, https://doi.org/10.1093/nar/gkr1065 (2012).
49. She, R. *et al*. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Research* **19**, 143–149, https://doi.org/10.1101/gr.082081.108 (2009).
50. Birney, E. *et al*. GeneWise and Genomewise. *Genome Research* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
51. Partners, C.-N. M. a. Database resources of the national genomics data center, China national center for bioinformation in 2024. *Nucleic Acids Research* **52**, D18–d32, https://doi.org/10.1093/nar/gkad1078 (2024).
52. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA025251 (2025).
53. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA025199 (2025).
54. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA025200 (2025).
55. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA025201 (2025).
56. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBPSJW010000000 (2025).
57. *NGDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/93569/show (2025).
58. Lv, Z. Chromosome-level genome assembly of *Camellia pitardii*. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.29533658.v1 (2025).
59. Lv, Z. *et al*. The genome assembly of *Camellia pitardii*. *Zenodo* https://doi.org/10.5281/zenodo.15744655 (2025).
60. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).


## Acknowledgements

This work was sponsored by the Natural Science Foundation of Chongqing, China (CSTB2024NSCQ-MSX0886) and Chongqing Finance Bureau Special Project. We thank Associate Professor Bi Ma from State Key Laboratory of Resource Insects, Southwest University for his comments and suggestions on the manuscript. We thank Biomarker Technologies Co., Ltd for assisting in sequencing and bioinformatics analysis.

## Author contributions

Z.L. and L.A. designed the project and designed the experiments. S.Z. and L.A. supervised the work and revised the manuscript. Z.L., S.Z., J.W. and L.Z. collected the samples and performed the data analysis. Z.L., J.W. and L.Z wrote the manuscript. All authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05764-5.

**Correspondence** and requests for materials should be addressed to L.A.

**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.