scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A telomere-to-telomere genome assembly of *Camellia nitidissima*

Xin-Feng Wang[1,2,3,4,7], Tong-Jian Liu[1,2,3,4,7], Tian Feng[1,2,3,4,5,7], Hui-Run Huang[1,2,3,4], Pu Zou[4], Xiao Wei[6], Xing Wu[4] ✉, Sheng-Feng Chai[6] ✉ & Hai-Fei Yan[1,2,3,4] ✉

Camellia nitidissima **is the model species of the** ***Camellia*** **sect.** ***Chrysantha*** **Chang, the only lineage within the genus** ***Camellia*** **known to produce golden-yellow flowers. This species holds high aesthetic, germplasm and medical value. Unfortunately, due to excessive collection and habitat loss,** ***C. nitidissima*** **is classified as a critically endangered plant. In this study, we assembled a telomere-to-telomere (T2T) genome of** ***C. nitidissima*** **by incorporating PacBio HiFi and Hi-C data. The assembled genome consisted of 15 pseudo-chromosomes, with a total size estimated to be 2.72 Gb. The GC content and repetitive sequences occupied 38.05% and 84.38% of the assembled genome, respectively. In total, 35,701 protein-coding genes were annotated. Multiple evaluation methods confirmed the contiguity (contig N50: 81.74 Mb), completeness (BUSCOs: 98.80%) and high LTR Assembly Index (LAI: 14.57) of the genome. This high-quality T2T genome will provide valuable insights into the genomic characteristics of** ***C. nitidissima*** **and facilitate conservation efforts as well as functional genomic studies in** ***Camellia*** **sect.** ***Chrysantha*** **species.**

Background & Summary

Camellia sect. *Chrysantha* Chang, belonging to the family Theaceae and genus *Camellia*, is recognized for its striking golden-yellow flowers and is often referred to as the "Queen of the Tea Family" and the "Giant Panda of the Plant"[1–5]. Members of sect. *Chrysantha* possess considerable ornamental value, and some species also exhibit medicinal properties[1–5]. All members within this section are classified as rare and endangered plants, and are designated as Class II protected plants in China (https://www.gov.cn/zhengce/zhengceku/2021-09/09/content_5636409.htm). In recent years, the wild populations of sect. *Chrysantha* species have suffered extensive damage, highlighting the urgent need for conservation efforts[2,3,6,7]. Genomic studies on the genus *Camellia* have predominantly focused on sect. *Oleifera* Chang Tax., sect. *Thea* (L.) Dyer, sect.*Camellia* (L.) Dyer and sect. *Furfuracea* Chang Tax.[8–16]. To date, only one genome of *C. limonia* in sect. *Chrysantha* has been reported[17], while no genome data is publicly available at present. The limited availability of genomic data has hindered comprehensive investigation into the evolutionary and conservation genomics of the sect. *Chrysantha*.

Camellia nitidissima Chi, a member of *Camellia* sect. *Chrysantha*, is endemic to the southern region of Guangxi, China, and Northern Vietnam[2–4]. This species is characterized by its golden-yellow flowers, with significant ornamental value (Fig. 1). *C. nitidissima* is rich in bioactive compounds such as polyphenols, polysaccharides, flavonoids, saponins, and alkaloids, conferring extensive medicinal properties[1,3,5,18,19]. Pharmacological investigations have indicated that *C. nitidissima* possesses antioxidant, hypoglycemic, hypolipidemic, anti-tumor and other pharmacological activities[3,20–23]. However, the species faces threats from climate change, leading to habitat loss, as well as human activities such as excessive collection. In 2020, *C. nitidissima* was designated as a Critically Endangered (CR) species in China. Therefore, the generation of a high-quality genome of

[1]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [2]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [3]Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510650, China. [4]South China National Botanical Garden, Guangzhou, 510650, China. [5]University of Chinese Academy of Sciences, Beijing, 100049, China. [6]Guangxi Key Laboratory of Plant Functional Phytochemicals and Sustainable Utilization, Guangxi Institute of Botany, Guangxi Zhuang Autonomous Region and Chinese Academy of Sciences, Guilin, Guangxi, 541006, China. [7]These authors contributed equally: Xin-Feng Wang, Tong-Jian Liu, Tian Feng. ✉e-mail: wuxing@scbg.ac.cn; sfchai@163.com; yanhaifei@scbg.ac.cn

Fig. 1 Example photo captions for different blooming stages of *Camellia nitidissima*. (**a**) flower bud initiation stage, at the onset of bud development, newly emerging flower buds are green. (**b**) early flower bud stage, the bud transitions to a yellow hue. (**c**) flower bud enlargement stage, the bud continues to grow and enlarge, turning a golden yellow as it approaches anthesis. (**d**) early flowering stage, the petals gradually unfurl as the flower begins to open. (**e**) blooming stage, the petals are fully expanded, marking the full anthesis.

C. nitidissima is essential for promoting its conservation and utilization, as well as for clarifying the phylogenetic relationships and their evolutionary histories among species within the genus *Camellia*.

In this study, we successfully assembled and annotated the T2T genome of *C. nitidissima*, using PacBio HiFi reads, next-generation sequencing (NGS) reads, high-throughput chromosome conformation capture (Hi-C) reads, and RNA-seq reads. After assembly, closing gaps and assessment, we attained the T2T assembly of 2.72 Gb, containing 15 pseudo-chromosomes, with a contig N50 of 81.74 Mb, a scaffold N50 of 187.52 Mb, a BUSCO completeness score of 98.80%, and a LTR Assembly Index (LAI) of 14.57 (Fig. 2, Tables 1–3 and Supplementary Tables S1, S2). Additionally, we performed a synteny analysis with genomes from other *Camellia* species, and all pseudo-chromosomes exhibited highly continuous synteny (Supplementary Fig. S1). Both telomeres were recognized in 12 pseudo-chromosomes, while a single telomere was found in each of the remaining three pseudo-chromosomes (Table 2). We further predicted the centromere regions of all 15 pseudo-chromosome (Fig. 2 and Supplementary Table S3).

Annotation of repeat elements revealed that 84.38% (~2.30 Gb) of the *C. nitidissima* genome comprises repeat elements, with long terminal repeat (LTR) retrotransposon accounting for 51.14% (1.39 Gb) of the genome (Table 1 and Table 3). This is relatively high in *Camellia* genomes, which have a repeat content ranging from approximately 69.03% to 86.60%[8–16]. The extensive presence of LTRs in *C. nitidissima* suggests an active transposable element landscape, potentially contributing to genome expansion. Our analyses predicted a total of 35,701 protein-coding genes and 42,715 transcripts (Table 1). We performed functional annotation for all gene transcripts in the genome using multiple databases. The results showed that out of 42,715 proteins, 41,510 proteins (97.18%) were annotated at least once. Among them, 36,727 proteins (85.98%) were annotated by Gene Ontology (GO), and 21,460 proteins (50.24%) were annotated by KEGG Orthology (KO) (Table 1 and Supplementary Table S4). We further analyzed the genes potentially affected by LTRs and their functions (Supplementary Table S5 and Fig. S2). Based on whether LTRs are located within 1 Kb or 5 Kb upstream or downstream of the genes, we found that more than half of the genes were influenced by LTRs. These genes exhibit diverse functions, suggesting potential roles in species growth and adaptation (Supplementary Table S5 and Fig. S2). Overall, the characterization of the genome revealed previously unknown features of the *C. nitidissima* genome (Fig. 2, Tables 1–3, Supplementary Fig. S1 and Tables S1–S4). The assembly demonstrates important improvements in contiguity and completeness, enabling more accurate analyses of genomic structure and function. The high-quality genome assembly not only provides a valuable resource for conservation genomics but also facilitates deeper insights into the evolutionary dynamics of *Camellia* species.

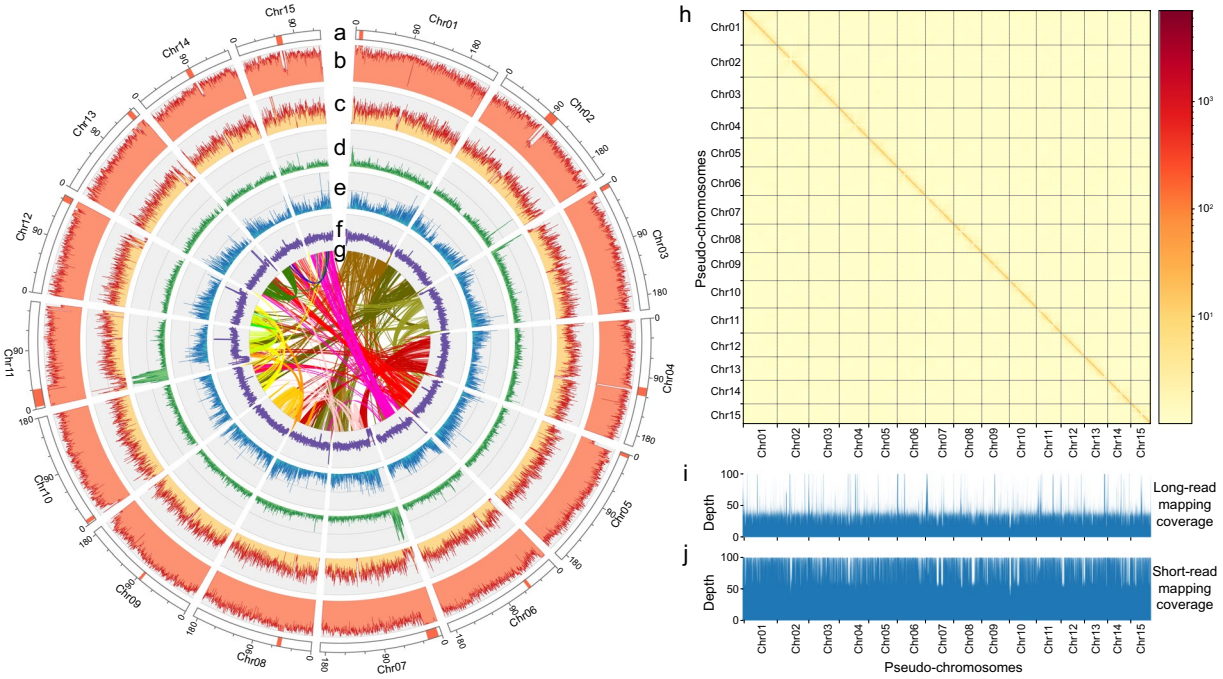

Fig. 2 Features of the *Camellia nitidissima* genome. (**a**–**g**) Circular tracks represent, from outer to inner, 15 chromosomal-level scaffolds (Chr01-Chr15), percentage of repeats (0%–100%), *Gypsy* (0%–99.96%), *Copia* (0%–99.80%), gene density (0–26), GC content (23.39%–61.69%) and the spectrum of collinear analysis (each line connects one pair of homologous genes and a cluster of such lines represents one collinear block). All statistics are calculated in 200-Kb windows. (**h**) The Hi-C heatmap of the *Camellia nitidissima* genome. Each box represents a pseudo-chromosome. (**i**) Long-read mapping coverage of all pseudo-chromosomes. (**j**) Short-read mapping coverage of all pseudo-chromosomes.

Methods

Sample collection and sequencing. The *C. nitidissima* sample was collected from Nawan Village (N21.7908, E108.1921), Fangcheng District, Fangchenggang City, Guangxi Zhuang Autonomous Region, China. The sampled individual of *C. nitidissima* was mature wild plant. Tissue samples, including fresh young leaves and flowers, were immediately frozen in liquid nitrogen after collection and preserved at −80 °C for further DNA and RNA extraction.

The frozen samples were sent to Novogene Co., Ltd. (Tianjin, China) for genomic DNA and RNA extraction and sequencing. Specifically, total DNA was extracted using a plant genomic DNA extraction kit (Novogene) and assessed via 0.75% gel electrophoresis experiment on an Agilent 4200 TapeStation system. The main band of the DNA was observed to be greater than 30 kb, indicating high completeness. DNA purity and concentration were evaluated using a NanoDrop One UV–Vis spectrophotometer and a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), respectively.

We utilized Single Molecule Real-Time (SMRT) sequencing technology to generate highly accurate long-read sequencing data (HiFi reads) based on the advanced PacBio Revio sequencing platform. A PCR-free SMRT bell library was prepared, and HiFi reads were generated using Circular Consensus Sequencing (CCS) mode. This sequencing approach yielded consensus sequences from multiple subreads of the same DNA molecule with quality scores exceeding 99%. A total of 90.67 Gb of data (~33.3 × coverage) was obtained from 4,895,607 high accuracy HiFi reads, with a mean read length of 18,497 bp.

For next-generation sequencing (NGS), a short-read library (2 × 150 bp) with an insert size of 300–500 bp was prepared using the TruSeq Nano DNA HT Sample Preparation Kit. Sequencing was conducted on an Illumina Novaseq system, producing 730,591,634 paired reads with 219.18 Gb (~80.6 × coverage) of raw data.

To achieve a chromosome-level assembly, we utilized high-throughput chromosome conformation capture (Hi-C) technology to generate interaction signals among adjacent sequences, facilitating the scaffolding of contigs into chromosomes. The Hi-C library was constructed from cross-linked DNA after digestion, biotinylation, ligation, enrichment, shearing, blunt-end repair. The size and the integrity of the insert fragments in the Hi-C library was subsequently evaluated using an Agilent 2100 Bioanalyzer. Additionally, quantitative PCR (qPCR) was employed to precisely quantify the effective concentration of the library, with a threshold of >2 nM considered acceptable. Sequencing of the qualified Hi-C library, with pair-end read lengths of 150 bp, was conducted on the Illumina Novaseq 6000 platform, resulting in a total of 151.67 Gb (~55.8 × coverage) of Hi-C data.

For transcriptome sequencing, three pair-end RNA libraries (2 × 150 bp) were prepared according to the TruSeq RNA Library Preparation Kit instructions and sequenced on the Illumina Novaseq 6000 platform, yielding 3.54 Gb, 3.65 Gb and 11.78 Gb of RNA-seq data for the leaf and flower samples, respectively (Table 1). All sequencing was conducted at Novogene Co., Ltd. (Tianjin, China). For NGS, Hi-C, and RNA-seq data, adapters and low-quality reads were removed using fastp v0.23.3[24].

Category		*Camellia nitidissima*
Assembly	Assembled contigs	561
	Contigs (≥50000 bp)	247
	Total contig size (bp)	2,834,860,250
	Largest contig (bp)	226,869,717
	Contig N50 (bp)	81,738,692
	Contig N90 (bp)	35,337,314
	Contig L50	13
	Contig L90	33
	GC content (%)	38.31
	Total scaffolding length (bp)	2,834,865,050
	Number of scaffolds	572
	Scaffold N50 (bp)	187,518,794
	Cumulative length of chromosomes (bp)	2,720,093,003
	Number of chromosomes	15
	Longest chromosome (bp)	226,869,817
	Shortest chromosome (bp)	130,101,769
	GC content (%)	38.05
Assessment	Completeness BUSCOs (%)	98.80
	LAI (LTR Assembly Index)	14.57
	Illumina mapping rate (%)	98.56
	Illumina breadth of coverage (%)	99.99
	Illumina average depth (rmdup)	77.57
	Long reads mapping rate (%)	99.83
	Long reads breadth of coverage (%)	99.99
	Long reads average depth (rmdup)	32.64
Annotation	Repeat contents (%)	84.38
	Number of genes	35,701
	Number of transcripts	42,715
	Average coding length (bp)	1,210.3
	Average number of exons per gene	5.23
	Number of genes with GO annotation	36,727
	Number of genes with KEGG annotation	21,460

Table 1. Summary of assembly, assessment and annotation information for the *Camellia nitidissima* genome.

Estimation of genome profiles. The k-mer frequency distribution was calculated using Jellyfish v2.3.0 (-m 21)[25] with the NGS data. The resulting file was utilized to predict genomic features using GenomeScope v2.0[26], with k-mer and read length set at 21 and 150, respectively. The ploidy level and max_kmercov were set to 2 and 1×10^7. Based on this analysis, the genome of *C. nitidissima* was estimated to be 2,632.29 Mb, with a heterozygosity rate of 1.12% (Model Fit: 94.96%).

Chromosome-level genome assembly. The genome was assembled using Hifiasm v0.19.9[27] with default parameters, which leverages long-read sequencing data (i.e. Pacbio HiFi reads) and Hi-C data to produce high-quality single-sample telomere-to-telomere assemblies. The assembled draft genome was 2834.86 Mb, consisting of 561 contigs with a contig N50 length of 81.74 Mb (Table 1). Hi-C reads were subsequently aligned to the draft genome, identifying 165,831,057 valid interaction paired signals using HiC-Pro v3.1.0 pipeline[28]. The aligned BAM file was sorted and the PCR duplicates were marked using biobambam[29]. The draft genome assembly was scaffold into chromosomes based on the Hi-C interaction information using Yahs[30] with default parameters, followed by manual adjustments made in Juicebox v1.11.08[31]. Gaps in the genome assembly were processed using QuarTeT v1.2.1[32], utilizing draft assembled contigs (including primary contigs and both haplotype contigs). A total of 14 gaps were closed, resulting in a filled length of 1,130,719 bp; after gap-filling, 25 gaps still remained. Telomeres of each pseudo-chromosome were identified using the TeloExplorer function in QuarTeT v1.2.1[32]. The detection of continuous telomeric repeat (AAACCCT) structures within 10,000 base pairs at both ends of a pseudo-chromosome is considered to indicate the presence of telomeres (Table 2). Centromeres of each pseudo-chromosome were identified using the Centromics v0.4 (https://github.com/ShuaiNIEgithub/Centromics; Supplementary Table S3).

The final genome assembly was 2.83 Gb in length and consisted of 572 scaffolds, with a scaffold N50 length of 187.52 Mb (Table 1). A total of 95.95% (2.72 Gb, 55 scaffolds) of the sequences were anchored to 15 pseudo-chromosomes. The lengths of individual chromosomes ranged from 130.1 Mb to 226.87 Mb (Table 1). The circular plot of chromosomes was visualized by Circos v0.69-8[33] (Fig. 2). GC content of the genome was

Chr id	Chr length (bp)	Status	Left num	Left direction	Right num	Right direction
Chr01	226,869,817	right	0	/	147	–
Chr02	211,477,848	both	519	+	332	–
Chr03	203,026,860	both	658	+	282	–
Chr04	197,457,834	both	444	+	317	–
Chr05	190,254,444	both	704	+	44	–
Chr06	188,766,305	right	0	/	54	–
Chr07	188,465,104	both	95	+	26	–
Chr08	187,518,794	both	374	+	59	–
Chr09	184,644,962	both	1099	+	658	–
Chr10	180,037,700	both	47	—	270	—
Chr11	165,004,885	right	0	/	16	+
Chr12	156,687,462	both	27	+	39	–
Chr13	155,067,501	both	168	+	11	–
Chr14	154,711,718	both	697	—	543	—
Chr15	130,101,769	both	389	+	14	–

Table 2. Summary of telomere information for the *Camellia nitidissima* genome.

Class		Count	Total Length (bp)	Pro. of genome
LTR	Caulimovirus	23,694	37,059,367	1.36%
	Copia	293,223	243,713,619	8.96%
	ERV1	4,107	742,458	0.03%
	ERVK	2,974	2,388,162	0.09%
	Gypsy	882,066	1,091,978,273	40.14%
	Ngaro	2,196	420,568	0.02%
	Unclassified	28,635	14,650,551	0.54%
	Total	1,236,895	1,390,952,998	51.14%
LINE	L1	108,148	56,462,538	2.08%
	RTE-BovB	51,948	13,921,059	0.51%
	Unclassified	17,239	5,194,770	0.19%
	Total	177,335	75,578,367	2.78%
DNA	CMC-EnSpm	25,219	17,762,648	0.65%
	Crypton-H	860	316,214	0.01%
	hAT-Ac	13,578	8,767,452	0.32%
	hAT-Tag1	6,347	2,751,366	0.10%
	hAT-Tip100	24,218	7,619,077	0.28%
	IS3EU	989	607,553	0.02%
	Merlin	2,185	1,636,605	0.06%
	MULE-MuDR	92,106	68,816,529	2.53%
	P	773	392,028	0.01%
	PIF-Harbinger	22,582	12,209,824	0.45%
	PIF-ISL2EU	1,443	199,586	0.01%
	TcMar-Pogo	2,297	3,983,768	0.15%
	TcMar-Tc1	1,835	687,945	0.03%
	Zisupton	8,666	1,293,549	0.05%
	Unclassified	4,415	1,101,547	0.04%
	Total	207,513	128,145,691	4.71%
Rolling Circle	Helitron	72,356	22,326,595	0.82%
Other	Low Complexity	84,998	4,635,620	0.17%
	Satellite	48,503	86,205,773	3.17%
	Simple Repeat	597,466	22,688,389	0.83%
	Total	730,967	113,529,782	4.17%
Unknown		1,549,239	561,734,951	20.65%

Table 3. Summary of repetitive regions of the *Camellia nitidissima* genome.

calculated using bedtools v2.31.1[34], revealing that individual chromosomes had GC contents ranging from 36.60% to 38.88%, with an overall mean of 38.05% (Fig. 2).

Repeat and gene annotation. The transposable element (TE) annotation pipeline EarlGrey v4.4.5[35] was employed to identify multiple repeat elements in the C. *nitidissima* genome. EarlGrey generated a non-redundant transposable element (TE) library using RepeatModeler, and subsequently utilized RepeatMasker to obtain TE annotation results for the C. *nitidissima* genome based on this TE library. Meanwhile, EarlGrey employed LTR_FINDER to identify full-length LTR retrotransposons with intact structures.

A total of 3,974,305 repetitive sequences were identified, representing 84.27% (~2.29 Gb) of the genome (Table 3). Among these repeats, LTRs were the most prevalent, constituting 51.14% (1.39 Gb) of the genome (Table 3). *Gypsy* elements (40.14%) were the predominant LTRs, followed by *Copia* elements (8.96%). The total lengths of the non-LTR elements LINE and DNA transposons were 75.58 Mb and 150.47 Mb, respectively. These elements accounted for 2.78% and 5.53% of the genome (Table 3).

The transposable element (TE) library generated by EarlGrey can facilitate the production of a soft-masked genome. Gene prediction and functional annotation for the soft-masked genome were performed using *de novo*, homology protein-based, and transcriptome-based methods via the Braker3 v3.0.8[36] pipeline with default parameters. RNA-seq data were mapped to the *C. nitidissima* genome using HISAT2[37] spliced aligner to generate sorted bam files. These RNA-Seq alignments, along with curated proteins from OrthoDB v10 database, were utilized to train gene prediction models using GeneMark-ETP[38] and Augustus[39]. Subsequently, these data were employed to achieve reliable predictions of protein-coding genes, with TSEBRA[40] producing a combination results. For functional annotation, the predicted genes were queried against public databases, using the InterProScan v5.62–94.0[41], EggNOG-mapper v2.1.11[42], PANNZER2[43], and Mercator4[44] pipelines.

In total, 35,701 genes encoding 42,715 proteins were predicted, distributed across the genome with an average length of 76,190.95 bp per gene, with a mean coding length of 1210.3 bp (Table 1). On average, each gene consists of 5.23 exons (Table 1). Among the 35,701 protein-coding genes, 35,201 (98.60%) were functionally annotated using the online EggNOG-mapper pipeline (http://eggnog-mapper.embl.de/), of which 16,793 genes (accounting for 47.03%) were classified into 11,492 GO terms, and 16,212 genes (occupying of 45.40%) were classified into 4,079 KO terms (Table 1). The collinear gene pairs and blocks were identified using MCScanX -master[45] (Fig. 2).

Data Records

The raw data, including Illumina short reads, HiFi reads, Hi-C reads, and RNA short reads, has been deposited to the Genome Sequence Archive (GSA) in the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB)[46,47] with the accession number of CRA023969[48] under BioProject PRJCA034778. The final genome assembly, annotation, and protein-coding sequences are accessible via Figshare[49] and have also been uploaded to the Genome Assembly Sequences and Annotations (GWH) in NGDC with the accession number of SAMC4541821[50] under the BioProject PRJCA034778[46,51]. The genome assembly has also been submitted to the National Center for Biotechnology Information (NCBI) with the accession number of JBMIOJ000000000[52] under BioProject PRJNA1240963.

Technical Validation

Multiple methods were used to evaluate the quality of the *C. nitidissima* genome. First, the Hi-C data visualized by HiCExplorer[53] showed high consistency across all 15 pseudo-chromosomes, confirming the accuracy of the ordering and orientation (Fig. 2). Second, the Pacbio HiFi reads and Illumina NGS reads were mapped to the assembly using minimap2[54,55] with the parameters "-ax map-hifi" and "-ax sr", respectively. The mapping rate, mean depth, and breadth of coverage was calculated using SAMtools v2.5.1[56] with custom scripts. A total of 99.83% of the HiFi reads and 98.56% of the NGS reads were successfully aligned (Table 1). The breadth of alignment coverage of these two datasets across the genome reached as high as 99.99% (Table 1). Third, BUSCO v5.6.1[57] was used for the evaluation of assembly completeness of the genome with the embryophyta_odb10 database. The complete BUSCOs for *C. nitidissima* were 98.80% (Table 1). Additionally, the genome quality was evaluated by calculating the LTR assembly index (LAI) using the LTR_retriever v3.0.1[58,59], yielding a LAI value of 14.57 (Table 1). Overall, these results demonstrated the high quality, accuracy, and reliability of the *C. nitidissima* reference genome.

Code availability

All software and programs used in this study were described and cited in the Methods section. If no detailed parameters were mentioned, default parameters were used.

Received: 22 January 2025; Accepted: 8 May 2025;
Published online: 18 May 2025

References

1. Liang, S.-Y. *Yellow Camellias*. (Beijing: Chinese Forestry Press, 1993).
2. Chai, S. *et al*. Eco-physiological basis of shade adaptation of Camellia nitidissima, a rare and endangered forest understory plant of Southeast Asia. *BMC Ecol.* **18**, 5 (2018).
3. LIU, Q. *et al*. Yellow Camellia: Resource Status and Research Progress in Modern Studies. *Mod. Chinese Med.* **23**, 727–733 (2021).
4. Hung-Ta, C. & Shan-Xiang, R. *Flora reipublicae popularis sinicae, vol. 49. 3rd ed. Beijing*: *Science Press* (1998).
5. Jiang, L. *et al*. Elucidation of the key pathway for flavonol biosynthesis in golden Camellia and its application in genetic modification of tomato fruit metabolism. *Hortic. Res*. uhae308 https://doi.org/10.1093/hr/uhae308 (2024).
6. Xiao, W. *et al*. Seed reproduction and biological characteristics of Camellia nitidissima. *Guihaia* **30**, 215–219 (2010).

7. Xiao, W., Shui-yuan, J., Yun-sheng, J., Hui, T. & Hong-lin, C. Research Progress of Camellia nitidssiam,a Rare and Endangered Plant. *J. Fujian For. Sci. Technol.* **33**, 169–174 (2006).
8. Chen, S. *et al*. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis. *Nat. Plants* **9**, 1986–1999 (2023).
9. Gong, W. *et al*. Chromosome-level genome of Camellia lanceoleosa provides a valuable resource for understanding genome evolution and self-incompatibility. *Plant J.* **110**, 881–898 (2022).
10. Lin, P. *et al*. The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. *Genome Biol.* **23**, 14 (2022).
11. Shen, T. F. *et al*. The reference genome of Camellia chekiangoleosa provides insights into Camellia evolution and tea oil biosynthesis. *Hortic. Res.* **9**, uhab083 (2022).
12. Wei, C. *et al*. Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. USA.* **115**, E4151–E4158 (2018).
13. Xia, E. *et al*. The Reference Genome of Tea Plant and Resequencing of 81 Diverse Accessions Provide Insights into Its Genome Evolution and Adaptation. *Mol. Plant* **13**, 1013–1026 (2020).
14. Zhang, W. *et al*. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
15. Zhang, X. *et al*. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. *Nat. Genet.* **53**, 1250–1259 (2021).
16. Zhang, F., Feng, L. Y., Lin, P. F., Jia, J. J. & Gao, L. Z. Chromosome-scale genome assembly of oil-tea tree Camellia crapnelliana. *Sci. Data* **11**, 1–8 (2024).
17. Lu, Y. *et al*. Chromosome-scale assembly and analysis of yellow Camellia (Camellia limonia) genome reveal plant adaptation mechanism and flavonoid biosynthesis in karst region. *Glob. Ecol. Conserv.* **56**, e03296 (2024).
18. Lin, J. N. *et al*. Chemical constituents and anticancer activity of yellow camellias against MDA-MB-231 human breast cancer cells. *J. Agric. Food Chem.* **61**, 9638–9644 (2013).
19. Li, X. L. *et al*. Flavonoid components and their relationship with flower colors in five species of Camellia section Chrysantha. *Chinese J. Ecol.* **38**, 961–966 (2019).
20. Kong, G., Du, H., Yuan, S. & Sun, L. Study Effect of Extrative Fraction of Camellia chrysantha(Hu) Tuyama from n-butylalcohol on Lung Carcinogenesis Induced by Urethane. *Asia-Pacific Tradit. Med.* **11**, 4–7 (2015).
21. CHENG, J. *et al*. *In vitro* Antioxidant Experiment Research of Total Saponins in the Flower of Camellia nitidissima Chi. *Chinese J. Ethnomedicine Ethnopharmacy* **25**, 27–30 (2016).
22. XIA, X. *et al*. Effect of Camellia nitidissima Extract on Pancreatic Function in Diabetes Mice. *LISHIZHEN Med. MATERIA MEDICA RESEARCH* **24**, 2863–2865 (2013).
23. Wei, L., QIN, X., LIN, H., NING, E. & YANG, H. Study on the hypolipidemia activity of polysaccharides from the leaves of Camellia chrysantha (Hu) Tuyama. *Food Sci. Technol.* **201**, 247–249 (2008).
24. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
25. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
26. Vurture, G. W. *et al*. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
27. 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).
28. Servant, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
29. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).
30. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
31. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
32. Lin, Y. *et al*. QuarTeT: A telomere-To-Telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
33. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
34. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
35. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Mol. Biol. Evol.* **41**, msae068 (2024).
36. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
37. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
38. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data. *Genome Res.* **34**, 757–768 (2024).
39. Stanke, M. & Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
40. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
41. Jones, P. *et al*. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
42. Huerta-Cepas, J. *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
43. Törönen, P., Medlar, A. & Holm, L. PANNZER2: A rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).
44. Schwacke, R. *et al*. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. *Mol. Plant* **12**, 879–892 (2019).
45. Wang, Y. *et al*. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
46. Bao, Y. *et al*. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* **52**, D18–D32 (2024).
47. Chen, T. *et al*. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics, Proteomics Bioinforma.* **19**, 578–583 (2021).
48. *Genome Sequence Archive (GSA)*. https://ngdc.cncb.ac.cn/gsa/browse/CRA023969 (2025).
49. Wang, X.-F. Camellia nitidissima genome. *figshare*. https://doi.org/10.6084/m9.figshare.28202408.v1 (2025).
50. Wang, X.-F. Camellia nitidissima Genome. *Genome Warehouse (GWH)*. https://download.cncb.ac.cn/gwh/Plants/Camellia_nitidissima_Camellia_nitidissima_GWHFILD00000000.1 (2025).
51. Chen, M. *et al*. Genome Warehouse: A Public Repository Housing Genome-scale Data. *Genomics, Proteomics Bioinforma.* **19**, 584–589 (2021).

52. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_049201075.1 (2025).
53. Wolff, J., Backofen, R. & Grüning, B. Loop detection using Hi-C data with HiCExplorer. *Gigascience* **11**, giac061 (2022).
54. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
55. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
57. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
58. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
59. 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 Guangdong Flagship Project of Basic and Applied Basic Research (2023B0303050001), the Guangdong Science and Technology Plan Project (2023B1212060046) and the National Natural Science Foundation of China (32460103).

Author contributions

H.F.Y., S.F.C. and X.W. designed and supervised the research; X.F.W., T.J.L. and H.F.Y. wrote the manuscript; X.F.W., T.J.L., T.F., H.R.H., P.Z. and X.W. analysed the data; H.F.Y. and T.J.L. collected the experimental materials. All authors contributed to the manuscript revision and approved the submitted version.

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/s41597-025-05157-8.

Correspondence and requests for materials should be addressed to X.W., S.-F.C. or H.-F.Y.