scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome level genome assembly of *Camellia sinensis* 'Yuwan Xiaoye'

Wei Zhang[1,2] & Yuqiong Chen[1] ✉

Tea is one of the oldest crops in the world and is of great economic value as a beverage. Its natural flavor and compounds related to health exhibit significant genetic diversity. This article reports the chromosome-scale genome assembly map of a new *Camellia sinensis* 'Yuwan Xiaoye' (YWXY) bred by our research center. The genome size is 3.18 Gb with a contig N50 of 181.8 Mb, and 93.43% of the assembled sequences were anchored to 15 chromosomes. The genome is predicted to contain 40,119 protein-coding genes, with 99.70% having functional annotations. Repeat elements account for approximately 82.21% of the genomic landscape. The completeness of YWXY genome assembly is highlighted by a BUSCO score of 99.07%. The assembled genome provides a critical resource for molecular breeding and functional studies in tea plants.

Background & Summary

Tea (*Camellia sinensis*, 2n = 30) is one of the most important economic crops worldwide, particularly in developing countries across Asia, Africa, and Latin America[1,2]. As the most widely consumed non-alcoholic beverage globally, tea is renowned for its significant health benefits attributed to characteristic compounds such as catechins, theanine, and caffeine[3]. From a genomic perspective, the tea plant exhibits distinctive features including a large genome size, high heterozygosity, and remarkable species diversity[4]. In recent years, with the rapid development of high-throughput sequencing technologies and bioinformatic methodologies, groundbreaking progress has been made in tea plant genomics research. To date, researchers have successfully assembled several chromosome-level tea plant genomes[5–8], constructed a comprehensive pan-genome[9,10], and conducted resequencing and transcriptome sequencing studies on hundreds of different genotypes[11–16]. These genomic datasets have provided crucial insights into the evolutionary history of tea plants, revealing frequent intra- and interspecific gene introgression events. Furthermore, these studies have established a solid foundation for elucidating the genetic basis and molecular mechanisms underlying key agronomic traits in tea plants.

The tea cultivar *Camellia sinensis* 'Yuwan Xiaoye' (YWXY) was developed through systematic breeding programs, demonstrating superior agronomic characteristics including early maturation, high yield, and exceptional stress tolerance (Fig. 1a). These outstanding traits reflect unique genetic adaptations of the area north of the Yangtze River. Currently, YWXY has been successfully cultivated more than 667 hectares, validating the practical agricultural value of its distinctive genomic features. The emergence of high-quality genome assemblies has significantly accelerated the process of plant breeding and provided new tools for molecular design in *Camellia sinensis*. By leveraging genomic information, researchers can now identify and select desirable traits with unprecedented precision and efficiency. Consequently, we anticipate rapid advancements in tea plant genetic research as functional genes associated with key agronomic traits continue to be discovered. The elucidation of genomic structures and gene functions will facilitate in-depth studies on the genetic basis of economically important traits in tea plants. This progress will ultimately contribute to the conservation of tea genetic resources, genetic selection, and artificial breeding programs.

In this study, we generated a high-quality chromosome-scale genome assembly of *C. sinensis* using PacBio HiFi long-read sequencing with Hi-C chromatin interaction data (Table 1). The estimated genome size was 3,072 Mb (Fig. 1b; Table 2), while the final assembly reached 3,175 Mb with a contig N50 of 181.8 Mb (Table 2). Of the assembled sequences, 2,966.67 Mb (93.43%) were anchored to 15 chromosomes, representing a complete

[1]National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, 430070, China. [2]Dabie Mountain Laboratory, College of Life Sciences, Xinyang Normal University, Xinyang, 464000, China. ✉e-mail: chenyq@mail.hzau.edu.cn

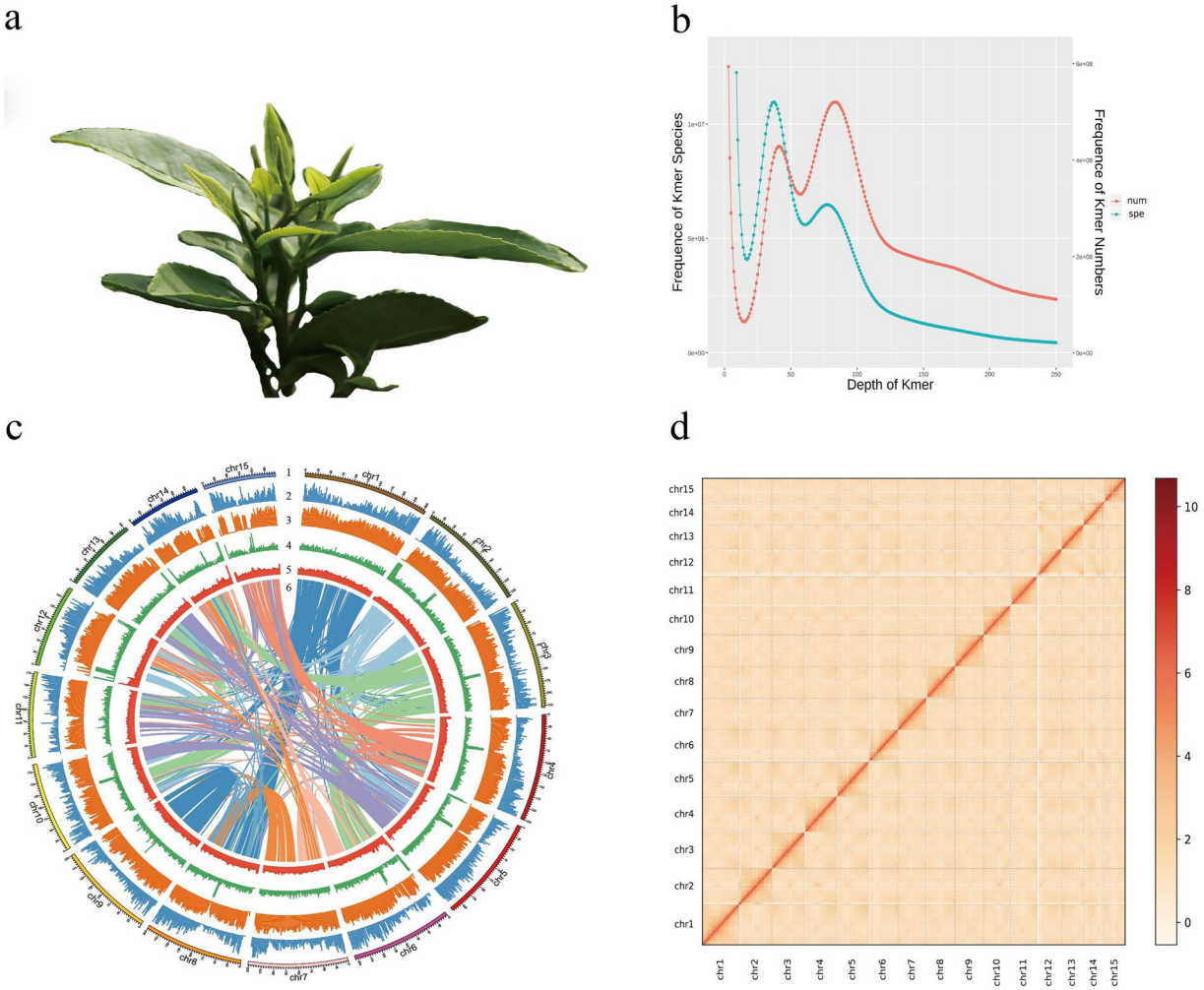

Fig. 1 Characterization of assembled *C. sinensis* genome. (**a**) The picture of the *C. sinensis* from XinYang (HeNan province, China). (**b**) K-mer (17-mer) frequency distribution curve. The x-axis represents k-mer depth (sequencing coverage), and the y-axis represents two metrics: spe (number of distinct k-mer species at each depth) and num (total count of all k-mers occurring at each depth). (**c**) Circos plot of the *C. sinensis* genome, with visualization of chromosome (1) gene density (2) TE repeat (3) TRF repeat (4) GC content (5) genomic synteny (6), in order from outside to inside. (**d**) Hi-C interactions between the 15 chromosomes of *C. sinensis*. The color scale bar, ranging from light yellow to dark red, represents the frequency of Hi-C interaction links from low to high.

Libraries	Insert sizes	Clean data (Gb)	Sequencing coverage (×)
MGI	300 bp	254.47	80.14×
PacBio	10–15 kb	210.12	66.18×
Hi-C	300 bp	551.11	173.57×
RNA	300 bp	12.74	—

Table 1. Statistics of the sequencing data used for genome assembly.

and contiguous reference genome (Table 2). To assess assembly quality and completeness, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis, which revealed a high completeness rate of 99.07% for the *C. sinensis* genome, confirming its high quality (Table 2). Gene annotation predicted 40,119 protein-coding genes (Table 2), along with extensive repetitive sequences spanning 2,610.37 Mb (82.21%) of the genome (Fig. 1c; Table 3). Additionally, we identified 3,563 non-coding RNA (ncRNA) genes, including 2,367 rRNAs, and 1,196 other ncRNAs (Table 4).

We anticipate that this high-quality genome assembly will make significant contributions to improving tea plant traits, accelerating molecular breeding processes, and enhancing our understanding of the evolutionary history within the *C. sinensis*.

Genome assembly	
Estimated the genome size	3,071,996,588
Total length of assembly (bp)	3,175,206,952
Number of scaffolds	591
scaffold N50 (bp)	200,885,500
Contig N50 (bp)	181,816,000
Chromosome number	15
Number of gaps	9
Anchor rate (%)	93.43%
GC content (%)	38.78%
BUSCO (%)	99.07%
single-copy BUSCOs (%)	85.87%
duplicated BUSCOs (%)	13.20%
fragmented BUSCOs (%)	0.43%
missing BUSCOs (%)	0.50%
Genome annotation	
Total length of repeats (bp)	2,610,367,426
Repeats percentage of assembly (%)	82.21%
Number of protein-coding genes	40,119
Average gene length (bp)	7,209.91
Average CDS length(bp)	1,179.63
Average intron length (bp)	1,367.95
Average exon length (bp)	272.88
BUSCO (%)	98.1%
Functional annotation	
GO	28,701 (71.54%)
KEGG_KO	15,137 (37.73%)
InterPro	29,737 (74.12%)
Swissprot	30,350 (75.65%)
KEGG_ALL	39,551 (98.58%)
TrEMBL	39,973 (99.64%)
NR	39,987 (99.67%)
Annotated	39,999 (99.70%)

Table 2. Statistics of the *C. sinensis* genome assembly and annotation.

Repeat Types	Length (bp)	Percent (%)
DNA transposons	281,969,115	8.88
LINEs	120,216,109	3.79
SINEs	2,192,069	0.07
LTRs	1,449,849,645	45.66
Unknown	1,118,998,530	35.24
Total TEs	2,501,252,794	78.77
Total Repeats	2,610,367,426	82.21

Table 3. Repetitive elements and their proportions in *C. sinensis* genome. Note: LINEs mean long interspersed nuclear elements; SINEs mean short interspersed nuclear elements; LTRs mean long terminal repeat retrotransposons.

Methods

Plant materials, sample collection, and sequencing. Fresh young leaves (one bud and two tender leaves) of *C. sinensis* were collected from the tea germplasm resource garden of Henan Jiuhuashan Tea Co., Ltd, Xinyang City, Henan Province, China (32°20′ N, 116°21′ E, altitude 400 m) in April 2025 during the spring tea harvesting season. The tea plants were 5 years old, grown under standard cultivation practices without water or nutrient stress, and were in healthy condition with no visible disease or pest damage. Fresh leaves were sampled, immediately frozen in liquid nitrogen, and stored at −80 °C until DNA/RNA extraction.

High-quality genomic DNA was extracted using the CTAB method[17]. DNA quality was verified by 0.8% agarose gel electrophoresis and absorbance measurements (NanoDrop Spectrophotometer, Thermo Fisher Scientific, USA), followed by quantification using Qubit Fluorometer (Invitrogen, USA). Whole-genome

	Type	Copy	AverageLength (bp)	TotalLength (bp)	% of genome
miRNA		124	124.04	15,381	0.0005
tRNA		801	73.58	58,936	0.0019
rRNA	rRNA	2,367	4,082.85	9,664,098	0.3044
	18S	1,191	1,718.85	2,047,155	0.0645
	28S	1,176	6,476.99	7,616,943	0.2399
	5S	0	0	0	0
snRNA	snRNA	271	125.29	33,954	0.0011
	CD-box	107	106.9	11,438	0.0004
	HACA-box	21	128.48	2,698	0.0001
	splicing	143	138.59	19,818	0.0006
	scaRNA	0	0	0	0

Table 4. Statistics of non-coding RNA annotation.

sequencing was performed by Wuhan GeneRead Biotechnology Co., Ltd, including PacBio HiFi long-read sequencing, Hi-C chromatin interaction sequencing and MGI hort-read sequencing.

Raw sequencing data were processed using quality control tools: CCS (v6.0.0, -min-passes 3 -min-snr2.5 -top-passes 60) for PacBio SMRT sequencing[18], Trimmomatic (v0.40, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50) for Hi-C read processing[19], and SOAPnuke (v2.1.0, --lowQual 20 --qualRate 0.5 --nRate 0.005) for short-read trimming[20]. For MGI sequencing, genomic DNA was randomly fragmented into 300–500 bp segments, and paired-end genomic libraries were prepared following the manufacturer's protocol. The libraries were then sequenced on the MGI DNBSEQ-T7 platform using a 150 bp paired-end configuration for genome survey, we obtained 254.47 Gb (80.14 × coverage) of MGI paired-end reads (Table 1). For PacBio long-read sequencing, SMRTbell libraries were constructed from the genomic DNA and sequenced on the PacBio Revio platform, and generated approximately 210.12 Gb of PacBio long-read data (66.18 × coverage) (Table 1).To generate chromosomal-level assembly of the *C. sinensis* genome, Hi-C library was constructed from the leaf of the plant. After crosslinking, cell wall lysis, chromatin digestion, biotin labelling, proximal chromatin DNA ligation,DNA purification and sequencing. Finally, we obtained 551.11 Gb of Hi-C data, providing 173.57 × coverage of the genome (Table 1).To assist genome annotation, total RNA was extracted from the above sample. Purity and integrity were assessed using NanoDrop spectrophotometer (Thermo Fisher Scientific, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, USA). RNA libraries were constructed using the VAHTS Universal V10 RNA-seq Library Prep Kit for MGI (NRM606) following the manufacturer's protocol and sequenced on MGI DNBSEQ-T7 platform, yielding a total of 12.74 Gb data (Table 1).

Chromosome-level genome assembly. First, the raw sequencing data from the MGI DNBSEQ-T7 platform underwent quality control using SOAPnuke (v2.1.0, --lowQual 20 --qualRate 0.5 --nRate 0.005)[20], and the filtered data were used to estimate genome size. K-mer analysis was performed using GCE (v1.0.2)[21] with a k-mer size of 17. The results estimated the genome size to be 3,072 Mb (Fig. 1b; Table 2). The PacBio HiFi reads were *de novo* assembled by using hifiasm (v0.19.5, Default parameters)[22]. Based on these sequencing data, the genome assembly had a total size of 3.18 Gb, with contig N50 of 181.82 Mb (Table 2). Chromosome-level genome scaffolding was performed using Hi-C technology. Hi-C reads were aligned to the *C. sinensis* contigs using BWA (v0.7.17, -SP5M) with default parameters. Read pairs mapped to different contigs were used for Hi-C-based scaffolding. Self-ligation, non-ligation, and other invalid reads were filtered out. 3D-DNA pipeline (https://github.com/aidenlab/3d-dna, -r 2)[23] was then employed to order and orient the contigs[24]. Ultimately, 9 gaps remained across the 15 assembled chromosomes.

Repeat annotation. We annotated the repetitive sequences in the genome using de novo prediction and homology-based approaches. Using the RepeatModeler[25] pipeline, we integrated multiple tools including RECON (v1.08)[26], RepeatScout (v1.0.5)[27], LTR-FINDER (v1.0.5)[28], and TRF (v4.0.935)[29] to detect and classify repetitive elements in the genome assembly. For homology-based annotation, RepeatMasker (open-4.0.9) and RepeatProteinMask (open-4.0.9) were used to screen against the Repbase library[30,31], respectively, to identify known transposable elements (TEs) in *C. sinensis* genome. The results revealed that repetitive sequences spanned 2,610.37 Mb, accounting for 82.21% of the assembled genome. Of these, 78.77% were TEs, including SINEs (0.07%), LINEs (3.79%), LTRs (45.66%), and DNA transposons (8.88%) (Fig. 1c; Table 3).

Protein-coding gene prediction and annotation. In this study, we employed three complementary approaches for protein-coding gene prediction in the *Camellia sinensis* genome, these are *ab initio* prediction, homology-based prediction, and transcriptome-based prediction. Prior to gene prediction, the assembled *C. sinensis* genome underwent both hard and soft masking using RepeatMasker. *Ab initio* Gene Prediction were performed using Augustus (v3.3.1)[32,33], Genescan (v1.0)[34] and Glimmer HMM (v3.0.4)[35]. Homology-based prediction was performed by Exonerate (v2.2.0)[36], and protein sequences from three theaceae species, including Anjibaicha (AJ hereafter, a temperature-sensitive albino cultivar), Zijuan (ZJ hereafter, an *assamica* variety from the Yunnan province of China), and L618 (wild accession)[10]. Transcriptome-based prediction was performed by StringTie (v2.1.1)[37] and PASA (v2.4.1). First, high-quality RNA-Seq reads were assembled into transcripts using StringTie (v2.1.1)[37], and then gene structures subsequently annotated using PASA (v2.4.1)[38]. The results

from these three methods were integrated using MAKER (v3.00)[39]. In total, we successfully predicted 40,119 protein-coding genes in the genome (Table 2). These predicted genes exhibited an average coding sequence length of 1,179.63 bp, with an average gene length of 7,209.91 bp (Table 2).

For functional annotation of predicted genes, BLASTP (v2.6.0, E-value < = 1e-5)[40,41] was employed to align the putative genes against multiple databases including the Kyoto Encyclopedia of Genes and Genomes (KEGG)[42], Gene Ontology (GO)[43], NCBI non-redundant protein database (NR), Swiss-Prot[44], TrEMBL[45], and InterPro[46]. Furthermore, functional annotations were successfully assigned to 39,999 genes, representing 99.70% of the total predicted genes (Table 2).

Annotation of non-coding RNA genes. The tRNAscan-SE (v1.3.1)[47] was used with default parameters to identify tRNAs. We annotated non-coding rRNAs (5S, 18S, and 28S) using RNAmmer (v1.2, -S euk -multi -m lsu,ssu,tsu)[48]. For miRNA and snRNA annotation, we employed Infernal (v1.1.2)[49] with default parameters to search against the Rfam (v14.1)[50] database. In total, we identified 2,367 rRNAs, 801 tRNAs, 124 miRNAs and 271 snRNAs (Table 4).

Data Records

The raw sequencing data (MGI, PacBio, Hi-C) have been deposited in the NCBI Sequence Read Archive (SRA) database under BioProject PRJNA1427946, with the accession number SRP679934[51]. The genome assembly of *C. sinensis* has been deposited into NCBI Datasets Genome, and can be accessed according to accession number JBVOCX000000000[52]. Additionally, the chromosome-level genome assembly and annotations are available in the Genome Warehouse (GWH) at the National Genomics Data Center, China, under accession GWHGGGI00000000.1[53].

Technical Validation

The BUSCO was used to evaluate the quality of the genome assembly. We employed BUSCO to evaluate the completeness of the genome assembly. Using BUSCO (v5.1, -m genome)[54] with the embryophyta reference gene set (n = 1,614), the final genome assembly achieved a completeness score of 99.07%. Specifically, the analysis identified 1,386 (85.87%) single-copy BUSCOs, 213 (13.20%) duplicated BUSCOs, 7 (0.43%) fragmented BUSCOs, and 8 (0.50%) missing BUSCOs (Table 2).

Data availability

All software and pipelines were executed according to the manual and protocols of the published bioinformatic tools. All software used in this work is publicly available, with versions and parameters clearly described in Methods. If no detailed parameters were mentioned for a software, the default parameters suggested by the developer were used. No custom code was used during this study for the curation and/or validation of the datasets.

Code availability

All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatics software. No specific code has been developed for this study.

Received: 29 August 2025; Accepted: 26 March 2026;
Published online: 01 April 2026

References

1. Xia, E. *et al.* The Tea Tree Genome Provides Insights into Tea Flavor and Independent Evolution of Caffeine Biosynthesis. *Mol. Plant.* **10**, 866–877 (2017).
2. Wei, K. *et al.* A coupled role for CsMYB75 and CsGSTF1 in anthocyanin hyperaccumulation in purple tea. *The Plant Journal: For Cell and Molecular Biology.* **97**, 825–840 (2019).
3. Pastoriza, S., Mesías, M., Cabrera, C. & Rufián-Henares, J. A. Healthy properties of green and white teas: an update. *Food Funct.* **8**, 2650–2662 (2017).
4. Zhang, Z. *et al.* Understanding the Origin and Evolution of Tea (Camellia sinensis [L.]): Genomic Advances in Tea. *J. Mol. Evol.* **91**, 156–168 (2023).
5. Zhang, Q. *et al.* The Chromosome-Level Reference Genome of Tea Tree Unveils Recent Bursts of Non-autonomous LTR Retrotransposons in Driving Genome Size Evolution. pp. 935–938 (2020).
6. Zhang, W. *et al.* Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
7. 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).
8. Kong, W., Yu, J., Yang, J., Zhang, Y. & Zhang, X. The high-resolution three-dimensional (3D) chromatin map of the tea plant (Camellia sinensis). *Hortic. Res.* **10**, uhad179 (2023).
9. 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).
10. Tariq, A. *et al.* In-depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of Camellia sinensis. *The Plant Journal: For Cell and Molecular Biology.* **119**, 2096–2115 (2024).
11. Yu, X. *et al.* Metabolite signatures of diverse Camellia sinensis tea populations. *Nat. Commun.* **11**, 5586 (2020).
12. Jeyaraj, A. *et al.* Genome-wide identification of conserved and novel microRNAs in one bud and two tender leaves of tea plant (Camellia sinensis) by small RNA sequencing, microarray-based hybridization and genome survey scaffold sequences. *Bmc Plant Biol.* **17**, 212 (2017).
13. Wang, X. *et al.* Population sequencing enhances understanding of tea plant evolution. *Nat. Commun.* **11**, 4447 (2020).
14. 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).

15. Kong, W. *et al*. Pan-transcriptome assembly combined with multiple association analysis provides new insights into the regulatory network of specialized metabolites in the tea plant Camellia sinensis. *Hortic. Res.* **9**, uhac100 (2022).
16. Kong, W. *et al*. Genomic analysis of 1,325 Camellia accessions sheds light on agronomic and metabolic traits for tea plant improvement. *Nat. Genet.* **57**, 997–1007 (2025).
17. Xiao, C. *et al*. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat. Methods.* **14**, 1072–1074 (2017).
18. Chin, C. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods.* **13**, 1050–1054 (2016).
19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).
20. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience.* **7**, 1–6 (2018).
21. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* (2013).
22. 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).
23. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (New York, N.Y.).* **356**, 92–95 (2017).
24. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
25. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
26. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
27. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics.* **21**(Suppl 1), i351–i358 (2005).
28. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic. Acids. Res.* **35**, W265–W268 (2007).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic. Acids. Res.* **27**, 573–580 (1999).
30. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
31. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends in Genetics: Tig.* **16**, 418–420 (2000).
32. 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).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic. Acids. Res.* **34**, W435–W439 (2006).
34. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
35. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England).* **20**, 2878–2879 (2004).
36. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *Bmc Bioinformatics.* **6**, 31 (2005).
37. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
38. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic. Acids. Res.* **31**, 5654–5666 (2003).
39. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
40. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
41. Camacho, C. *et al*. BLAST+: architecture and applications. *Bmc Bioinformatics.* **10**, 421 (2009).
42. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic. Acids. Res.* **40**, D109–D114 (2012).
43. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
44. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic. Acids. Res.* **31**, 365–370 (2003).
45. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic. Acids. Res.* **28**, 45–48 (2000).
46. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic. Acids. Res.* **43**, D213–D221 (2015).
47. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic. Acids. Res.* **25**, 955–964 (1997).
48. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic. Acids. Res.* **35**, 3100–3108 (2007).
49. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England).* **25**, 1335–1337 (2009).
50. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic. Acids. Res.* **33**, D121–D124 (2005).
51. *NCBI Sequence Read Archiv* https://www.ncbi.nlm.nih.gov/sra/SRP679934 (2026).
52. Zhang, W. *Camellia sinensis* isolate YWXY, whole genome shotgun sequencing project. *Genebank* https://identifiers.org/ncbi/insdc:JBVOCX000000000 (2026).
53. *CNCB Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/98825/show (2026).
54. 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).

Acknowledgements

This work was supported by the Henan Province Central Leading Local Science and Technology Development Fund Project Funding (Z20231811160).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.C.