scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Nothapodytes nimmoniana*

Jiahui Wang[1,2,3], Xiaojie Ma[1,2,3], Xingyu Yang[1,2,3], Danni Yang[1,2], Tianping Huang[1,2], Yunqiang Yang[1,2] ✉ & Yongping Yang[1,2] ✉

Nothapodytes nimmoniana **is a plant species belonging to the genus** ***Nothapodytes*** **in the family** ***Icacinaceae*****. This species holds significant medicinal value due to its camptothecin content. In this study, we present the first chromosome-level genome assembly of** ***N. nimmoniana*** **constructed using NGS, Hi-C, and HiFi sequencing technologies. The assembled genome spans 3.53 Gb across 14 chromosomes, with an N50 length of 248.74 Mb. Genome annotation revealed that repetitive sequences constitute 80.82% of the genome size, and 83,269 protein-coding genes were predicted. Additionally, 4,360,538 bp of non-coding RNA were annotated. This genomic resource provides a foundation for further investigation into camptothecin biosynthesis pathways and plant phylogeny in** ***N. nimmoniana*****.**

Backgrounds & Summary

Nothapodytes nimmoniana is a perennial evergreen tree belonging to the genus *Nothapodytes* in the family *Icacinaceae*. This species is primarily distributed across India, Sri Lanka, Myanmar, Thailand, and Taiwan[1,2]. *N. nimmoniana* holds significant medicinal value due to its high content of camptothecin and its derivatives, which are potent anticancer compounds[3–5]. Camptothecin, a cytotoxic quinoline alkaloid, functions by inhibiting DNA topoisomerase and demonstrates substantial antitumour activity against various cancer types[6–10]. Consequently, *N. nimmoniana* serves as a valuable source for the extraction of camptothecin and its derivative compounds, underscoring its importance in the development and utilization of plant-based medicinal resources[4,11,12].

The absence of comprehensive genomic data for *N. nimmoniana* has hindered a thorough understanding of its genetic makeup and evolutionary relationships. This study, therefore, aims to construct a chromosome-level diploid whole genome sequence of *N. nimmoniana*. The primary objectives are to conduct an in-depth analysis of the biosynthetic pathways of significant medicinal alkaloids, such as camptothecin, and to elucidate other regulatory mechanisms. This genomic information is anticipated to expedite the development of novel camptothecin derivatives, potentially expanding the options for cancer treatment. Additionally, the genome sequence will contribute to clarifying the phylogenetic position of *N. nimmoniana* within the plant kingdom and enhance our comprehension of the evolutionary history of the genus *Nothapodytes*.

Methods

Plant materials and genome sequencing. To investigate the genome of *N. nimmoniana*, fresh young leaves, stems, and roots were collected from a single plant at the Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences (101°15′3″E, 21°55′41″N). High-quality DNA was extracted from the leaves using a modified cetyl trimethylammonium bromide (CTAB) method[13], purified, and used to construct a DNA library. Pair-end sequencing was then performed on the DNBSEQ-T7 platform. A total of 168.73 Gb (48x depth) of raw sequencing data was obtained, with a Q30 value exceeding 95% (Table 1). Additionally, RNA was extracted from various tissues of *N. nimmoniana* using The E.Z.N.A.® HP Plant RNA Kit, and mRNA was enriched through PolyA screening. Following fragmentation and length screening, the mRNA was reverse-transcribed to obtain

[1]Yunnan International Joint Laboratory for the Conservation and Utilization of Tropical Timber Tree Species, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan, 666303, China. [2]Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan, 666303, China. [3]University of Chinese Academy of Sciences, Beijing, 100049, China. ✉e-mail: yangyunqiang@xtbg.ac.cn; yangyp@xtbg.ac.cn

Sequence	Platform	Total reads	Total bases	GC content (%)	Q20 (%)	Q30 (%)	Sequence depth (×)
DNA reads	DNBSEQ-T7	562437332	168731199600	33.98	99.12	96.44	48.00
RNA reads	DNBSEQ-T7	402011176	120603352800	43.43	97.49	93.01	34.00
Hi-C reads	DNBSEQ-T7	1313976022	394192806600	35.40	99.51	97.32	112.00

Table 1. Characteristics of NGS data for genome assembly.

Sequence	Platform	Total reads	Total bases	Sequence depth (×)	N50 (bp)	N90 (bp)
HiFi reads	PacBio Revio	9777149	206266430391	58.00	21952	15058

Table 2. Characteristics of HiFi data for genome assembly.

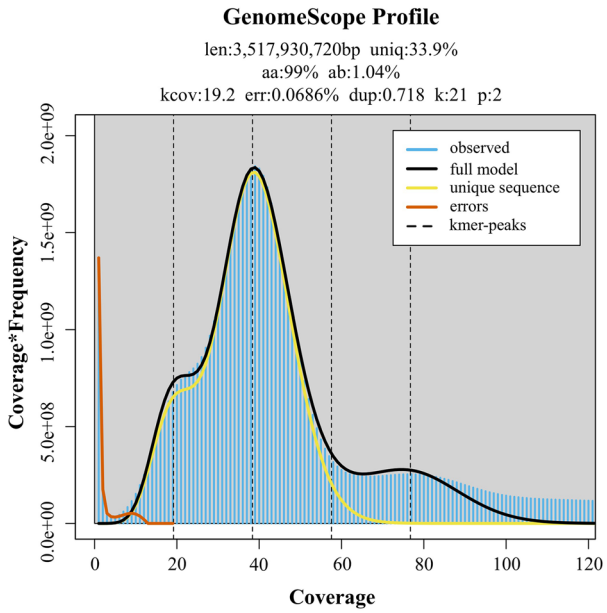

Fig. 1 K-mer distribution (K = 21) of *N. nimmoniana* genome using GenomeScope 2.

cDNA. These quality-controlled cDNA libraries were also subjected to pair-end sequencing on the DNBSEQ-T7 platform, yielding 120.60 Gb (34x depth) of raw sequencing data (Table 1).

To facilitate genome assembly and chromosome mounting, we conducted Hi-C library construction and sequencing. The process involved cross-linking and fixing the chromatin using formaldehyde, followed by treatment with *Dpn*II. Subsequently, biotin labelling and end repair were performed on the enzyme section, and the DNA library was purified and constructed. Pair-end sequencing on the DNBSEQ-T7 platform yielded 394.19 Gb (112x depth) of raw sequencing data, with a Q30 of 97.32% (Table 1), indicating high sequencing quality suitable for subsequent data analysis. Additionally, we constructed a HiFi library and performed long-read sequencing using Circular Consensus Sequencing (CCS) mode on the PacBio Revio platform, obtaining a total of 206.27 Gb of raw sequencing data (Table 2).

Estimate of genome size. Initially, fastp v0.23.2[14] (-n 0 -l 140) was employed to screen and filter low-quality fragments from NGS raw reads, yielding 157.62 Gb of NGS clean reads. Subsequently, Jellyfish's count and histogram[15] sub-commands were utilized to count the NGS clean reads with 21-kmer frequency distribution, and model fitting was performed using Genomescope v2.029[16] to predict the genome size, heterozygosity, and genome percentage. The genome size of *N. nimmoniana* was estimated at 3.51 Gb, with a heterozygosity of 1.04% and a repetitive sequence percentage of 66.1% (Fig. 1).

Genome *de novo* assembly. The genome assembly of *N. nimmoniana* was performed by integrating NGS, Hi-C, and HiFi data. Initially, fastp v0.23.2[14] (-n 0 -l 140) was utilized to screen and filter low-quality fragments from Hi-C raw reads, producing Hi-C clean reads. Subsequently, a preliminary genome assembly was conducted using hifiasm v0.19.8-r603[17–19] in conjunction with HiFi reads. This process yielded a draft genome of 3.75 Gb, comprising 618 contigs with an N50 of 231.94 Mb (Table 3). Further chromosome anchoring was executed using HiC-Pro v3.1.0[20], while 3D-DNA[21] and Juicebox[22] were employed to cluster, order, and orient contigs, resulting in a scaffold-level genome of *N. nimmoniana*. The final genome size was 3.53 Gb, consisting of 14 chromosomes with

Features	Statistics
Sequenced genome size (Gb)	3.75
Number of contigs	618
Contig N50 (bp)	231938903
Contig N90 (bp)	104429113
Max contig size (bp)	372361515

Table 3. Characteristics of the *N. nimmoniana* genome at contig level.

Features	Statistics
Number of Chromosomes	14
Scaffold N50 (bp)	248744538
Scaffold N90 (bp)	210258015
GC content (%)	32.28
Max scoffold size (bp)	372361515
Total Size (Gb)	3.53

Table 4. Characteristics of the N. nimmoniana genome at scaffold level.

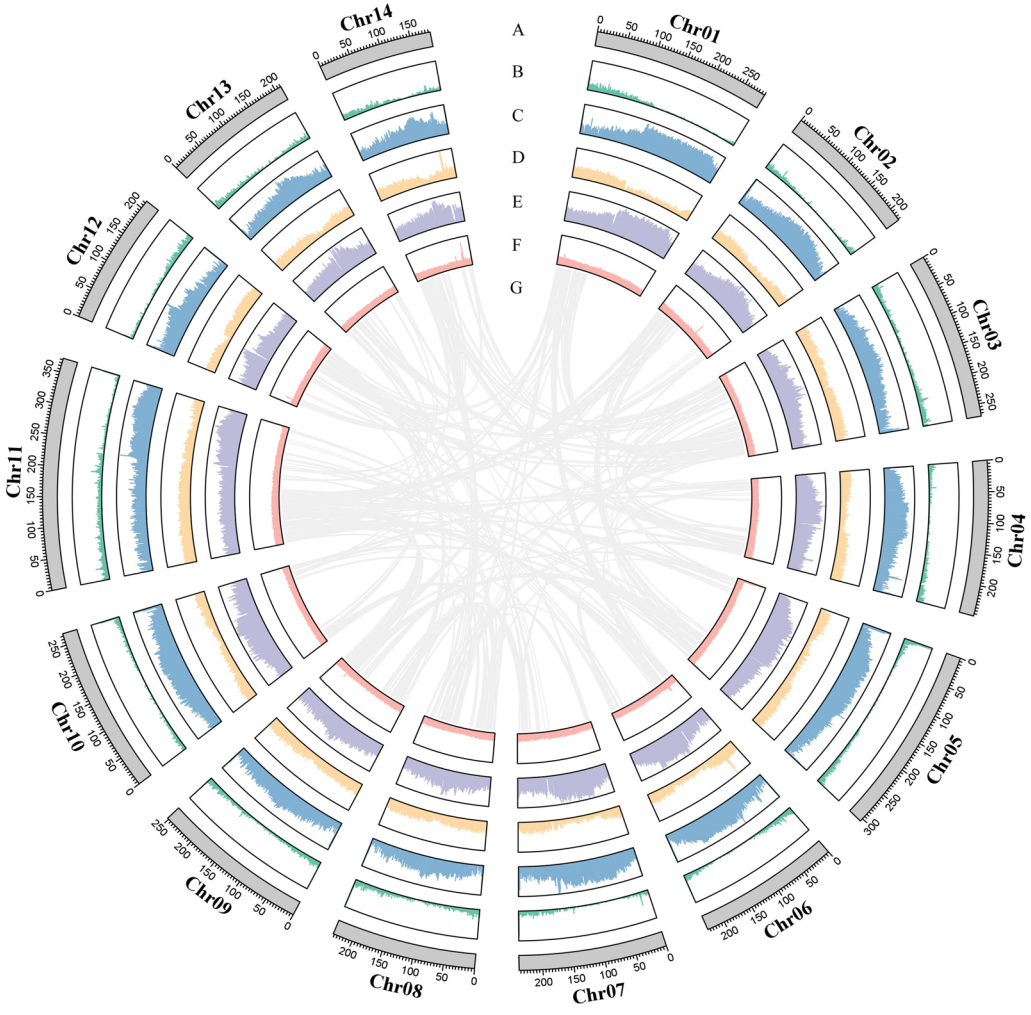

Fig. 2 Circos plot illustrating the distribution of genomic elements in *N. nimmoniana*. The statistics are calculated for every 1700 kb window across the genome sequence. The tracks depict: (**A**) chromosome lengths, (**B**) gene distribution across chromosomes, (**C**) transposable element distribution across chromosomes, (**D**) Copia element distribution across chromosomes, (**E**) Gypsy element distribution across chromosomes, (**F**) GC content of chromosomes, and (**G**) collinearity blocks between chromosomes.

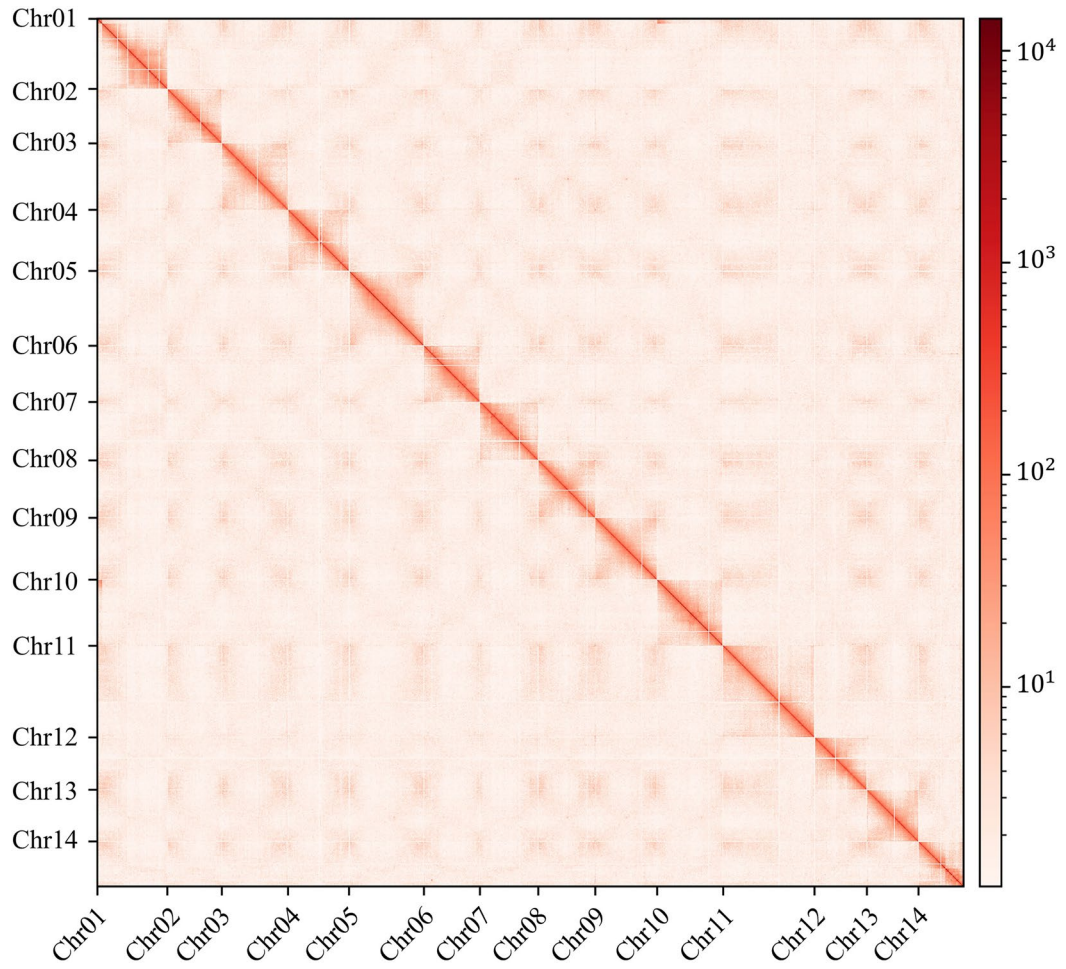

Fig. 3 Heatmap of genome-wide Hi-C data of *N. nimmoniana* chromosomes.

an N50 length of 248.74 Mb (Table 4). TBtools[23] was used to generate Circos plots (Fig. 2), while hicexplorer[24–27] was employed to create whole genome Hi-C data heatmaps (Fig. 3).

Repetitive elements identification. Following genome acquisition, we employed Extensive de-novo TE Annotator (EDTA) v2.1.2[28] to identify and annotate TEs, complemented by TEsorter v1.4.6[29] and DeepTE[30] for detailed classification. Subsequently, EDTA was utilized to quantify the repetitive sequences. The analysis revealed 2,852,972,798 bp of repetitive sequences, constituting 80.82% of the entire genome. The TEs were categorized into five groups: LTR, tandem inverted repeats (TIR), non-LTR, non-TIR, and others. The LTRs were further subdivided into Copia and Gypsy, representing 17.58% and 35.16% of the genome size, respectively (Table 5 and Fig. 2).

Protein-coding genes prediction. Following the annotation of repetitive sequences, we utilized RepeatMasker v4.1.2-p1[31] to mask these sequences in the genome. The resulting masked genome served as input for identifying protein-coding genes through three methodologies: homologous protein prediction, transcriptome prediction, and *de novo* prediction. The OrthoDB v10[32] database constituted the homologous protein library. Initially, RNA-seq clean reads from *N. nimmonian* leaves, stems, and roots were aligned to the genome using HISAT2 v2.2.1[33] to generate bam files. Subsequently, AUGUSTUS v3.4.0[34], integrated with GeneMark-ES Suite version 4.69_lic[35], was employed via BRAKER2[36] to perform homologous protein and transcriptome prediction for model training. The results were then integrated using TSEBRA v1.0.3[37]. Finally, MAKER v3.01.04[38] and GenomeTools[39] were used to correct and format the annotation files, yielding the protein-coding gene annotation files for the *N. nimmonian* genome. The analysis revealed 83,269 predicted protein-coding genes in the *N. nimmonian* genome, with gene lengths ranging from 156 to 48,075 bp (Fig. 2).

Genes function and non-coding RNA annotation. The predicted genes underwent functional annotation through homology searches of public databases, including COG[40], KOG, GO[41], KEGG[42], Swissprot[43], NR, and Pfam[44], utilizing BLASTP[45] (e-value = 1e-10). In total, 80.23% of the genes received functional annotations. The annotation rates for genes in the COG[40], KOG, GO[41], KEGG[42], Swissprot[43], NR, and Pfam[44] databases

Type		Count	Masked (bp)	Masked (%)
LTR	Copia	821701	620509802	17.58
	Gypsy	1638016	1241097411	35.16
	unknown	957197	471002927	13.34
TIR	CACTA	219688	100924982	2.86
	Mutator	371628	174122511	4.93
	PIF_Harbinger	140502	48351089	1.37
	Tc1_Mariner	120424	24924254	0.71
	hAT	162137	44270124	1.25
non-LTR	DIRS_YR	214	53428	0.00
	L1_LINE	8517	1579604	0.04
	LINE_element	53862	32651362	0.92
	Penelope	3836	980145	0.03
	tRNA_SINE	3085	417761	0.01
	unknown	38458	17825848	0.50
non-TIR	helitron	101608	35042147	0.99
others	DNA_transposon	151100	26920622	0.76
	low_complexity	150	423140	0.01
	repeat_region	57649	11645304	0.33
	retrotransposon	1363	230337	0.01
Total		4851135	2852972798	80.82

Table 5. Summary of transposable elements in *N. nimmonian* genome.

Database	Anno_num	Ratio (%)
COG	49122	58.99
KOG	23809	28.59
GO	26508	31.83
KEGG	14914	17.91
Swissprot	41268	49.56
NR	56003	67.26
Pfam	66806	80.23
Total_annotated	66806	80.23

Table 6. Statistical analysis of the functional gene annotations of the *N. nimmonian* genome.

Type	Counts	Masked (bp)
miRNA	276	36332
tRNA	738	53955
snoRNA	30982	3265367
rRNA	4647	956943
orthers	363	47941
Total	37006	4360538

Table 7. Classification of non-coding RNA in the *N. nimmonian* genome.

were 58.99%, 28.59%, 31.83%, 17.91%, 49.56%, 67.26%, and 80.23%, respectively (Table 6). Additionally, Cmscan[46] was employed to annotate non-coding RNAs in the *N. nimmonian* genome against the Rfam[47] database. This process identified 37,006 non-coding RNAs (4,360,538 bp), comprising 276 miRNAs (36,332 bp), 738 tRNAs (53,955 bp), 30,982 snoRNAs (3,265,367 bp), 4,647 rRNAs (956,943 bp), and 363 other types of RNAs (47,941 bp) (Table 7).

Data Records

The raw sequencing data are publicly available in the Genome Sequence Archive (GSA) in National Genomics Data Center (https://ngdc.cncb.ac.cn/gsa)[48,49] under the number CRA020913[50]. The genome assembly sequences and annotation files have been deposited in Figshare[51] and NCBI GenBank database[52].

BUSCO	%
Genome Complete Buscos	98.9
Complete and aingle-copy Buscos (S)	81.8
Complete and duplicated Buscos (D)	17.1
Fragemented Buscos (F)	0.4
Missing Buscos (M)	0.7

Table 8. Statistics for genome assessment using BUSCO.

Technical Validation

Furthermore, we used Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.4.3)[53,54] with the embryophyta odb10 dataset to assess the quality of genome assembly. The dataset includes 1614 conserved single-copy genes analyzed under default parameters. The analysis revealed that 98.9% of these benchmark genes were fully represented, with 81.8% identified as intact single copies and only 0.7% absent (Table 8). Collectively, these BUSCO metrics demonstrate the high-quality assembly of the *N. nimmoniana* genome, reflecting robust coverage of conserved eukaryotic gene content.

Code availability

No specific code was developed in this work.

Received: 27 February 2025; Accepted: 27 June 2025;
Published online: 08 July 2025

References

1. Editorial Committee of Flora of China, Chinese Academy of Sciences. Flora of China. *Beijing:Science Press*. (1959–2004).
2. Ito, Y. *et al*. Molecular species delimitation reveals underestimated diversity in the tree genus *Nothapodytes (Icacinaceae)*. *Plant Syst. Evol.* **308**, 3, https://doi.org/10.1007/s00606-021-01797-6 (2022).
3. Isah, T. & Mujib, A. Camptothecin from *Nothapodytes nimmoniana*: review on biotechnology applications. *Acta Physiol. Plant.* **37**, 6, https://doi.org/10.1007/s11738-015-1854-3 (2015).
4. Ao, M. Z. *et al*. Camptothecin distribution and content in *Nothapodytes nimmoniana*. *Nat. Prod. Commun.* **6**(2), 197–200, https://doi.org/10.1177/1934578X1100600210 (2011).
5. Kumara, M. S. *et al*. Biotechnology of camptothecin production in *Nothapodytes nimmoniana*, *Ophiorrhiza sp*. and *Camptotheca acuminata*. *Appl. Microbiol. Biotechnol.* **105**(24), 9089–9102, https://doi.org/10.1007/s00253-021-11700-5 (2021).
6. Liu, L. F. *et al*. Mechanism of action of camptothecin. *Ann. N. Y. Acad. Sci.* **922**, 1–10, https://doi.org/10.1111/j.1749-6632.2000.tb07020.x (2000).
7. Thomas, C. J., Rahier, N. J. & Hecht, S. M. Camptothecin: Current perspectives. *Bioorgan. Med. Chem.* **12**, 1585–1604, https://doi.org/10.1016/j.bmc.2003.11.036 (2004).
8. Wang, X., Zhuang, Y., Wang, Y., Jiang, M. & Yao, L. Recent advances in camptothecin and its derivatives as potential antitumor agents. *Eur. J. Med. Chem.* **234**, 115710, https://doi.org/10.1016/j.ejmech.2023.115710 (2023).
9. Wall, M. E. *et al*. Plant antitumor agents. I. The isolation and structure of camptothecin, a novel alkaloidal leukemia and tumor inhibitor from *Camptotheca acuminata*. *J.Am.Chem.Soc.* **88**(16), 3888–3890, https://doi.org/10.1021/ja00968a057 (1966).
10. Ramesha, B. T. *et al*. Prospecting for camptothecines from *Nothapodytes nimmoniana* in the Western Ghats, south India: identification of high-yielding sources of camptothecin and new families of camptothecines. *J. Chromatogr. Sci.* **46**(4), 362–368, https://doi.org/10.1093/chromsci/46.4.362 (2008).
11. Shrivastava, V., Sharma, N., Shrivastava, V. & Sharma, A. Review on camptothecin producing medicinal plant: *Nothapodytes nimmoniana*. *Biomed. Pharmacol. J.* **14**(4), 1799–1813, https://doi.org/10.13005/bpj/2279 (2021).
12. Zhang, Y. B. *et al*. Biotechnology of camptothecin production in *Nothapodytes nimmoniana, Ophiorrhiza sp*. and *Camptotheca acuminata*. *Front. Pharmacol.* **105**(24), 9089–9102, https://doi.org/10.1007/s00253-021-11700-5 (2021).
13. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **15**, 8–15, https://doi.org/10.1007/BF02772108 (1997).
14. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
15. Marçais, G. & Kingsford, C. 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).
16. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
17. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
18. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335, https://doi.org/10.1038/s41587-022-01261-x (2022).
19. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat Methods.* **21**, 967–970, https://doi.org/10.1038/s41592-024-02269-8 (2024).
20. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259, https://doi.org/10.1186/s13059-015-0831-x (2015).
21. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scafolds. *Sci.* **356**, 92–95, https://doi.org/10.1126/science.aal332 (2017).
22. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
23. Chen, C. *et al*. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant.* **16**, 1733–1742, https://doi.org/10.1016/j.molp.2023.09.010 (2023).
24. Joachim, W., Rolf, B. & Björn, G. Loop detection using Hi-C data with HiCExplorer. *GigaScience.* **11**, giac061, https://doi.org/10.1093/gigascience/giac061 (2022).
25. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**(W1), W177–W184, https://doi.org/10.1093/nar/gkaa220 (2020).

26. Joachim, W. *et al*. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**(W1), W11–W16, https://doi.org/10.1093/nar/gky504 (2018).
27. Fidel, R. *et al*. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**(1), 189, https://doi.org/10.1038/s41467-017-02525-w (2018).
28. Ou, S. J. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**(1), 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
29. Zhang, R. G. *et al*. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic. Res.* **9**, uhac017, https://doi.org/10.1093/hr/uhac017 (2022).
30. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network. *Bioinformatics*. **36**, 4269–4275, https://doi.org/10.1093/bioinformatics/btaa519 (2020).
31. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics.* **21**, i351–i358, https://doi.org/10.1093/bioinformatics/bti1018 (2005).
32. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**(D1), D807–D811, https://doi.org/10.1093/nar/gky1053 (2019).
33. Daehwan, K., Ben, L. & Steven, L. S. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods.* **12**(4), 357–360, https://doi.org/10.1038/nmeth.331 (2015).
34. Mario, S., Mark, D., Robert, B. & David, H. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* **24**(5), 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
35. Tomáš, B., Alexandre, L. & Mark, B. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.* **2**(2), 1–14, https://doi.org/10.1093/nargab/lqaa026 (2020).
36. Tomáš, B. *et al*. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform*. **3**(1), 1–11, https://doi.org/10.1093/nargab/lqaa108 (2021).
37. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics.* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
38. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics.* **48**, 4.11.1–14.11.39, https://doi.org/10.1002/0471250953.bi0411s48 (2014).
39. Gordon, G., Sascha, S. & Stefan, K. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE ACM T COMPUT BI.* **10**(3), 645–656, https://doi.org/10.1109/TCBB.2013.68 (2013).
40. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science.* **278**, 631–637, https://doi.org/10.1126/science.278.5338.631 (1997).
41. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
42. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462, https://doi.org/10.1093/nar/gkv1070 (2016).
43. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
44. Typhaine, P. *et al*. InterPro in 2022. *Nucleic Acids Res.* **51**(D1), D418–D427, https://doi.org/10.1093/nar/gkac993 (2023).
45. Kent, W. K. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664, https://doi.org/10.1101/gr.229202 (2002).
46. Madeira, F. *et al*. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279, https://doi.org/10.1093/nar/gkac240 (2022).
47. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **1**, D121–D124, https://doi.org/10.1093/nar/gki081 (2005).
48. Chen, T. *et al*. The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics.* **19**(4), 578–583, https://doi.org/10.1016/j.gpb.2021.08.001 (2021).
49. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* **52**(D1), D18–D32, https://doi.org/10.1093/nar/gkad1078 (2024).
50. *CNCB-NGDC Genome Sequence Archive.* https://ngdc.cncb.ac.cn/gsa/browse/CRA020913 (2025).
51. Wang, J. H. The assembled genome and related data of *Nothapodytes nimmoniana*. *figshare*. https://doi.org/10.6084/m9.figshare.27952785.v2 (2024).
52. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBNVYP000000000 (2025).
53. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**(19), 3210–2, https://doi.org/10.1093/bioinformatics/btv351 (2015).
54. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **1**(12), e323, https://doi.org/10.1002/cpz1.323 (2021).

Acknowledgements

This research was supported by the Major Program of National Natural Science Foundation of China (31590820, 31590823), the National Natural Science Foundation of China (31601999 and 41771123), the West Light Foundation of the Chinese Academy of Sciences (to YQY), Yunling Scholar Project to Yongping Yang (QYXTZX-RKZ2022-01).

Author contributions

Yongping Yang and Yunqiang Yang conceived the study and supervised the project. Jiahui Wang and Xiaojie Ma wrote the manuscript and participated in the data analysis. Xingyu Yang, Danni Yang, Tianping Huang, Yunqiang Yang collected the samples, performed the figures drawing and upload the data. All authors have read, revised, and approved the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

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