scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the medicinal plant *Ophiorrhiza japonica* Blume

Xiaoxin Tang[1,2,4], Yunjing Liu[1,4], Yiying Liao[3,4], Ming Tang[1] ✉, Tuo Yang[3] ✉ & Yin Yi[1] ✉

***Ophiorrhiza japonica*, a medicinal plant of Rubiaceae, has been selected as a model plant for the study of MIA biosynthesis and regulation, as well as a sustainable source of camptothecin. Here, we performed an assembly and annotation of *O. japonica* genome. To achieve this, we employed a range of advanced techniques, including flow cytometry, PacBio HiFi sequencing, ONT RNA-sequencing and Hi-C technology. This approach enabled us to construct a high quality, chromosome-level genome of *O. japonica*. The assembled *O. japonica* genome spanned 549.81 Mb with a contig N50 size of 43 Mb and a scaffold N50 size of 46.45 Mb. The 24 contigs, representing 99.42% of the total assembled genome, were anchored to 11 chromosomes using Hi-C scaffolding. A total of 313.49 Mb of repeat sequences were identified and 28,182 protein-coding genes were predicted. The findings of this study provide invaluable genomic resources that will facilitate a deeper understanding of species evolution and enable the investigation of a range of crucial traits.**

Background & Summary

Camptothecin (CPT) is a potent anticancer compound whose sustainable supply is hampered by the limited and endangered botanical sources, primarily *Camptotheca acuminata*[1,2]. The herbaceous genus *Ophiorrhiza* (Rubiaceae), particularly *O. japonica*, has emerged as a promising alternative model due to its shorter life cycle and genetic tractability[3,4]. However, the lack of a high-quality genomic resource has hindered in-depth studies on its CPT biosynthetic capacity and the evolutionary origins of monoterpenoid indole alkaloids (MIAs) in this lineage.

Here, we present the first chromosome-scale genome assembly of *Ophiorrhiza japonica*, generated by integrating PacBio HiFi, Hi-C, and Oxford Nanopore RNA-seq data. The assembly spans 549.81 Mb with a scaffold N50 of 46.45 Mb, and 99.42% of the sequences are anchored to 11 chromosomes. We annotated 28,182 protein-coding genes, with both assembly and annotation showing exceptional completeness (BUSCO scores of 98.2% and 98.8%, respectively).

Beyond providing a fundamental resource, our analyses yielded key evolutionary insights. Comparative genomics revealed that the establishment of strictosidine biosynthesis-the central gateway to MIAs-predates the diversification of downstream CPT-specific pathways. This critical evolutionary innovation occurred after the ancient γ whole-genome triplication event (~110 Mya) shared with *O. pumila*[5], but before the more recent whole-genome duplication in *C. acuminata*. Furthermore, by integrating genomic, transcriptomic, and metabolomic data, we delineated putative metabolic gene clusters associated with CPT biosynthesis and propose an evolutionary model for MIA diversification in *Ophiorrhiza*.

This high-quality genome assembly not only establishes *O. japonica* as a robust model for engineering camptothecin production but also provides a pivotal genomic framework for elucidating the evolutionary history of medicinal alkaloid biosynthesis in plants.

[1]Key Laboratory of National Forestry and Grassland Administration on Biodiversity Conservation in Karst Mountainous Areas of Southwestern China, School of Life Science, Guizhou Normal University, Guiyang, China. [2]School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore, Singapore. [3]Key Laboratory of Southern Subtropical Plant Diversity, Fairy Lake Botanical Garden, Shenzhen & Chinese Academy of Science, Shenzhen, China. [4]These authors contributed equally: Xiaoxin Tang, Yunjing Liu, Yiying Liao. ✉e-mail: mingtang@gznu.edu.cn; yangtuo@szbg.ac.cn; gzklppdr@gznu.edu.cn

Sample No.	Reference	Reference Intensity	Sample Intensity	Ratio	Size(Gb)
L-L-W-1	Tomato	15.62	10.28	0.66	0.58
L-L-W-1	Maize B73	39.19	9.63	0.25	0.57
L-L-W-2	Tomato	15.11	9.99	0.66	0.58
L-L-W-3	Tomato	14.42	9.55	0.66	0.58

Table 1. Flow cytometry analysis of the *Ophiorrhiza japonica* genome size.

Type	Platform	Number of Reads	Total bases (bp)
Long-reads WGS	Sequell	2,151,148	32,245,225,122
High quality Iso-seq	ONT	5,816,111	7,846,541,110
Hi-C	Novaseq	170,105,636	50,882,591,642

Table 2. Sequencing data statistics for long-read and Hi-C data of the *Ophiorrhiza japonica*.

Methods

Flow cytometry (FCM) analyses for nuclear genome size estimation. Herein, we aimed to estimate the genome size of *O. japonica* using flow cytometry protocol with propidium iodide (PI)-stained nuclei, following the described protocol. Mature leaf tissue were sampledh from an *O. japonica* individual collected in Guizhou Province, China. The samples were placed in 0.8 mL of pre-cooled MGB dissociation solution, containing 45 mM $MgCl_2{\cdot}6H_2O$, 20 mM MOJS, 30 mM sodium citrate, 1% (w/v) PVP 40, and 0.2% (v/v) Tritonx-100. To this mixture, 0.8 mL of 10 mM Na_2EDTA and 20 μL/mL β-mercaptoethanol (pH 7.5) was added, and the samples were rapidly chopped vertically with a sharp razor blade. Subsequently, the homogenate was incubated on ice for 10 minutes in the dissociation solution, followed by filtration through a 40 μm pore size filter in order to obtain the cell nucleus suspension.

For staining, the nuclei suspension was mixed with an appropriate volume of pre-cooled propidium iodide (PI) (1 mg/mL) and an appropriate volume of RNAase solution (1 mg/mL) at a final working concerntration of 50 μg/mL for both reagents[6]. The stained suspension was incubated on ice in the dark for 30–60 minutes. In order to facilitate a comparative analysis of genome size, the tomato and maize B73, which have a known genome size of 900 Mbp and 2.3 Gbp, respectively, were processed following the identical protocol. Flow cytometric analysis yielded an genome size approximately 0.5775 ± 0.0043 Gb (Table 1)

Plant material, DNA library construction and sequencing. In December 2022, dozens of *O. japonica* plants were collected in Guiyang City, Guizhou Province, China for genome sequencing. Fresh and healthy leaves weretaken from the most vigorous individual and promptly frozen in liquid nitrogen. The samples were then stored at −80 °C until DNA extraction.

High quality genomic DNA was isolated from the leaves using a modified cetyltrimethylammonium bromide (CTAB) method, following the protocol of[7]. The integrity and concentration of the DNA were assessed by 0.75% agarose gel electrophoresis and a NanoDrop One spectrophotometer (Thermo Fisher Scientific). A PacBio HiFi library was prepared using the SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences, PacBio) according to the manufacturer's instructions. This library was then sequenced on the PacBio Sequel II sequencing platform. This process yielded 32.24 Gb of raw PacBio continuous long reads (HiFi) with an average length of 14.98 kb (Table 2).

Plant material, RNA extraction and sequencing. For the purpose of genome annotation, we sampled roots, stems, leaves and flowers from *O. japonica* plants. The tissues were rapidly frozen in liquid nitrogen and stored at −80 °C freezers until RNA extraction. The tissues (roots, stems, leaves, flowers) were combined and subsequently subjected to Oxford Nanopore cDNA sequencing. cDNA library construction was conducted using the cDNA-PCR Sequencing Kit (SQK-PCS109), followed by sequencing on a PromethION sequencer. yielded 7.84 Gb of transcriptome vato support for gene annotations (Table 2).

Hi-C library construction and sequencing. After formaldehyde cross-linking, the library construction was initiated. After passing quality control, the high-throughput sequencing was conducted on the Illumina NovaSeq. 6000, generating a total of 50.88 Gb of paired-end clean reads, each 150 bp in length (Table 2).

Nuclear genome assembly. We generated a primary contig assembly using hifiasm v.0.16.1 with default parameters and using the HiFi reads as input[8]. The resulting contig assembly was then scaffolded using Hi-C reads via ALLHIC v0.9.8[9]. The final chromosome-level assemblies were generated after manually adjusting and correcting assembling errors in Juicebox v3.0[10]. After Hi-C-assisted assembly, the total length of chromosome-anchored sequences accounted 99.42% of the entire genome (Table 3), which were assembled into 11 chromosomes (Figs. 2 and 3) with lengths ranging from 41.87 Mb to 66.62 Mb and an N50 size of 46.45 Mb (Table 3 and Table 4). We used BUSCO v5.50[11] with the eukaryotic ortholog database (Eukaryota_odb10) to evaluate the completeness of the genome. The BUSCO completeness metrics were employed for the purpose of facilitating a comparison of genome statistics for *O. japonica* with those for another CPT or CPT analogue-producing organism, namely *Camptotheca acuminata* (Nyssaceae) and *O. pumila* (Rubiaceae), for which whole genome data is available. (Table 5)

Chromosome	Length (bp)	Contig number
chr1	66626000	1
chr2	62292267	4
chr3	60488254	6
chr4	50315677	2
chr5	46450834	1
chr6	45784361	1
chr7	44576775	2
chr8	44404086	1
chr9	44176543	1
chr10	42820813	4
chr11	41875368	1
chrUn	3179363	43
Chr	549810978	24
Total	552990341	67

Table 3. Chromosome and corresponding statistical results after Hi-C assisted assembly.

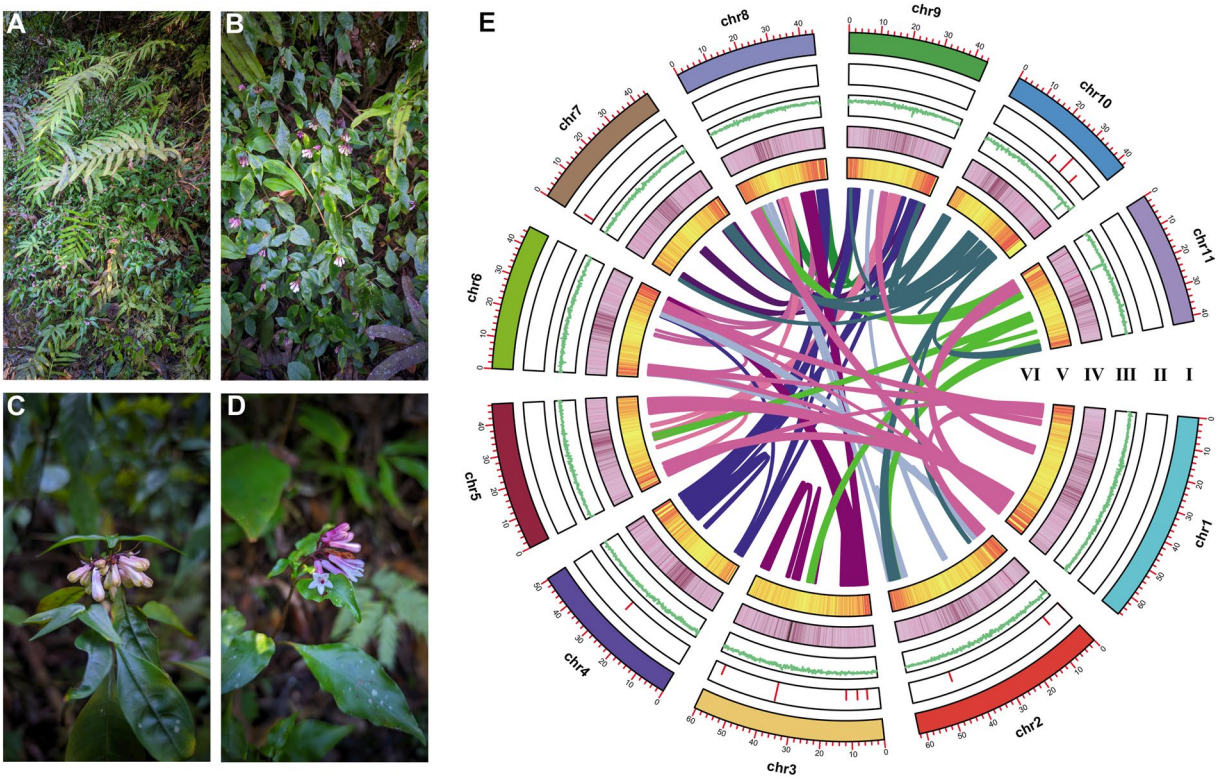

Fig. 1 The habitat, morphological characteristics and genomic features of *Ophiorrhiza japonica*. (**A**) The understory habitat. (**B**) The populations. (**C**) The inflorescence. (**D**) The long-styly flower. (**E**) Features of the *Ophiorrhiza japonica* genome. From outside to inside: (I) the 11 pseudo-chromosomes, (II) N content, (III) GC skew, (IV) GC content, (V) gene density, (VI) syntenic blocks.

Genome annotation. Genome annotation encompasses three core components: the identification of repetitive sequences, the annotation of mRNA and non-coding RNA, and functional annotation of mRNA.

To identify repetitive sequences, RepeatModeler v2.0.1[12] was initially employed to ascertain the presence of repeat elements within the *O. japonica* genome. Following this,, the predicted results were integrated with the RepBase database. Subsequently, RepeatMasker v4.1.0[13] was employed for the prediction of repetitive sequences within the genome. The results demonstrated that approximately 313.49 Mb of the repeat sequence in *O. japonica* constituted 56.69% of the entire genome. The most prevalent repetitive sequence is the long terminal repeat (LTR), which exceeds 196.16 Mb in length and accounts for 35.47% of the total. Additionally, other repetitive sequences comprise 25.62 Mb (4.63%) of DNA transposons,22.56 Mb (4.08%) of long interspersed elements (LINE), and 0.064 Mb (0.01%) of simple sequence repeats (SSRs) (Table 6).

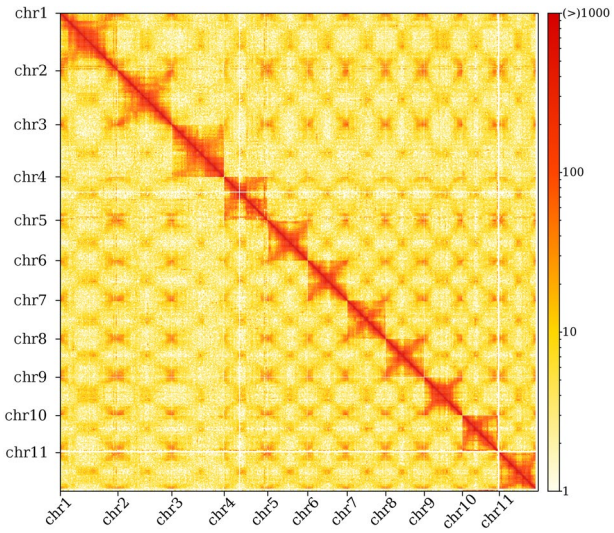

Fig. 2 Contact map plot of the *Ophiorrhiza japonica* genome. The Hi-C raw read pairs were aligned with the genome sequences; the x and y axes indicate their positions; the red dots indicate the position of the read pairs, and a high density of red dots denotes that they are located on the same chromosome.

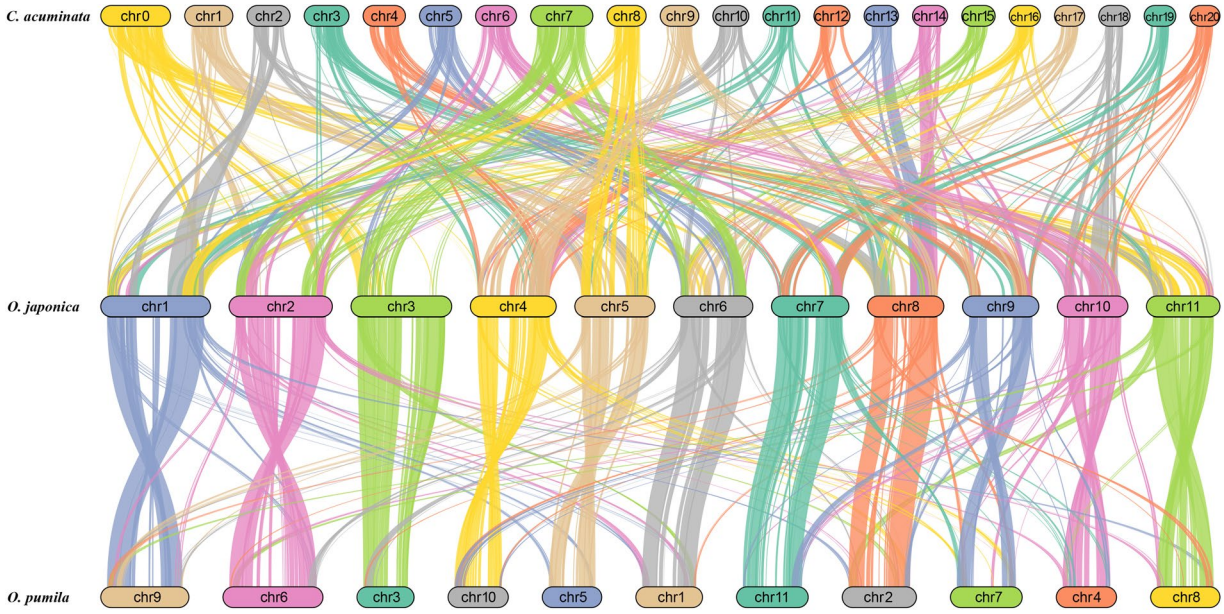

Fig. 3 Genome-wide synteny among three camptothecin-producing plants.

Three approaches were employed for the prediction of mRNA: *ab initio* gene prediction, homology-based prediction, and transcript prediction. The *ab initio* gene prediction was conducted using the Augustus v3.3.3[14] and GlimmerHMM v3.0.4[15] software. For homology-based prediction, we employed exonerate[16] and GeMoMa v1.9[6] to compare homologous proteins from five related species (*Coffea arabica*, *Coffea canephora*, *Coffea eugenioides*). The RNA-seq data were reconstructed using StringTie v2.1.3[17], and then TransDecoder v5.1.0[18] was employed to predict coding sequences. The resulting multiple datasets were integrated using EVidenceModeler v1.1.1[19]. Subsequently, PASA v2.5.2[20] was utilised to update the integrated data, incorporating UTR regions and identifying novel transcripts. Our predictions identified a total of 28,182 genes in the genome, with an average gene length of 4,152 bp, an average coding sequence length of 1,392 bp, and an average of 5.16 exons per gene (Table 7). The vast majority of these final gene models are supported by RNA-seq evidence, as the PASA-based curation step explicitly incorporates transcriptomic alignments for validation and refinement.

The nucleotide sequences of the longest predicted transcripts for each gene were aligned against the NR[21], Swiss-Prot[22], GO[23], COG, KOG, and KEGG[24] databases using BLAST v2.10.1[25]. The amino acid sequences were compared to the Pfam database[26] using HMMER v3.2.1[27], which provided functional annotation information

	Scaffold	Contig
Length (bp)	552990341	552988941
GC (%)	35.16	35.16
N50	46450834	43005162
L50	5	6
N75	44404086	24700748
L75	8	10
N90	42820813	15108609
L90	10	14
Longest	66626000	66626000

Table 4. Assembly statistics of the *Ophiorrhiza japonica* genome.

BUSCOs	*Ophiorrhiza japonica*	*Ophiorrhiza pumila*	*Camptotheca acuminata* V3.0
Complete (C)	98.20%	97.10%	98.50%
Complete and single-copy (S)	92.90%	95.00%	93.00%
Complete and duplicated (D)	5.30%	2.10%	5.50%
Fragmented (F)	0.70%	0.40%	0.60%
Missing (M)	1.10%	2.50%	0.90%

Table 5. BUSCO assessment results of available three genomes.

Type	*De novo* + repbase	(%)in genome
DNA	25624669	4.63
LINE	22568553	4.08
SINE	6120	0
LTR	196169995	35.47
Satellite	174247	0.03
Simple_repeat	64125	0.01
Other	6935391	1.25
Unknown	61950487	11.2
Total	313493587	56.69

Table 6. Summary of annotated transposable elements of the *Ophiorrhiza japonica*.

Method	Software	Species	Gene number	Average gene length(bp)	Average CDS length(bp)	Average exon number per gene	Average exon length(bp)	Average intron length(bp)
Ab initio	AUGUSTUS		34834	4039	1184	4.69	253	774
Ab initio	GlimmerHMM		44777	7762	837	4.27	196	2117
Homology-based	exonerate and GeMoMa	*Coffea arabica*	75152	11900	1191	4.99	239	2681
Homology-based	exonerate and GeMoMa	*Coffea canephora*	33282	10974	1000	4.22	237	3101
Homology-based	exonerate and GeMoMa	*Coffea eugenioides*	44438	12614	1163	4.8	242	3011
Integration	EVM		28236	3942	1341	4.67	287	710
Final set	PASA		28182	4152	1392	5.16	376	578

Table 7. Summary of structures annotations of the *Ophiorrhiza japonica* genome.

for the predicted genes. Of the total number of predicted genes in the *O. japonica* genome, 28,101 (99.71%) were successfully predicted, with an average CDS length of 1.3 kb (Fig. 4 and Table 8).

The prediction of non-coding RNAs is comprised of three distinct phases. The first phase involved the prediction of rRNA sequences, which were accomplished using Barrnap v0.9. The second phase focused on the prediction of tRNA sequences, which were predicted using tRNAscan v2.0.0[28]. Finally, Infernal v1.1.3[29] was employed to predict non-coding RNA by searching the Rfam database[30] (Table 9).

Genome-wide synteny analysis. We identified the syntenic blocks within the *O. japonica* genome, as well as between this genome and two published camptothecin-producing plants genomes, using the One Step MCScanX implemented in TBtools-II[31], with default parameters. Intra-genomic syntenic blocks were visualized using TBtools (Fig. 1E), while inter-genomic syntenic blocks were visualized using ChiPlot (https://www.chiplot.online/) (Fig. 3).

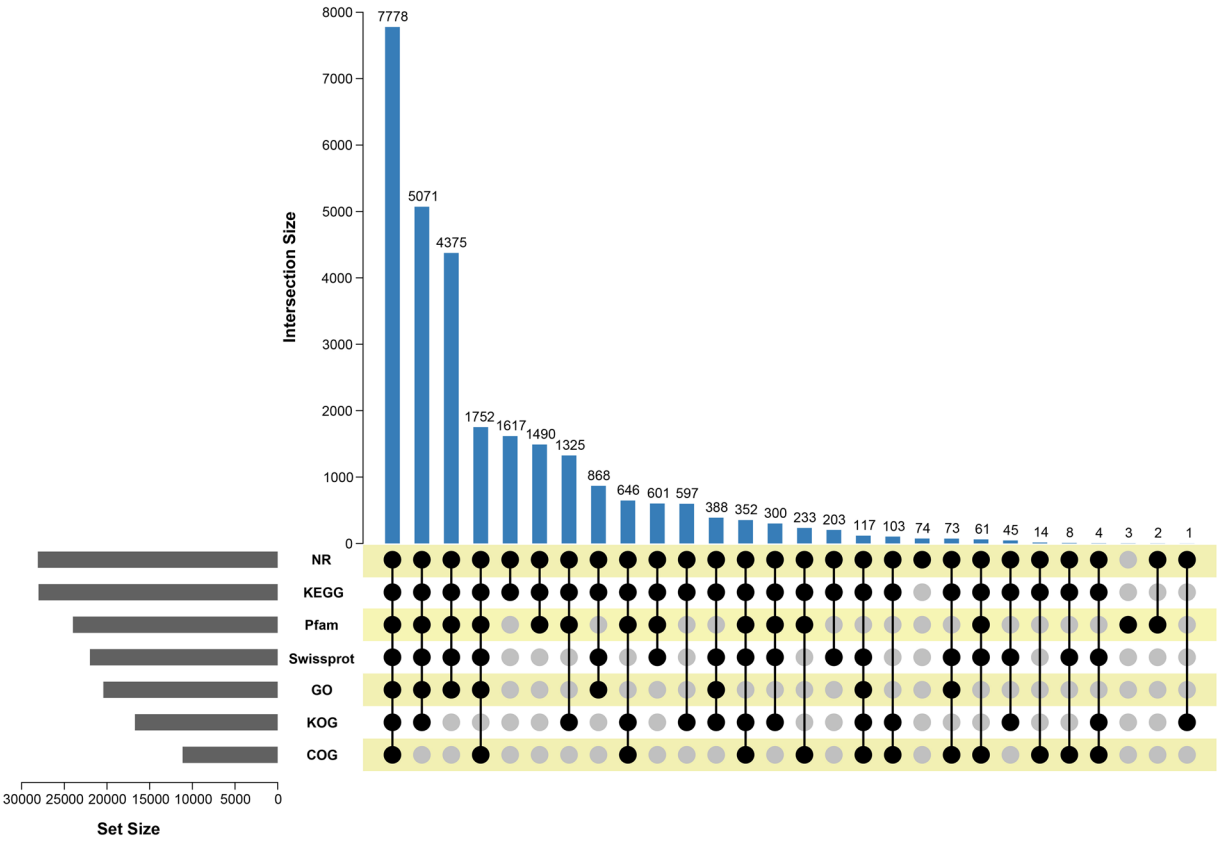

Fig. 4 UpSet plot summarizing the functional annotation of the genome. (Note: Bottom panel (Intersection size): The vertical bars represent the number of genes annotated in specific combinations of databases. Each of the 28 bars on the X-axis corresponds to a unique combination, as defined by the matrix above it. The height of the bar (Y-axis) indicates the number of genes shared by that exact combination. Middle panel (Intersection matrix): This matrix specifies the composition of each combination. Each row represents one of the six databases. A filled dot indicates that the database is part of the combination for the specific bar below it. For example, the tallest bar (far right) corresponds to the combination where all six databases have annotations for the same set of genes. Left panel (Set size): The horizontal bars show the total number of genes annotated in each individual database, irrespective of overlaps with other databases).

Database	Annotated_Number	Percent of all genes (%)	300 < = length < 1000	length > = 1000
COG	11141	39.53	3504	7435
GO	20422	72.46	7108	12780
KEGG	28021	99.42	10928	16059
KOG	16727	59.35	5827	10492
Pfam	23989	85.12	8484	14955
Swissprot	21996	78.04	7699	13721
NR	28098	99.7	10980	16066
All	28101	99.71	10983	16066

Table 8. Summary of functional annotations of the *Ophiorrhiza japonica* genome.

Data Record

The raw sequencing data of *Ophiorrhiza japonica* have been submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive(SRA) under accession numbers SRR31367429, SRR31367428, and SRR31367427[32–34]. The genome assembly is available at NCBI GenBank under accession JBNJRQ000000000.1[35]. The genome annotation files are available at Figshare under https://doi.org/10.6084/m9.figshare.29603738.v2[36]. In addition, the sequencing data are also available at the Genome Sequence Archive (GSA) in National Genomics Data Center (NGDC) under BioProject accession PRJCA027769 and dataset accession CRA020465[37–39].

Technical Validation

The quality of the *Ophiorrhiza japonica* genome is robustly supported by key metrics.

Type		Total number	Average length	Total length	(%) in genome
tRNA		624	74	46580	0.0084
sRNA		1	376	376	0.0001
miRNA		118	126	14918	0.0027
rRNA	18S_rRNA	63	1806	113823	0.0206
	28S_rRNA	85	3708	315258	0.057
	5.8S_rRNA	86	153	13222	0.0024
	5S_rRNA	335	111	37471	0.0068
	Total	569	843	479774	0.0868
snRNA	CD-box	1947	105	206352	0.0373
	HACA-box	79	137	10836	0.002
	scaRNA	0	0	0	0
	splicing	170	120	20432	0.0037
	snRNA-other	0	0	0	0
	total	2196	108	237620	0.043
Cis-reg	frameshift_element	0	0	0	0
	IRES	0	0	0	0
	leader	0	0	0	0
	riboswitch	1	133	133	0
	thermoregulator	0	0	0	0
	other	6	48	288	0.0001
	total	7	60	421	0.0001

Table 9. Statistics of non-coding RNA annotation.

Assembly completeness was exceptional, with a BUSCO score of 98.2% (embryophyta_odb10), surpassing the benchmark for high-quality plant genomes (Table 5).

Chromosome assembly was highly accurate. The Hi-C contact map shows strong intra-chromosomal interactions and clear boundaries (Fig. 2), with >99% of sequences anchored to chromosomes and minimal inter-chromosomal signal (Table 3).

The completeness of the gene annotation is high, with 98.8% of the conserved embryophyte BUSCOs identified in the annotated gene set.

Data availability

The raw sequencing data of *Ophiorrhiza japonica* have been submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive(SRA) under accession numbers SRP545588: https://www.ncbi.nlm.nih.gov/sra/ SRP545588 The genome assembly is available at NCBI GenBank under accession JBNJRQ000000000.1: https://www.ncbi.nlm.nih.gov/nuccore/JBNJRQ000000000.1 The genome annotation files are available at Figshare: https://doi.org/10.6084/m9.figshare.29603738.v2. All data has also been deposited in National Genomics Data Center under the accession number PRJCA027769: https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA027769.

Code availability

The manuscript did not use custom code to generate or process the data described. All software and pipelines were executed according to the manual and protocols of the published bioinformatic tools.

Received: 20 December 2024; Accepted: 31 January 2026;
Published online: 09 February 2026

References

1. Wall, M. E. *et al*. Plant Antitumor Agents. I. The Isolation and Structure of Camptothecin, a Novel Alkaloidal Leukemia and Tumor Inhibitor from Camptotheca acuminata1,2. *Journal of the American Chemical Society* **88**, 3888–3890, https://doi.org/10.1021/ja00968a057 (2002).
2. Pu, X. *et al*. Possible clues for camptothecin biosynthesis from the metabolites in camptothecin-producing plants. *Fitoterapia* **134**, 113–128, https://doi.org/10.1016/j.fitote.2019.02.014 (2019).
3. Sudo, H., Yamakawa, T., Yamazaki, M., Aimi, N. & Saito, K. Bioreactor production of camptothecin by hairy root cultures of Ophiorrhiza pumila. *Biotechnology Letters* **24**, 359–363, https://doi.org/10.1023/a:1014568904957 (2002).
4. Sirikantaramas, S., Yamazaki, M. & Saito, K. Mutations in topoisomerase I as a self-resistance mechanism coevolved with the production of the anticancer alkaloid camptothecin in plants. *Proc Natl Acad Sci USA* **105**, 6782–6786, https://doi.org/10.1073/pnas.0801038105 (2008).
5. Rai, A. *et al*. Chromosome-level genome assembly of Ophiorrhiza pumila reveals the evolution of camptothecin biosynthesis. *Nat Commun* **12**, 405, https://doi.org/10.1038/s41467-020-20508-2 (2021).
6. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods Mol Biol* **1962**, 161–177, https://doi.org/10.1007/978-1-4939-9173-0_9 (2019).
7. Abu Almakarem, A. S., Heilman, K. L., Conger, H. L., Shtarkman, Y. M. & Rogers, S. O. Extraction of DNA from plant and fungus tissues *in situ*. *BMC Res Notes* **5**, 266, https://doi.org/10.1186/1756-0500-5-266 (2012).

8. 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).
9. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
10. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256–258 e251, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
11. Manni, M., Berkeley, M. R., Seppey, M., Simao, 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, https://doi.org/10.1093/molbev/msab199 (2021).
12. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
13. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4.10.11–14.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
14. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
15. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
16. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, https://doi.org/10.1186/1471-2105-6-31 (2005).
17. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
18. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
19. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
20. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
21. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–65, https://doi.org/10.1093/nar/gkl842 (2007).
22. Dimmer, E. C. *et al*. The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**, D565–570, https://doi.org/10.1093/nar/gkr1048 (2012).
23. 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).
24. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
26. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222–230, https://doi.org/10.1093/nar/gkt1223 (2014).
27. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205–211, https://doi.org/10.1142/9781848165632_0019 (2009).
28. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
29. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
30. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–124, https://doi.org/10.1093/nar/gki081 (2005).
31. 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).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31367427 (2024).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31367429 (2024).
34. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31367428 (2024).
35. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBNJRQ000000000.1 (2025).
36. Tang, X. *et al*. Chromosome-level Genome Assembly and Annotation of Ophiorrhiza japonica. *figshare* https://doi.org/10.6084/m9.figshare.29603738.v2 (2025).
37. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA020465 (2024).
38. Chen, T. *et al*. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics Proteomics Bioinformatics* **19**, 578–583, https://doi.org/10.1016/j.gpb.2021.08.001 (2021).
39. Members, C.-N., Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res* **52**, D18–D32, https://doi.org/10.1093/nar/gkad1078 (2024).

Acknowledgements

This work was funded by various sources, including the National Natural Science Foundation of China (32360330), China Scholarship Council (202308520101), Joint Fund of the National Natural Science Foundation of China and the Karst Science Research Center of Guizhou Province, (U1812401), Integration and Demonstration of Karst Rocky Desertification Ecological Restoration Technologies (QianKeHeChengGuo [2022]010) and Water- Fertilizer Coupling and Biodiversity Restoration in Karst Rocky Desertification (QianJiaoJi[2023]004), UrbanAdministration and Law Enforcement Bureau of Shenzhen Municipality ResearchProgram (202406).

Author contributions

Y.Y., T.Y. and M.T. conceived and designed the study. X.X.T., Y.J.L. and Y.Y.L. provided the samples, assembled the genome data, performed bioinformatics analysis, wrote the manuscript and prepared the figures and revised and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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