scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Telomere-to-telomere gapless genome assembly of *Acorus tatarinowii*

Lei Huang[1,2], Zijie Wei[1,2] & Xinwei Xu[1] ✉

***Acorus tatarinowii*, a medicinal herb belonging to the Acoraceae family, is widely used in aromatherapy and cognitive enhancement due to its rich repertoire of over 160 natural products. Phylogenetically, it holds a key position as the sister lineage to all other monocots, underscoring the need for a high-quality genome assembly to facilitate in-depth evolutionary and functional studies. In this study, we present a telomere-to-telomere (T2T) genome assembly of *A. tatarinowii*, spanning 359.36 Mb with a scaffold N50 of 32.54 Mb. The assembly demonstrates reference-grade quality, supported by a consensus quality value (QV) of 45.41 and a long terminal repeat assembly index (LAI) of 10.04. BUSCO analysis of the genome revealed 97.6% complete alignments, confirming the assembly's high continuity and completeness. Genome annotation identified 24,879 protein-coding genes, with repetitive sequences comprising 57.51% of the genome. BUSCO assessment of gene models further validated 97.2% completeness. This high-quality T2T genome assembly not only elucidates the genomic features of this early-diverging monocot species but also provides a robust foundation for future functional genomics research.**

Background & Summary

Acorus tatarinowii Schott (2n = 24), a perennial herb within the Acoraceae family, grows in wetlands and streams in Eastern and Southern Asia[1]. Renowned for its medicinal properties, this species has been traditionally used in Chinese medicinal to treat various ailments, including depression, epilepsy, fever, dizziness, heartache, and gastrointestinal disorders[2–4]. In recent decades, *A. tatarinowii* has garnered significant research attention due to its pharmacological potential[4–8]. Over 160 natural products—including phenylpropanoids, terpenoids, lignans, and alkaloids—have been isolated and characterized from this plant[9]. Beyond its medicinal applications, *A. tatarinowii* is also employed in artificial wetlands for phytoremediation, leveraging its rapid growth, large biomass production, anti-algae properties, and nutrient absorption capacity[3,10].

Recently, molecular genetic studies have been conducted on *A. tatarinowii* to assess its genetic diversity[11] and assemble its genome[12]. Because *Acorus* holds a pivotal phylogenetic position as the sister lineage to all other monocots, the genomes of *Acorus* species are important for exploring the evolution of early monocots[12–14]. With the recent rapid advances in genome sequencing technology, there has been a surge in deciphering gapless telomere-to-telomere (T2T) genomes for more and more species[15–18]. A chromosome-level genome assembly of *A. tatarinowii* was recently published[12]. However, a complete T2T genome assembly of *A. tatarinowii* has not yet been reported.

Here, we assembled a gapless T2T genome of *A. tatarinowii* using a combination of Pacific Biosciences (PacBio) HiFi sequencing, Oxford Nanopore Technologies (ONT) ultralong sequencing, and Hi-C technology. The final assembly spans 359.36 Mb and a scaffold N50 of 32.54 Mb (Fig. 1). Compared to the previously published genome assembly[12], our assembly demonstrates substantial improvements in completeness (97.6% vs. 92.4% of complete BUSCO groups) and accuracy (99.31% vs. 96.44% HiFi data mapping rate). This high-quality genome assembly and annotation not only provides critical insights into the evolutionary genomics of early monocots, but also establishes a robust foundation for elucidating the biosynthetic pathways of medicinally valuable compounds in *A. tatarinowii*.

[1]National Field Station of Freshwater Ecosystem of Liangzi Lake, College of Life Sciences, Wuhan University, Wuhan, P. R. China. [2]These authors contributed equally: Lei Huang, Zijie Wei. ✉e-mail: xuxw@whu.edu.cn

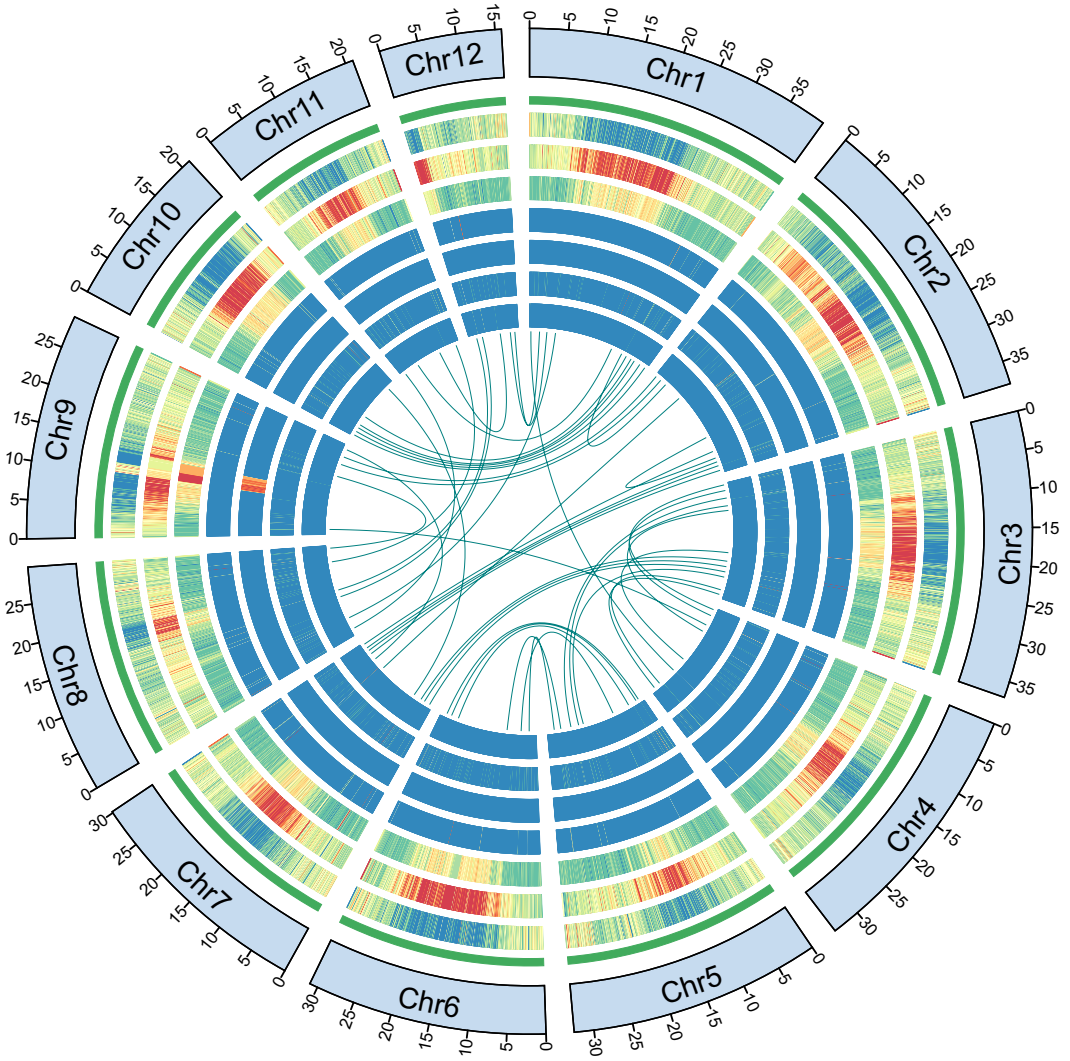

Fig. 1 Circos plot of the genome of *Acorus tatarinowii*. The circles, from the outermost to the innermost, represent: (I) chromosome length, (II) scaffold number of each chromosome, (III) gene density, (IV) TE density, (V) GC density, (VI) miRNA density, (VII) rRNA density, (VIII) snRNA density, (IX) tRNA density, and (X) genomic synteny.

Methods

Library construction and genome sequencing. The entire *A. tatarinowii* plant (NCBI Taxonomy ID: 123564)[19] was collected from a stream (111.0866° E, 25.2572° N) in the Doupangling Mountain, Hunan, China, in 2022. A specimen was deposited at the herbarium of Wuhan University (contact person: Xinwei Xu, xuxw@whu.edu.cn) under the voucher number HN20220102001.

We employed a multi-platform sequencing strategy to achieve comprehensive genome coverage and high-quality assembly. Initial short-read sequencing library was constructed using the VAHTS® Universal Plus DNA Library Prep Kit and sequenced on the DNBSEQ-T7 platform in paired-end 150 bp mode, generating 63 Gb of raw data primarily used for genome size estimation and error correction during *de novo* assembly. For ultra-long Oxford Nanopore technologies (ONT) sequencing, high-quality DNA was ultrasonically fragmented using a Covaris instrument. Libraries were then prepared using the SQK-LSK109 ligation kit and sequenced on the Oxford Nanopore PromethION platform, yielding 19.48 Gb of data with an average read length of 47,978 bp and N50 of 51,665 bp. For PacBio HiFi sequencing, SMRT cells with 15–20 kb insert fragments were constructed and sequenced on the PacBio Sequel II platform, generating 23.7 Gb of high-fidelity data with an N50 of 17,279 bp. Finally, chromosome-scale scaffolding was achieved through Hi-C chromatin conformation capture using DpnII restriction enzyme digestion following Belton *et al.*[20], with libraries PCR-amplified and sequenced on the DNBSEQ-T7 platform to generate 56.03 Gb of raw data.

To obtain a comprehensive transcriptome profile, fresh tissues from the root, stem, leaf, and flower of *A. tatarinowii* were pooled for total RNA extraction using the RNA prep Pure Plant Plus Kit (Tiangen Biotech, Beijing, China). Full-length cDNA libraries were constructed via strand-switching reverse transcription, followed by

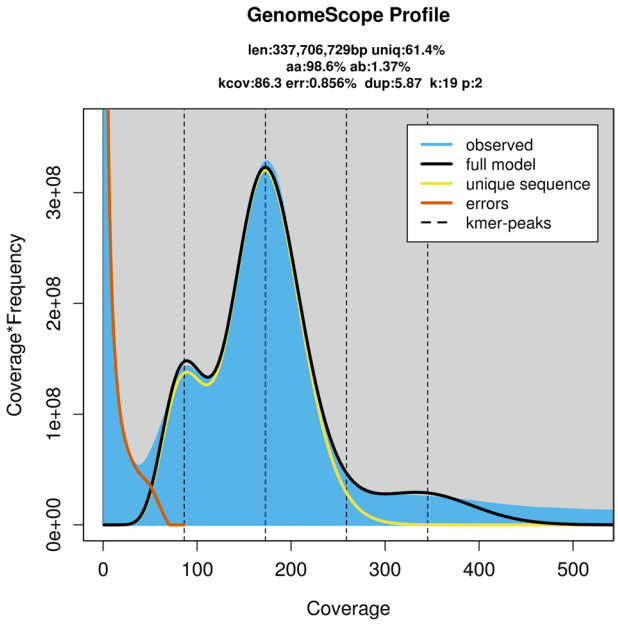

Fig. 2 *K*-mer frequency spectrum and genome characteristics estimated by GenomeScope2.

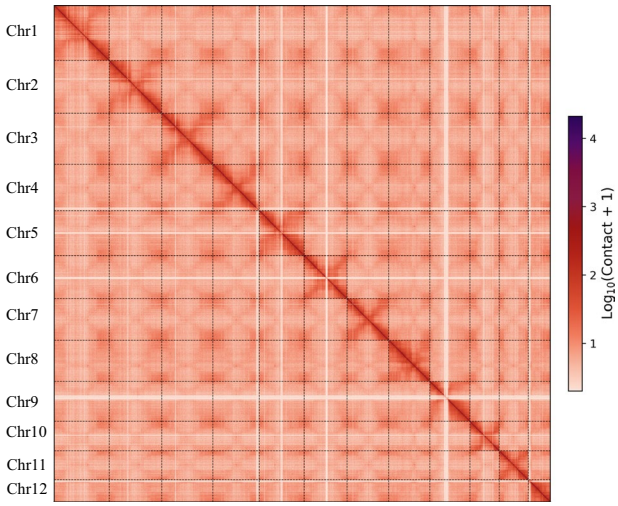

Fig. 3 The genome-wide Hi-C heatmap illustrates the interaction frequencies among different chromosomes of *Acorus tatarinowii*.

PCR amplification with the cDNA-PCR Sequencing Kit (SQK-PCS109). Sequencing was performed on the Oxford Nanopore PromethION platform, yielding 7.68 Gb of raw data.

Genome survey. To assess genome characteristics, we performed a *K*-mer analysis[21] on short-read sequencing data. Raw reads were quality-filtered using fastp v0.24.0[22] with default parameters, and clean reads were processed with Jellyfish v2.2.10[21] for *K*-mer ($K = 19$) frequency distribution profiling. Genome size and heterozygosity was estimated using GenomeScope v2.1.0[23], revealing an approximate genome size of 337 Mb and a heterozygosity rate of 1.37% (Fig. 2).

Genome assembly. To achieve a T2T genome assembly of *A. tatarinowii*, we integrated ultra-long ONT and PacBio HiFi sequencing data. Raw reads from both platforms were filtered using Nanofilt v2.8.0[24] with the parameters: -q 10 -l 500. The ultra-long ONT data were initially assembled using Nextdenovo v2.5.015, while PacBio HiFi data were processed using hifiasm v0.19.5[25], yielding a preliminary assembly (381.63 Mb, N50 = 24.44 Mb, 284 contigs).

To refine the assembly, we removed duplicate sequences and reduced heterozygosity using Purge_haplotigs v1.0.4[26] and Purge_dups v1.2.5[27]. Contaminating chloroplast and mitochondrial sequences were identified and filtered out via minimap2 v2.17-r94121[28] (aligning against organellar genomes; >50% identity matches removed; Table S2, see supplementary xlsx file).

ID	length	Contig number	ID	length	Gap number
Chr1	39809690	1	Chr1	39931050	0
Chr2	37905596	2	Chr2	38001515	0
Chr3	36966297	2	Chr3	37062551	0
Chr4	33538814	1	Chr4	33643077	0
Chr5	32415925	2	Chr5	32535662	0
Chr6	31020328	1	Chr6	31067262	0
Chr7	30079287	1	Chr7	30161589	0
Chr8	29678992	2	Chr8	29749569	0
Chr9	29078875	1	Chr9	29187282	0
Chr10	21152744	1	Chr10	21192928	0
Chr11	20929042	1	Chr11	21001344	0
Chr12	15784882	1	Chr12	15829076	0
chrUnn	2215627	4			

Table 1. Assembly statistics of each chromosome before and after gap filling in the *Acorus tatarinowii* genome.

Type	*De novo* + Repbase		TE Proteins		Combined TEs	
	Length (bp)	%	Length (bp)	%	Length (bp)	%
DNA	10,771,578	2.98	2,270,762	0.63	11,080,109	3.06
LINE	6,938,612	1.92	2,242,635	0.62	7,470,667	2.07
SINE	12,220	0.00	0	0.00	12,220	0.00
LTR	90,955,257	25.16	19,716,380	5.45	91,643,645	25.35
LTR-Gypsy	58,190,201	16.09	12,272,386	3.39	58,404,602	16.15
LTR-Copia	13,754,347	3.80	7,050,498	1.95	13,977,851	3.87
Satellite	265,451	0.07	0	0.00	265,451	0.07
Simple repeat	20,168	0.01	0	0.00	20,168	0.01
Other	519	0.00	0	0.00	519	0.00
Unknown	101,524,657	28.08	9,159	0.00	101,533,816	28.08
Total	202,340,428	55.96	24,234,309	6.70	207,931,278	57.51

Table 2. Statistics of repetitive sequence annotation result.

For chromosome-level scaffolding, Hi-C reads were filtered using fastp v0.24.0[22], aligned to contigs using HiCUP v0.8.0[29], and scaffolded with ALLHIC v0.9.8[30]. Chromosome contact maps were generated with Juicer v1.6[31] and 3D-DNA v201008[32], followed by manual curation in JuiceBox v1.11.08[33] to correct misassemblies (Fig. 3).

Gaps were filled using Winnowmap v1.11[34] by aligning the HiFi and ONT data to the Hi-C validated, chromosome-level genome. The gap-filling process followed a prioritization scheme of HiFi data over ONT data. Telomeric regions were extended using ONT reads, resulting in a gap-free (0 gaps) genome. The final genome assembly of *A. tatarinowii* spans 359.36 Mb across 12 chromosomes (Fig. 1, Table 1), with a scaffold N50 of 32.54 Mb.

Repetitive sequence annotation. To comprehensively analyse the repetitive sequences in the *A. tatarinowii* genome, we integrated *de novo* prediction with homology-based searches against the Repbase database[35]. A custom repeat library was constructed using RepeatModeler v1.0.11[36], LTR_FINDER v1.1[37], and LTR_retriever v3.0.4[38], followed by comprehensive annotation with RepeatMasker v4.0.6[39] in conjunction with the Repbase library. RepeatProteinMask within RepeatMasker was additionally employed to identify transposable elements (TE) protein-type repetitive sequences. After merging and deduplication, we identified 207.93 Mb of repetitive sequence (57.51% of the genome), with long terminal repeats (LTRs, 25.35%), long interspersed nuclear elements (LINEs, 2.07%), and satellite (0.07%) representing the major components (Fig. 1, Table 2).

Gene prediction and functional annotation. We employed an integrative approach combining transcriptomic, homology-based, and *ab initio* methods for comprehensive gene prediction. For transcriptional evidence, TransDecoder v5.5.0 (https://github.com/TransDecoder) was used to identify protein-coding genes based on full-length RNA-seq data processed through the HISAT2-StringTie v2.2.0 pipeline[40,41]. Homology-based predictions were performed with BLAST v2.13.0[42] and Exonerate v2.4.0[43] against four monocot reference genomes: *Acorus americanus* (https://phytozome-prod.jgi.doe.gov/info/Aamericanus_v1_1), *Zostera marina* (GCA_001185155.1), *Zea mays* (B73 RefGen_v4), and *Oryza sativa* (GCF_000005425). For *de novo* prediction, Augustus v3.5.0[44] and GlimmerHMM v3.0.4[45] were applied to the repeat-masked genome. These predictions were integrated using Maker v3.01.03[46], yielding a final gene set with an average of 5.42 exons per gene and a mean exon length of 283.38 bp.

Item	Count	Percentage
All	24,879	100.00%
Annotation	22,833	91.78%
KEGG	5,933	23.85%
Pathway	4,864	19.55%
NR	20,919	84.08%
Uniprot	20,893	83.98%
GO	12222	49.13%
Pfam	16312	65.57%
Interpro	22523	90.53%

Table 3. Functional annotation results of protein-coding genes.

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		102	117	11,908	0.003293
tRNA		500	75	37,410	0.010346
rRNA	rRNA	2,863	200	573,441	0.158594
	18S	132	1774	234,161	0.064761
	28S	127	196	24,949	0.006900
	5.8S	129	157	20,253	0.005601
	5S	2,475	119	294,078	0.081332
snRNA	snRNA	534	113	60,160	0.016638
	CD-box	308	102	31,332	0.008665
	HACA-box	56	133	7,453	0.002061
	splicing	170	126	21,375	0.005912
	scaRNA	0	0	0	0.000000

Table 4. Annotation results of noncoding RNAs.

Item	Assembly		Annotation	
	Proteins	Percent (%)	Proteins	Percent (%)
Complete BUSCOs (C)	1,575	97.6	1,572	97.2
Complete and single-copy BUSCOs (S)	1,513	93.7	1,511	93.6
Complete and duplicated BUSCOs (D)	62	3.8	58	3.6
Fragmented BUSCOs (F)	2	0.1	5	0.3
Missing BUSCOs (M)	37	2.3	40	2.5
Total BUSCO groups searched	1,614	100.0	1614	100.0

Table 5. BUSCO assessment results of the *Acorus tatarinowii* genome assembly and annotation.

Functional gene annotation was performed by aligning protein sequences against various databases based on sequence and motif similarity. Diamond v2.0.15.153[47] was used to search Uniprot[48], NR[49], GO[50], and KOG databases, applying a stringent e-value threshold of 10^{-5}. KOBAS v3.0[51] was employed to associate potential genes with the KEGG[52] pathway by ID. Motifs and domains were identified by searching the InterPro[53] and Pfam[54] databases using InterProScan v5.55–88.0[55]. This comprehensive analysis functionally annotated 22,833 genes that exhibited specific functions across various metabolic pathways (Table 3).

For noncoding RNA prediction, tRNAscan-SE v2.0.12[56] and RNAmmer v1.2[57] were employed to predict tRNAs and rRNAs, respectively. Other types of noncoding RNAs, including small snRNAs and miRNAs, were identified using Infernal v1.1.4[58] based on the Rfam database[59]. We identified 500 tRNAs, 2,863 rRNAs, 534 snRNAs, and 102 miRNAs (Table 4), providing a foundation for further functional studies in *A. tatarinowii*.

Data Records

The raw sequencing data generated in this study have been deposited in the Genome Sequence Archive (GSA) at the National Genomics Data Centre (NGDC), part of the China National Centre for Bioinformation (CNCB)/Beijing Genomics Institute, Chinese Academy of Sciences, under the accession number CRA016787[60]. The associated BioProject is registered under the accession number PRJCA026616[61].The assembled genome sequences have been deposited in the NCBI GenBank with the accession number JBPSQV000000000.1[62]. The genome assembly and annotation files have been deposited in the Genome Warehouse (GWH) at the National Genomics Data Centre (NGDC) under the accession number GWHGGEW00000000.1[63], as part of BioProject PRJCA026616[61], and are also available in the Figshare database[64].

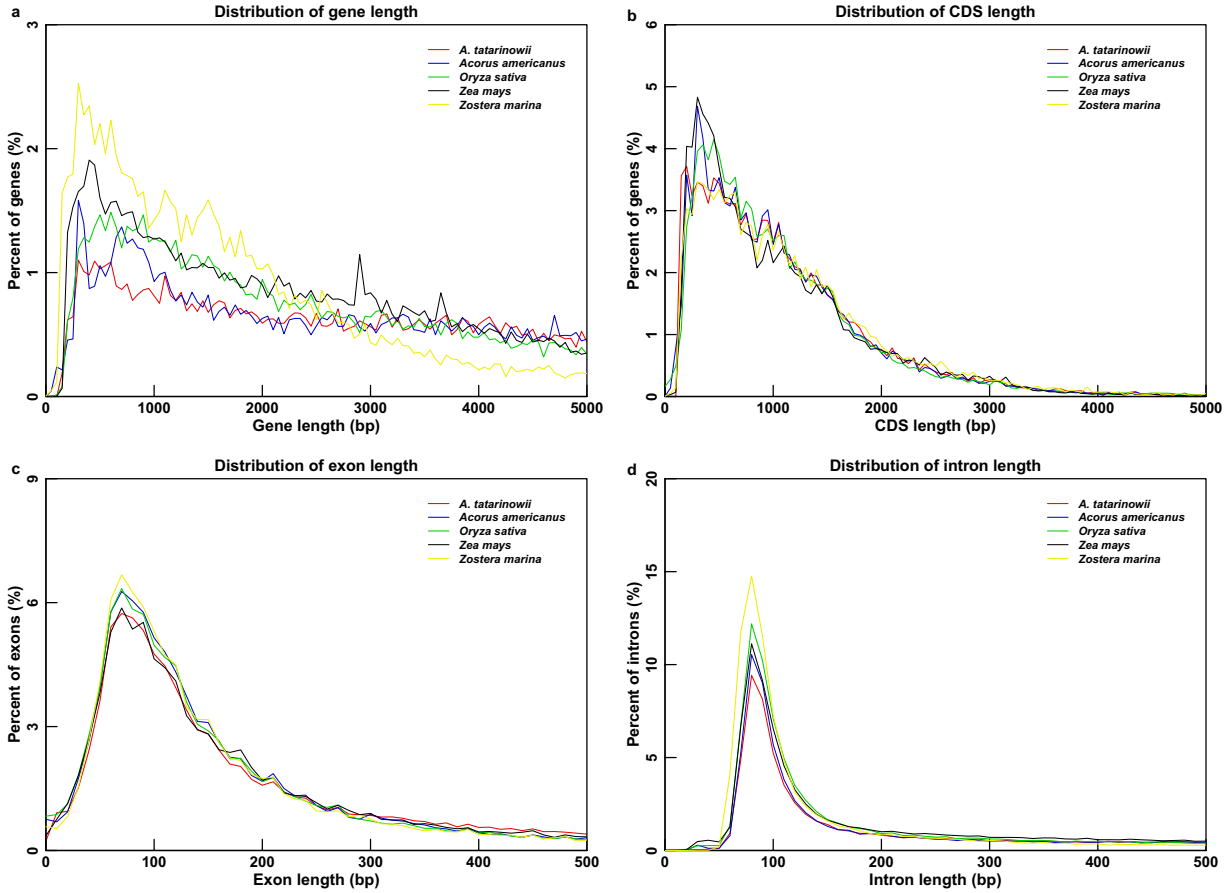

Fig. 4 Distribution of gene element lengths across *Acorus tatarinowii*, *Acorus americanus*, *Oryza sativa*, *Zea mays*, and *Zostera marina*. (**a**) Distribution of gene length. (**b**) Distribution of CDS length. (**c**) Distribution of exon length. (**d**) Distribution of intron length.

Item	Start	End	QV	Error rate
Genome	196403	359362689	45.4103	2.88E-05
Chr1	23167	39931032	45.1507	3.05E-05
Chr2	17680	38001497	46.1098	2.45E-05
Chr3	16975	37062533	46.1778	2.41E-05
Chr4	18093	33643059	45.4803	2.83E-05
Chr5	25932	32535644	43.7711	4.20E-05
Chr6	11652	31067244	47.0458	1.97E-05
Chr7	18343	30161571	44.9461	3.20E-05
Chr8	18361	29749551	44.8821	3.25E-05
Chr9	14236	29187264	45.9046	2.57E-05
Chr10	10242	21192910	45.9446	2.54E-05
Chr11	11673	21001326	45.337	2.93E-05
Chr12	10049	15829058	44.7596	3.34E-05

Table 6. Consensus quality value (QV) of the *Acorus tatarinowii* genome assembly.

Technical Validation

The genome assembly of *A. tatarinowii* was comprehensively evaluated using multiple quality assessment approaches. The assembly achieved complete chromosomal with zero gaps across all 12 chromosomes (Fig. 1), demonstrating exceptional continuity as evidenced by a scaffold N50 of 32.54 Mb. The embryophyta_odb10 BUSCO v4.1.4[65] assessment revealed 1,564 complete core genes (97.6%) out of 1,614 benchmarked orthologs (Table 5), confirming high assembly completeness. The BUSCO evaluation demonstrated high annotation quality for the gene set, showing 97.2% BUSCO completeness (comprising 93.6% single-copy and 3.6% duplicated complete BUSCOs) (Table 5). Annotation quality was also assessed using the agat_sp_statistics.pl script from AGAT package v1.0.0[66], identifying 24,596 protein-coding genes

ID	upstream	downstream	UpStream_TelStart	UpStream_TelEnd	DownStream_TelStart	DownStream_TelEnd
Chr1	1250	1658	1	11275	39914892	39931047
Chr2	2494	1707	7	19403	37983640	38001509
Chr3	1461	156	1	14201	37045469	37062550
Chr4	2795	490	7	19995	33634550	33643077
Chr5	1181	870	4	8717	32517684	32535656
Chr6	1	1631	9666	9672	31048923	31067259
Chr7	801	1026	2	15670	30153929	30161589
Chr8	227	1873	1	10358	29734813	29749566
Chr9	997	2758	7	19297	29167283	29187279
Chr10	2814	2563	3	19868	21174347	21192927
Chr11	410	429	3	14254	20988676	21001307
Chr12	1122	726	1	9437	15822262	15829076

Table 7. Quantification and genomic coordinates of telomeric repeats at the ends of chromosomes in *Acorus tatarinowii*. *Note:* Telomeric regions were identified based on the conserved plant telomeric repeat sequences: 5′-CCCTAAA and 3′-TTTAGGG.ID refers to the chromosome number; UpStream and DownStream represent the number of telomeric repeat units detected at the upstream and downstream ends of each chromosome, respectively. UpStream_TelStart and UpStream_TelEnd indicate the genomic coordinates of the upstream telomeric region, while DownStream_TelStart and DownStream_TelEnd denote the coordinates of the downstream telomeric region.

with well-defined structural features: 59.4% of mRNAs contained at least one UTR (52.7% possessed both 5′ and 3′ UTRs), while genes exhibited an average of 5.4 exons and 4.2 introns per mRNA, with 19.2% being single-exon genes (Table S1, see supplementary xlsx file). Comparative genomic analysis showed highly consistent distribution patterns of gene structures (CDS length, exon length, and intron length) in homologous species (Fig. 4). The assembly achieved reference-grade quality as indicated by an LTR assembly index (LAI) score of 10.04 (Table S3, see supplementary xlsx file), obtained using LTR_retriever v3.0.4[38] and an overall consensus quality value (QV) of 45.41 (Table 6 and Figure S1), obtained using Merqury v1.3.6[67]. Sequencing data alignment confirmed exceptional accuracy, with 99.82% of ONT ultralong reads and 99.31% of PacBio HiFi reads successfully mapping to the assembly. Telomere analysis identified complete terminal sequences on all chromosomes (Table 7), while Hi-C interaction heatmaps showed no aberrant signals (Fig. 3). The assembly was validated by BlobTools v1.1.1[68], which detected no substantial non-plant contamination (Figure S2). These comprehensive evaluations collectively demonstrate that the *A. tatarinowii* genome assembly exhibits outstanding metrics in continuity, completeness, accuracy and structural integrity, providing a reliable genomic resource for future studies.

Data availability

The Genome Survey Sequencing data, the ultra-long ONT sequencing data, the Hi-C sequencing data, the HiFi sequencing data, and the RNA sequencing data were deposited in Genome Sequence Archive (GSA) of NGDC with accession number CRR1170421[69], CRR1170419[70], CRR1170417[71], CRR1170418[72] and CRR1170420[73]. The associated BioProject is registered under the accession number PRJCA026616[61].The assembled genome sequences have been deposited in the NCBI GenBank with the accession number JBPSQV000000000.1[62]. The genome assembly and annotation files have been deposited in the Genome Warehouse (GWH) at the National Genomics Data Centre (NGDC) under the accession number GWHGGEW00000000.1[63], and are also available in the Figshare database (https://doi.org/10.6084/m9.figshare.29764529.v3)[64].

Code availability

Custom software code was not developed in this study. All bioinformatics tools and pipelines were executed according to the manuals and protocols provided by their respective software developers. The software versions and their corresponding parameters used have been thoroughly described in the "Methods" section.

Received: 2 March 2025; Accepted: 4 September 2025;
Published online: 13 October 2025

References

1. Hen, L. Flora reipublicae popularis sinicae. *Inst Bot Kummingense Acad Sin* **13**, 1–12 (1979).
2. Cheng, Z. *et al*. From Folk Taxonomy to Species Confirmation of Acorus (Acoraceae): Evidences Based on Phylogenetic and Metabolomic Analyses. *Front. Plant Sci.* **11**, 965 (2020).
3. Zhao, Y. *et al*. Ethnic, Botanic, Phytochemistry and Pharmacology of the Acorus L. Genus: A Review. *Mol. Basel Switz.* **28**, 7117 (2023).
4. Wang, M. *et al*. *Acorus tatarinowii* Schott: A Review of Its Botany, Traditional Uses, Phytochemistry, and Pharmacology. *Mol. Basel Switz.* **28**, 4525 (2023).
5. Zhu, H. *et al*. Extraction and purification of cis/trans asarone from *Acorus tatarinowii* Schott: Accelerated solvent extraction and silver ion coordination high-speed counter-current chromatography. *J. Chromatogr. A* **1643**, 462080 (2021).

6. Zhong, J., Qiu, X., Yu, Q., Chen, H. & Yan, C. A novel polysaccharide from *Acorus tatarinowii* protects against LPS-induced neuroinflammation and neurotoxicity by inhibiting TLR4-mediated MyD88/NF-κB and PI3K/Akt signaling pathways. *Int. J. Biol. Macromol.* **163**, 464–475 (2020).
7. Pu, D. *et al*. Extraction and *in vitro* active evaluation of essential oil of *Acorus tatarinowii* Schott rhizome rich in β-asarone using enzymatic pretreatment and solvent-free microwave-assisted method. *J. Essent. Oil Bear. Plants* **26**, 1563–1575 (2023).
8. Huang, L. *et al*. Chemical constituents from the rhizome of *Acorus tatarinowii* Schott and their anti-inflammatory activities in LPS induced PC12 cells. *Pharmacol. Res. - Mod. Chin. Med.* **8**, 100269 (2023).
9. Yang, L. *et al*. A review of the botany, traditional uses, phytochemistry, and pharmacology of the Flos Inulae. *J. Ethnopharmacol.* **276**, 114125 (2021).
10. He, C. & Wang, C. Allelopathic effect of *Acorus tatarinowii* upon algae. *J. Environ. Sci. China* **13**, 481–484 (2001).
11. Qin, Q. *et al*. Assessment of genetic diversity and construction of core germplasm in populations of *Acorus tatarinowii* based on SNP markers. *J. Appl. Res. Med. Aromat. Plants* **44**, 100605 (2025).
12. Shi, T. *et al*. The slow-evolving *Acorus tatarinowii* genome sheds light on ancestral monocot evolution. *Nat. Plants* **8**, 764–777 (2022).
13. Guo, X. *et al*. The genome of Acorus deciphers insights into early monocot evolution. *Nat. Commun.* **14**, 3662 (2023).
14. Ma, L. *et al*. Diploid and tetraploid genomes of Acorus and the evolution of monocots. *Nat. Commun.* **14**, 3661 (2023).
15. Huang, G. *et al*. A telomere-to-telomere cotton genome assembly reveals centromere evolution and a Mutator transposon-linked module regulating embryo development. *Nat. Genet.* **56**, 1953–1963 (2024).
16. Bi, G. *et al*. Near telomere-to-telomere genome of the model plant Physcomitrium patens. *Nat. Plants* **10**, 327–343 (2024).
17. Deng, Y. *et al*. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* **15**, 1268–1284 (2022).
18. Zhang, Y. *et al*. Telomere-to-telomere Citrullus super-pangenome provides direction for watermelon breeding. *Nat. Genet.* **56**, 1750–1761 (2024).
19. Schoch, C. L. *et al*. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
20. Belton, J.-M. *et al*. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k -mers. *Bioinformatics* **27**, 764–770 (2011).
22. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
23. 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 (2020).
24. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
25. 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).
26. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
27. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
29. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
30. 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 (2019).
31. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
32. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
33. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
34. Jain, C. *et al*. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
35. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. Dna* **6**, 1–6 (2015).
36. Smit, A. F. A., Hubley, R. & Green, P. RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology. (2015).
37. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
38. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126 (2018).
39. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **5**, 4.10. 1–4.10. 14 (2004).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
41. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
42. Camacho, C. *et al*. BLAST plus: architecture and applications. *BMC Bioinformatics* **10**, 10–1186.
43. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
44. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
45. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
46. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
48. Apweiler, R. *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119 (2004).
49. 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–D65 (2007).
50. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
51. Bu, D. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).
52. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
54. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
55. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
56. 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 (2021).
57. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).

58. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
59. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
60. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787 (2025).
61. *NGDC BioProject* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA026616 (2025).
62. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051622315.1 (2025).
63. *NGDC Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/98806/show (2025).
64. Huang, L. *et al*. Genome Assembly and Annotation Files of *Acorus tatarinowii* Schott. *figshare* https://doi.org/10.6084/m9.figshare.29764529.v3 (2025).
65. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323 (2021).
66. Dainat, J., Hereñú, D. & Pucholt, P. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. *Zenodo*. (2021).
67. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
68. Laetsch, D. R. & Blaxter, M. L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
69. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787/CRR1170421 (2025).
70. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787/CRR1170419 (2025).
71. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787/CRR1170417 (2025).
72. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787/CRR1170418 (2025).
73. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA016787/CRR1170420 (2025).

Acknowledgements

This study was supported by the National Water Pollution Control and Treatment Science and Technology Major Project, China (No. 2015ZX07503005). The calculations in this paper were performed using the supercomputing system at the Supercomputing Center of Wuhan University.

Author contributions

X.X. designed the research; Z.W. carried out the field collections; Z.W. carried out the experiments and performed the data analysis; L.H. drafted the manuscript. X.X. revised the manuscript. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05933-6.

Correspondence and requests for materials should be addressed to X.X.