scientific data

Check for updates

OPEN

DATA DESCRIPTOR

High quality chromosome-level genome assembly of *Psammosilene tunicoides* (Caryophyllaceae), an endangered medicinal plant

Zongyi Xie[1,2,4], Yingmin Zhang[1,2,4], Congwei Yang[2], Ticao Zhang[3] ✉, Aili Zhang[1,2] ✉ & Guodong Li[1,2] ✉

***Psammosilene tunicoides*, a medicinal plant with significant anti-inflammatory and analgesic properties, is currently classified in China as a Vulnerable species and a Grade II Protected Wild Plant. In this study, a chromosome-level *de novo* genome assembly of *P. tunicoides* was performed for the first time, yielding a total genome size of approximately 1,463,250,424 bp, with a contig N50 of 91,918,445 bp and a scaffold N50 of 103,834,185 bp. Approximately 94.40% of the assembled sequences were anchored onto 14 pseudochromosomes. A total of 30,924 genes were predicted, with more than 95% of the genes annotated. Repetitive elements account for 83.40% of the genome. BUSCO analysis indicated high completeness for both the assembly and annotation, with scores of 98.40% and 98.60%, respectively. This high-quality genomic resource establishes a foundation for exploring adaptive evolution and genomic diversity in *P. tunicoides* and its relatives, while also supporting genetic improvement, resource conservation, and utilization initiatives.**

Background & Summary

Psammosilene tunicoides W. C. Wu & C. Y. Wu (2n = 2X = 28[1,2]) is a perennial herb in the family Caryophyllaceae and the sole species of its genus, native to Sichuan, Yunnan, Guizhou, and Tibet[3]. As a commonly used herb in traditional Chinese medicine from southwestern China, the dried roots of this plant are used for their purported medicinal properties and are recorded in the Chinese Pharmacopoeia[4] (Fig. 1A–C). Modern pharmacological studies suggest that extracts of this plant can exhibit analgesic, anti-inflammatory, and antioxidant properties[5–7]. *P. tunicoides* has gained considerable attention as a key component in certain proprietary Chinese medicines used to treat orthopaedic injuries[8,9].

During cultivation, this species is susceptible to root rot, resulting in unsatisfactory yields that do not meet market demands[10,11]. Consequently, overharvesting driven by commercial needs has led to the continuous decline of wild *P. tunicoides* populations[12]. In 2021, it was listed as a National Grade II Protected Wild Plant in the List of National Key Protected Wild Plants[13], and in 2023, *P. tunicoides* was classified as Vulnerable in the China Biodiversity Red List[14]. These threats underscore the urgency of conducting conservation breeding and sustainable resource management for this species. Unfortunately, the severe lack of fundamental research currently hinders the advancement of related conservation efforts.

This study employed an integrated approach combining PacBio HiFi long-read sequencing, next-generation sequencing (NGS), and High-throughput Chromosome Conformation Capture (Hi-C) to assemble the first high-quality, chromosome-level genome of *P. tunicoides*. The length of the final assembled genome was 1,463,250,424 bp, contig N50 of 91,918,445 bp, scaffold N50 of 103,834,185 bp, and anchors 94.40% of the sequences onto 14 putative chromosomes. A total of 30,924 genes were predicted, and repetitive sequences accounted for 83.40% of the genome. As the first high-quality reference genome for this species, this resource

[1]Yunnan Key Laboratory of Dai and Yi Medicines, Yunnan University of Chinese Medicine, Kunming, 650500, China. [2]College of Chinese Material Medica, Yunnan University of Chinese Medicine, Kunming, 650500, China. [3]State Key Laboratory of Phytochemistry and Natural Medicines, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. [4]These authors contributed equally: Zongyi Xie, Yingmin Zhang. ✉e-mail: zhangticao@mail.kib.ac.cn; zhangailibb@126.com; gammar116@163.com

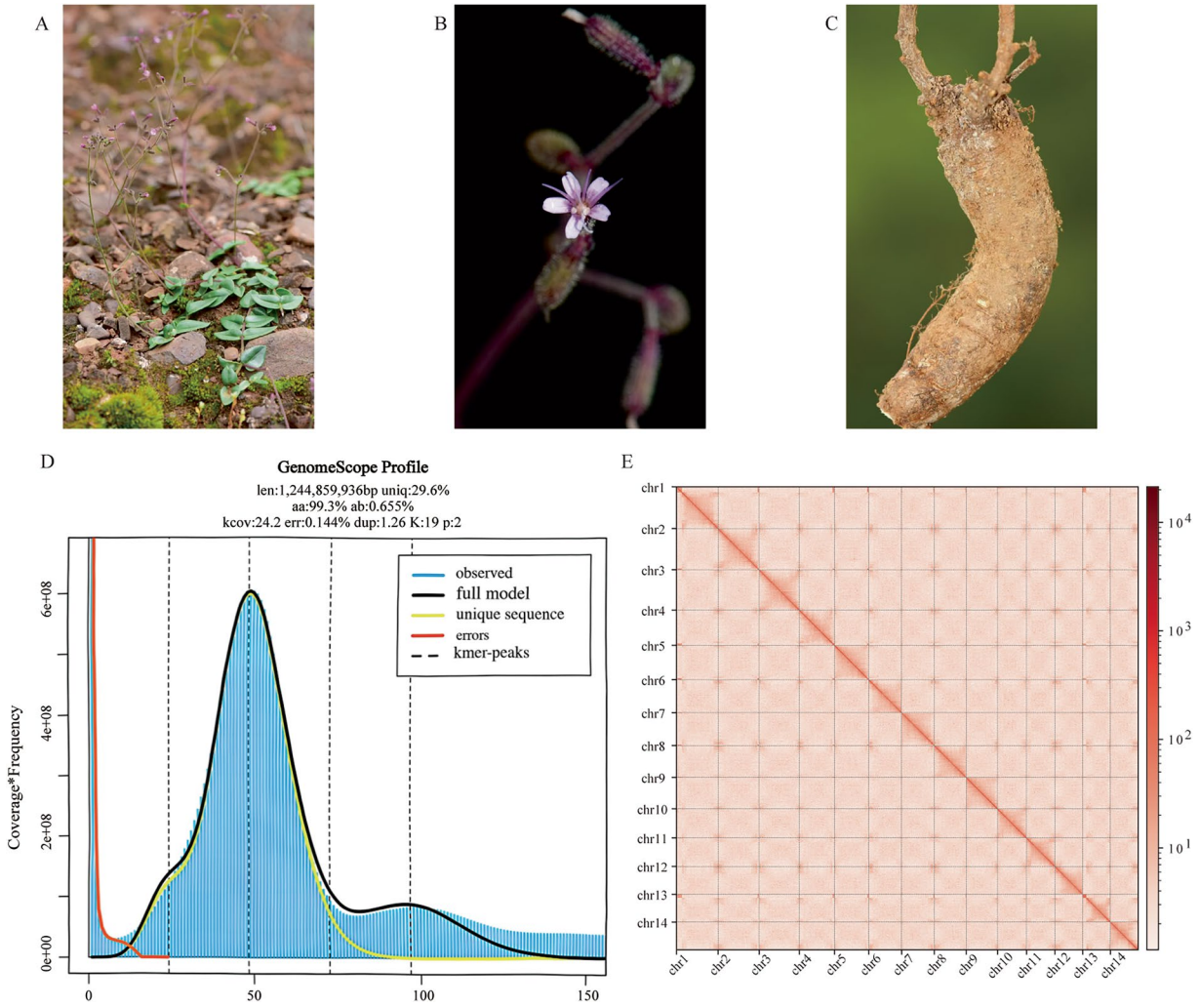

Fig. 1 Morphological characteristics and genome assembly evaluation of *Psammosilene tunicoides*. (**A**) Leaf of *P. tunicoides*; (**B**) Inflorescence of *P. tunicoides*; (**C**) Root of *P. tunicoides*; (**D**) Genome survey of *P. tunicoides* based on K-mer analysis (K = 19); (**E**) Hi-C interactive heatmap of *P. tunicoides* genome.

establishes a foundation for future research into the genetic improvement, resource protection, and sustainable utilization of *P. tunicoides*.

Methods

Sample collection and DNA sequencing. Samples of *P. tunicoides* were collected on 21 March 2025 from Maʾer Mountain, Jianchuan County, Dali Prefecture, Yunnan Province, China (100°3′N, 26°31′E). A voucher specimen (ID: 532931089) was deposited in the herbarium of Yunnan University of Chinese Medicine. Immediately after collection, fresh young leaves were frozen in liquid nitrogen and stored at −80 °C prior to DNA extraction. High-quality genomic DNA was extracted from leaf tissues using the DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA).

A considerable number of short-read sequences were generated through NGS on the DNBSEQ-T7 platform (MGI Tech, Shenzhen, China) following DNA library preparation. Quality control of the sequencing data was performed using fastp v0.21.0[15], which involved the removal of reads with N base content exceeding 5%, reads with more than 50% low-quality bases, reads with adapter contamination, and PCR duplicates to ensure data accuracy. This resulted in a final dataset of 75.67 Gb of clean short reads.

High-quality HiFi reads were obtained using PacBio single-molecule sequencing after the construction of a SMRTbell library. Following quality control filtering with SMRTLink v13.0 (https://www.pacb.com/smrt-link/), Polymerase reads were removed based on several criteria including a length of less than 50 bp, a quality value below 0.8, the presence of self-ligated adapters, or internal adapter sequences. This process yielded 51.24 Gb of clean HiFi reads.

To construct the Hi-C sequencing library, chromatin was first crosslinked with formaldehyde, followed by digestion with the DpnII restriction enzyme. Subsequently, biotinylated nucleotides were used to end-label and gap-fill the resulting DNA fragments. Proximity ligation was then performed to join interacting fragments, after which crosslinks were reversed and the DNA was purified. The purified DNA was sheared to a target size of

Item	Contig	Scaffold
Total number	945	14
total length without N (bp)	1,550,221,676	1,463,248,324
Total length (bp)	1,550,221,676	1,463,250,424
N90 (bp)	42,552,916	87,821,861
N50 (bp)	91,918,445	103,834,185
Min (bp)	14,928	87,724,098
Median (bp)	38,479	102,329,461.50
Max (bp)	146,408,198	131,987,497
GC content	35.74%	35.39%
Average (bp)	1,640,446.22	104,517,887.43

Table 1. Summary of the *Psammosilene tunicoides* genome assembly.

200–400 bp. Following end-repair, A-tailing, and adapter ligation, biotinylated fragments were enriched using streptavidin magnetic beads and amplified by PCR. Hi-C reads were sequenced on the Illumina NovaSeq 6000 platform (paired-end, 150 bp reads) after DNA library construction. The clean data were generated by filtering raw reads with fastp v0.23.2, which eliminated adapter sequences and low-quality reads, resulting in a total of 144.82 Gb of clean Hi-C reads.

Contamination screening. To assess potential contamination in the sequencing data, the first 50,000 reads were sampled from the NGS data and aligned against the NT database using blastn v2.11.0 (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.11.0/). The results indicated that 15.06% of the sampled reads (7,532 reads) had significant hits in the NT database. The target species, *P. tunicoides*, accounted for the highest number of aligned reads (2,476). The dominance of the target species suggests a low level of contamination, supporting the suitability of the data for subsequent analyses.

Genome survey. K-mer analysis was used to estimate the genome size and heterozygosity. The K-mer frequency and depth distribution was generated from quality-controlled short reads using Jellyfish v2.2.10[16]. This distribution provided the basis for estimating genome size, which was further evaluated with GenomeScope v2.0[17]. The results indicate a genome size of approximately 1,244.86 Mb for *P. tunicoides*, with a heterozygosity of 0.66% (Fig. 1D).

Genome and chromosome assembly. The HiFi reads were assembled using the Overlap-Layout-Consensus (OLC) algorithm implemented in hifiasm v.0.19.8-r603[18]. Subsequently, Haplotigs were removed with Purge Dups v.1.2.5[19], resulting in a preliminary draft genome assembly. The assembled genome size was 1,550,221,676 bp, with a contig N50 of 91,918,445 bp (Table 1). During the assembly process, an all-versus-all alignment of reads is first performed to identify overlaps among the reads, thereby clustering them in a manner analogous to a clustering operation. Subsequently, three rounds of self-correction are performed on the overlapping reads within the clusters. Following error correction, an all-versus-all alignment is conducted again on the corrected reads to re-identify overlaps among them, thereby constructing a string graph.

To obtain the chromosome-level genome of *P. tunicoides*, Hi-C data were integrated to further optimize the polished assembly. Initially, the clean Hi-C reads were aligned to the draft genome using the HiCUP v0.8.2[20]. Uniquely mapped paired-end reads were retained, and the proportions of self-circles, dangling ends, dumped pairs, and valid interaction pairs were assessed for quality control. Only valid interaction pairs were retained for subsequent analysis. Subsequently, interaction analysis was performed on the valid interaction pairs. The results were used to determine the degree of association between different contigs, thereby enabling contig clustering. Finally, the ALLHIC v0.9.8[21] was employed to cluster contig sequences into distinct chromosomal groups using agglomerative hierarchical clustering. The contigs within each chromosomal group were then ordered and oriented. The interaction relationships between pairs of contigs were converted into specified binary files using the 3D-DNA v201008[22] and Juicer v1.6[23]. Manual ordering and orientation of the contigs were subsequently performed using Juicebox v1.11.08[24]. Unknown sequences between contigs were filled with Ns, resulting in the final chromosome-level genome sequence. The final genome assembly had a total length of 1,463,250,424 bp, with a scaffold N50 of 103,834,185 bp (Table 1). A total of 94.40% of the assembled sequences were anchored onto 14 pseudochromosomes (Figs. 1E, 2), with chromosome lengths ranging from 86,973,352 bp to 131,987,497 bp.

Repeat annotation. To comprehensively analyze the characteristics of repetitive elements in the genome of *P. tunicoides*, this study employed both homology-based and *de novo* prediction annotation approaches.

For homology-based annotation, the RepeatMasker subtool RepeatProteinMask v4.1.7 (https://www.repeatmasker.org/RepeatProteinMask.html) was utilized to identify repetitive sequences of the transposable element (TE) protein type, generating TE protein annotations.

For *de novo* annotation, RepeatModeler v2.0.6[25] was applied to generate model sequences based on the genome assembly itself. Additionally, LTR Finder[26] and LTR harvest v1.6.5[27] were used to predict long terminal repeat (LTR) sequences, and LTR Retriever v3.0.1[28] was employed to remove redundancies from the predictions generated by these two methods, resulting in a non-redundant LTR sequence set. The *de novo*-derived sequences

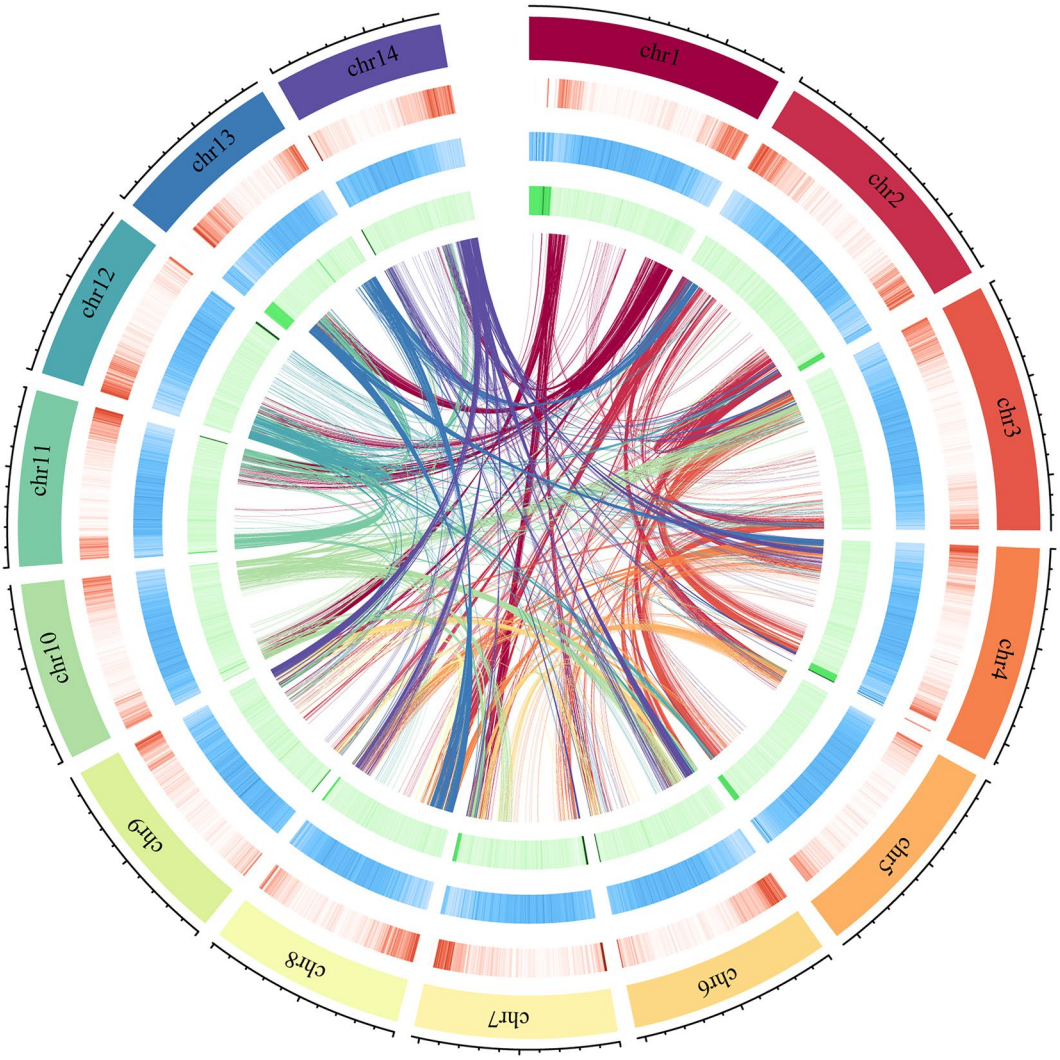

Fig. 2 Genomic features of *Psammosilene tunicoides*. From outer to inner circles: 14 chromosomes (chr1–chr14), gene density, repeat sequence density, GC content, and syntenic gene blocks represented by connecting lines within the genome (window size = 10 Mb).

were then merged to construct a *de novo* repeat library. Unknown elements within this library were re-annotated using the TEclass v2.1.3[29]. Subsequently, the RepBase library[30] and the *de novo* library were combined, and RepeatMasker v4.1.7 (https://www.repeatmasker.org/RepeatMasker/) was used to align and predict repetitive sequences, producing the "*De novo* + RepBase" results. Finally, all repetitive sequence predictions were merged and deduplicated to obtain the final set of genomic interspersed repeats, designated as the Combined TEs results.

The annotation indicated that the total length of repetitive sequences was approximately 1,220,316,436 bp, accounting for 83.40% of the entire genome. Among these, LTR retrotransposons constituted the largest proportion at 69.86% (1,022,186,873 bp), followed by DNA transposons at 9.87% (144,412,594 bp), and LINE elements at 2.83% (41,476,789 bp). This comprehensive analysis of repeat element characteristics provides an important foundation for studying genome evolution, structural variation, and gene regulatory mechanisms (Table 2).

Non-coding RNA annotation. To identify non-coding RNA in the genome of *P. tunicoides*, this study utilized several bioinformatics tools based on the structural characteristics of different RNA types. tRNA sequences were searched using tRNAscan-SE v2.0.12[31]. For rRNA prediction, the RNAmmer v1.2[32] was employed. Additionally, Infernal v1.1.4[33] was used, based on the Rfam database[34], to identify snRNA and miRNA in the genome. In total, 5,574 snRNAs, 3,730 rRNAs, 1,397 tRNAs, and 110 miRNAs were identified (Table 3).

Protein-coding gene prediction and functional annotation. This study employed a combination of transcriptome prediction, homology-based prediction, and *de novo* prediction methods for gene structure prediction. For transcriptome prediction, the StringTie v2.2.1[35] was used to merge predicted transcripts, followed by coding sequence prediction with TransDecoder v5.7.0 (https://github.com/TransDecoder/TransDecoder). In the homology prediction approach, protein sequences from five reference species (*Arabidopsis thaliana*[36], *Oryza sativa*[37], *Saponaria officinalis*[38], *Silene latifolia*[39], and *Amaranthus tricolor*[40]) were aligned to the *P. tunicoides*

Type	TE proteins	De novo + repbase	Combined TEs	
	Length (bp)	Length (bp)	Length (bp)	(%) in genome
DNA	959,562	144,388,577	144,412,594	9.87
LINE	3,498	41,476,789	41,476,789	2.83
SINE	0	2,068,393	2,068,393	0.14
LTR	187,718,706	1,007,360,106	1,022,186,873	69.86
LTR-Gypsy	146,080,907	461,105,483	476,086,212	32.54
LTR-Copia	40,979,518	67,897,862	78,391,313	5.36
Other	0	13,672	13,672	0
Unknown	0	47,964,040	47,964,040	3.28
Total	188,681,606	1,205,598,649	1,220,316,436	83.40

Table 2. Summary of the repetitive sequences in *Psammosilene tunicoides* genome assembly.

Type		Copy	Average length (bp)	Total length (bp)
miRNA		110	129	14,165
tRNA		1,397	75	104,452
rRNA	rRNA	3,730	594	2,214,677
	18S	318	1,824	580,002
	28S	320	3,882	1,242,283
	5.8S	712	154	109,927
	5S	2,380	119	282,465
snRNA	snRNA	5,574	110	613,073
	CD-box	5,135	108	553,065
	HACA-box	67	129	8,612
	splicing	372	138	51,396

Table 3. Annotation of non-coding RNAs in *Psammosilene tunicoides*.

Item	Number
Total number of genes	30,924
Average mRNA length (bp)	5,936.56
Average CDS length per gene (bp)	1,169.55
Average exon number per gene	5.22
Average exon length (bp)	336.57
Average intron length (bp)	989.62
Total number of exons	161,504
Total number of introns	130,580
Total intron length (bp)	129,224,715

Table 4. Summary of predicted protein-coding genes in *Psammosilene tunicoides*.

genome using the Miniprot v0.13[41] to obtain protein alignment evidence. Based on the repeat-masked *P. tunicoides* genome, *de novo* gene prediction was performed using Augustus v3.5.0[42] and GlimmerHMM v3.0.4[43] were utilized. Finally, the Maker/EVidenceModeler v3.01.03[44]/v2.1.0[45] was employed to integrate the gene sets obtained from the three methods, ensuring the accuracy and comprehensiveness of gene structure predictions. A total of 30,924 genes were predicted, with an average mRNA length of 5,936.56 bp, an average coding sequence length of 1,169.55 bp, and an average exon length of 336.57 bp. average of 5.22 exons (Table 4).

To comprehensively analyze gene functions, this study annotated gene functions and the metabolic pathways involved based on existing databases. The Diamond v2.1.8[46] was employed to perform homology alignment of the coding protein sequences in the gene set with established protein databases, including Uniprot (https://www.uniprot.org/) and NR (https://ftp.ncbi.nlm.nih.gov), to predict the biological functions of the protein sequences encoded by the genes. Additionally, InterProScan v5.55–88.0[47], InterPro v5.55–88.0[48], and HMMScan v3.3.2[49] were used to obtain information about conserved sequences, motifs, and domains within the proteins. Structural domain predictions were made, followed by Kyoto Encyclopedia of Genes and Genomes (KEGG)[50] annotation and Pfam annotation to further elucidate the functional roles of the predicted proteins in metabolic pathways and biological processes. Approximately 95.51% of the genes (29,535 genes) were functionally annotated in at least one database (Table 5).

Item	Count	Percentage
All	30,924	100.00%
Annotation	29,535	95.51%
KEGG	19,270	62.31%
Pathway	9,546	30.87%
NR	28,716	92.86%
Uniprot	28,355	91.69%
GO	21,536	69.64%
KOG	939	3.04%
Pfam	21,814	70.54%
Interpro	28,704	92.82%

Table 5. Annotation results of functional genes in *Psammosilene tunicoides*.

	Assembly		Annotation	
Features	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	1,588	98.40	1,592	98.60
Complete Single Copy BUSCOs	1,464	90.70	1,490	92.30
Complete Duplicated BUSCOs	124	7.70	102	6.30
Fragmented BUSCOs	9	0.60	7	0.40
Missing BUSCOs	17	1	15	0.90
Total BUSCO groups searched	1,614	100	1,614	100

Table 6. BUSCO assessment results for *Psammosilene tunicoides* genome assembly and annotation.

Item	Value
Unique Mapped Paired-end Pairs	419,798,773
Valid Paired-end Pairs	417,419,773
Valid Rate	99.43%
Raw Paired-end Pairs	484,096,738
Unique Mapped Paired-end Pairs	419,798,773
Unique Mapped Ratio	86.72%

Table 7. Statistics of Hi-C library information.

Data Overview

The sequencing generated 75.67 Gb of clean short reads, 51.24 Gb of clean HiFi reads, and 144.82 Gb of clean Hi-C reads. The initial assembly produced a 1,550,221,676 bp genome with a contig N50 of 91,918,445 bp. Subsequent Hi-C scaffolding yielded a chromosomal genome of 1,463,250,424 bp, featuring a scaffold N50 of 103,834,185 bp, with 94.40% of sequences anchored to 14 pseudochromosomes. Genome annotation revealed repetitive sequences spanning approximately 1,220,316,436 bp, constituting 83.40% of the assembly. We identified 5,574 snRNAs, 3,730 rRNAs, 1,397 tRNAs, and 110 miRNAs. A total of 30,924 protein-coding genes were predicted, of which 29,535 (95.51%) were functionally annotated.

Technical Validation

Assessment of the genome assembly. The quality of the genome assembly was evaluated in terms of continuity, consistency, and completeness. The contig N50 reached 91,918,445 bp, indicating a high degree of continuity (Table 1). Alignment of NGS data to the assembled genome yielded a mapping rate of 99.99% and a coverage of 99.77%, demonstrating high sequence consistency. Furthermore, the Complete BUSCO[51] (Embryophyta_odb10) score of 98.40% further confirmed its excellent completeness. Overall, the genome assembly of *P. tunicoides* was determined to be of high quality (Table 6).

Additionally, 94.40% of the assembled sequences were anchored to 14 chromosomes, reflecting a high degree of chromosomal-level continuity. To evaluate the accuracy of the chromosome assembly, we generated interaction maps using HiC Explorer v3.7.5[52], utilizing contig interaction intensity and positional information (Fig. 1E). The resulting Hi-C heatmaps exhibited well-defined chromosomal groupings with clear boundaries. Within each group, interaction signals were significantly stronger along the diagonal than in off-diagonal regions, indicating preferential interactions among adjacent sequences. This observation aligns with the principle of Hi-C-assisted assembly, wherein adjacent genomic segments exhibit stronger interactions. The observed interaction pattern, characterized by strong intra-chromosomal signals and weak inter-chromosomal associations, confirms the accuracy of genome anchoring and underscores the high quality of the chromosomal-level assembly (Tables 7, 8).

Category	All Ditags	Unique Ditags
Read Pairs	151,780,074	151.076,533
Cis-close (<10Kbp)	10,024,705	9,977,081
Cis-close (>10Kbp)	70,809,326	70,480,098
Trans	70,946,043	70,619,354

Table 8. Hi-C Interaction Statistics.

Lastly, Merqury[53] software was utilized to estimate the consensus QV of the assembly. Optimal k-mer length was determined as 19 via analysis with the "best_k.sh" script of the Merqury software. A k-mer spectrum of the whole genome was then constructed using NGS and CCS data with Meryl under default settings. Merqury evaluation, integrating the k-mer database with assembled sequences, showed k-mer completeness values of 96.32% and 96.16%, and a QV score of 46.51 and 80.26 based on k-mer consistency.

Assessment of the gene annotation. The quality of gene annotation was assessed using BUSCO[51] analysis. This analysis revealed that 98.60% of complete BUSCO genes were present in the annotation, indicating a high degree of completeness and accuracy (Table 6). Furthermore, the annotation file (gff) was validated using AGAT v2.0.0[54]. This analysis confirmed that the gff file was structurally sound, adhering to the standard hierarchical structure (gene-mRNA-UTR/CDS) and free of common formatting errors. No critical errors or warnings were reported during the AGAT validation.

Data availability

The NGS short reads, HiFi reads, and Hi-C reads data have been deposited in the NCBI Sequence Read Archive with accession numbers SRP638663[55], SRP638699[56], SRP637737[57]. The genome assembly has been deposited in GenBank under the accession number JBTWEM000000000[58]. The dataset of gene annotation files has been stored in the Figshare database[59].

Code availability

No custom scripts were utilized in this study. All data processing commands and pipelines were executed in strict accordance with the manuals and protocols provided by each bioinformatics software tool.

Received: 5 November 2025; Accepted: 28 February 2026;
Published online: 06 March 2026

References

1. Xin, Y. L. *et al.* Preliminary study on polyploid induction of *Psammosilene tunicoides*. *J. Chin. Med. Mater.* **40**(06), 1266–1269 (2017).
2. Pan, Y. Z., Yang, Z. & Y.&, G. X. The Chromosome Number of Psammosilene tunicoides Endemic to China. *Acta Bot. Yunnanica.* **26**(2), 204–206 (2004).
3. Qu, Y., Yu, H. & Zhou, X. L. Review on study advances on rare and endangered medicinal herb *Psammosilene tunicoides*. *China J. Tradit. Chin. Med. Pharm.* **26**(08), 795–7 (2011).
4. National Pharmacopoeia Commission of China. *Pharmacopoeia of the People's Republic of China* (China Medical Science Press, 2020).
5. Li, Y. H. *et al.* Triterpenoid saponins from *Psammosilene tunicoides* and their antinociceptive activities. *Phytochemistry.* **214**, 113795 (2023).
6. Tian, J. *et al.* Antifungal cyclic peptides from *Psammosilene tunicoides*. *J. Nat. Prod.* **73**(12), 1987–92 (2010).
7. Hou, Y. *et al.* Pharmacodynamics assessment of β-carboline from the roots of *Psammosilene tunicoides* as analgesic compound. *J. Ethnopharmacol.* **291**, 115163 (2022).
8. Zhao, B. S. *et al.* Research Progress in Chemical Compoents, Pharmacological Effectiveness and Toxicity of *Psammosilene tunicoides*. *Chin. J. Exp. Tradit. Med. Formulae* **17**(18), 288–291 (2011).
9. Li, R. *et al.* An old herbal medicine with a potentially new therapeutic application in inflammatory bowel disease. *Int. J. Clin. Exp. Med.* **4**(4), 309–19 (2011).
10. Li, Y. *et al.* Integrated analysis of transcriptomics and metabolomics and high-throughput amplicon sequencing reveals the synergistic effects of secondary metabolites and rhizosphere microbiota on root rot resistance in *Psammosilene tunicoides*. *Front. Microbiol.* **16**, 1554406 (2025).
11. Yang, W. T. *et al.* Comparative Analysis of Rhizosphere and Endophytic Microbial Communities Between Root Rot and Healthy Root of *Psammosilene tunicoides*. *Curr. Microbiol.* **80**(7), 215 (2023).
12. Li, J. *et al.* Application of barcode high-resolution melting for rapid authentication of the medicinal plant *Psammosilene tunicoides*. *Biotechnol. Biotechnol. Equip.* **30**(4), 790–796 (2016).
13. National Forestry and Grassland Administration. *List of National Key Protected Wild Plants* https://www.forestry.gov.cn/search/272693 (2021).
14. Ministry of Ecology and Environment of the People's Republic of China. China Biodiversity Red List https://www.mee.gov.cn/xxgk2018/xxgk/xxgk01/202305/t20230522_1030745.html (2023).
15. Chen, S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890 (2018).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
17. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**(14), 2202–2204 (2017).
18. Cheng, H. *et al.* Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**(2), 170–175 (2021).
19. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**(9), 2896–2898 (2020).
20. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
21. Zhang, X. *et al.* Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**(8), 833–845 (2019).

22. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromsome-length scaffolds. *Science* **356**(6333), 92–95 (2017).
23. Neva, C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**(1), 95–98 (2016).
24. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**(1), 99–101 (2016).
25. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**(17), 9451–7 (2020).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
27. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinform.* **9**, 18 (2008).
28. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **176**(2), 1410–1422 (2018).
29. Abrusán, G. *et al*. TEclass–a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**(10), 1329–1330 (2009).
30. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
31. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**(5), 955–964 (1997).
32. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**(9), 3100–3108 (2007).
33. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22), 2933–2935 (2013).
34. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**(D1), D192–D200 (2021).
35. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**(3), 290–295 (2015).
36. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001735.4/ (2018).
37. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_037179205.1/ (2024).
38. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_040167595.1/ (2019).
39. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_048544455.1/ (2025).
40. *NCBI Sequence Read Archive* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_026212465.1/ (2022).
41. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**(1), btad014 (2023).
42. Stanke, M. *et al*. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**(5), 637–644 (2008).
43. Delcher, A. L. *et al*. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**(6), 673–679 (2007).
44. Holt, C. & Yandell, M. an annotation pipeline and genome-database management tool for second-generation genome projects J. *BMC Bioinform.* **12**, 491 (2011).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**(1), R7 (2008).
46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**(1), 59–60 (2015).
47. Mulder, N. & Apweiler, R. InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.* **396**, 59–70 (2007).
48. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**(D1), D344–D354 (2021).
49. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**(1), 205–211 (2009).
50. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
51. Manni, M. *et al*. 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**(10), 4647–4654 (2021).
52. 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 (2020).
53. 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).
54. Dainat, J., Hereñú, D. & Pucholt, P. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. *Zenodo* https://doi.org/10.5281/zenodo.4205393 (2020).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP638663 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP638699 (2025).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP637737 (2025).
58. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBTWEM000000000 (2025).
59. Zhang, Y. M. Genome annotations of Psammosilene tunicoides. *Figshare* https://doi.org/10.6084/m9.figshare.30488117 (2025).

Acknowledgements

This work was supported by the National Key R&D Program of China (2024YFF1306700), the National Natural Science Foundation of China (82460751, 81560613), and Support Program for "Xindian" Talents of Yunnan - Study and Visiting Programs.

Author contributions

G.L. designed the study. Z.X., Y.Z., C.Y., T.Z. and A.Z. performed the experiments and analyzed the data. Z.X. and Y.Z. wrote the paper. G.L. and T.Z. revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.Z., A.Z. or G.L.