scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the alpine extremophyte Tibetan snow lotus, *Saussurea hypsipeta* Diels

Mingxia Wang[1], Guojing Hu[1], Lamu Yangjin[1], Siheng Zeng[1], Huiyan Xiong[2], Tao He[1]✉ & Ruijun Duan[1]✉

Saussurea hypsipeta **Diels, a perennial medicinal plant used in traditional Chinese medicine, is primarily distributed on the Qinghai-Tibet Plateau, where it serves as a key component of the alpine ecosystem. The lack of genetic information and high-quality genome assembly hinders in-depth research on this species as well as its conservation and further utilization. Here, we report the chromosome-level genome assembly of this diploid species, which was accomplished by integrating Illumina short-read, PacBio HiFi long-read, and Hi-C sequencing. The final assembly size was 3.18 Gb with excellent contiguity (contig N50 = 4.33 Mb; scaffold N50 = 182.60 Mb) and integrity (BUSCO score 97.46%). Among them, 98.91% of the sequences were anchored to 16 pairs of chromosomes, and the content of repetitive sequences was 87.15%. Furthermore, we predicted 70,374 genes, including 41,655 protein-coding genes, of which 94.13% were functionally annotated. This high-quality genomic resource provides a critical foundation for elucidating the molecular mechanisms of high-altitude adaptation, also offering genetic insights into its medicinal value.**

Background & Summary

As the most species-rich, ecologically adaptable, and economically valuable group of angiosperms, the Asteraceae family occupies a critical position in the global plant kingdom[1]. *Saussurea* DC. is one of the large genera within the Asteraceae family, primarily comprising perennial herbaceous species[2]. It is primarily distributed in the Hengduan and the Himalayas Mountains of Eurasia[3]. Approximately 33% of all *Saussurea* species occur in the Qinghai-Tibet Plateau and adjacent areas, with about 70% of species distributed between 4000 and 6000 meters[4]. Based on morphological characteristics such as pubescence, stem morphology, bract coloration, and involucral appendages, the genus is divided into six subgenera: *subg. Amphilaena*, *subg. Eriocoryne*, *subg. Jurinocera*, *subg. Theodorea*, *subg. Florovia*, and *subg. Saussurea*. Among these, *subg. Amphilaena* is characterized by conspicuously colored bracts and predominantly occurs in alpine meadows or scree slopes above 2000 m elevation[5]. *Subg. Eriocoryne* densely covered with woolly trichomes (partly referred to as "woolly plants")[6], a pubescence structure that effectively reduces transpiration and heat loss, provides insulation against low temperatures and strong UV radiation, and significantly enhances survival capacity in alpine environments. This unique adaptability fully demonstrates the genus's capacity to thrive in extreme alpine habitats[2].

Saussurea hypsipeta Diels, a perennial herb, belonging to the subg. *Eriocoryne* of the genus *Saussurea* within the Asteraceae family, is a characteristic and endemic species that constitutes a key component of the alpine ecosystem of the Qinghai-Tibet Plateau in China. In recent years, its unique woolly structure and extreme environmental adaptation mechanisms have attracted extensive attention from botanical, ecological, and conservation biology researchers. In Tibetan medicine, *S. hypsipeta* and *S. medusa* were commonly known as "Tibetan snow lotus"[7,8], and their above-ground parts can be used as medicine, which is effective in rheumatoid arthritis and gynecological diseases. At present, *S. hypsipeta* has a complete mitochondrial (GenBank accession number: ON584565.1)[9] and chloroplast genome (GenBank accession number: MT554929)[10]. However, the study of

[1]College of Eco-Environmental Engineering, Qinghai University, Xining, Qinghai, 810016, China. [2]College of Agriculture and Animal Husbandry, Qinghai University, Xining, Qinghai, 810016, China. ✉e-mail: hetaoxn@aliyun.com; ruijunduan@163.com

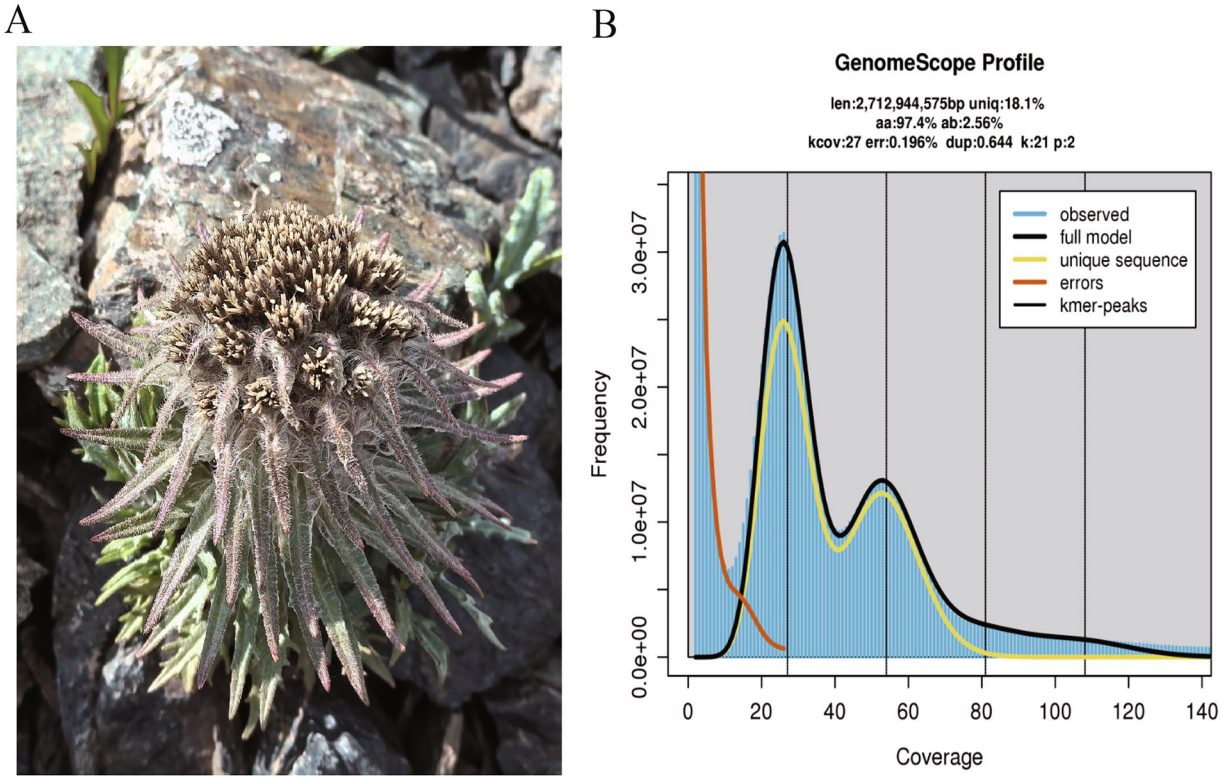

Fig. 1 Morphological characterization of *S. hypsipeta* (**A**) and GenomeScope K-mer distribution curves (Kmer size = 21) (**B**).

molecular adaptation mechanisms and pharmacological effects of this species has been limited for a long time due to the lack of high-quality genomic data at the chromosome level. In this context, the present research aims to provide deeper insights into the genomic study of *S. hypsipeta* by utilizing advanced sequencing technologies.

In this study, we performed de novo assembly of the chromosome-level genome of *Saussurea hypsipeta* using the haplotype-resolving assembler Hifiasm v0.15.5, combined with Illumina short-read, PacBio HiFi long-read sequencing, and Hi-C-assisted anchoring technologies. Initial attempts at chimeric assembly yielded suboptimal results, with low contiguity and high redundancy, due to the species' high heterozygosity. By switching to Hifiasm's haplotype-resolving strategy, we successfully separated two haplotypes (hap1 and hap2). Hap1 was selected as the reference genome for downstream analyses owing to its superior quality, with a total length of 3.18 Gb, a contig N50 of 4.33 Mb, and anchoring to 16 pairs of chromosomes via Hi-C. BUSCO assessment indicated a genome completeness of 97.46%. Repetitive sequences accounted for 87.15% of the genome, with long terminal repeats (LTRs) being the most abundant (58.45%). Gene prediction identified 70,374 genes, including 41,655 protein-coding genes (94.13% functionally annotated) and 28,719 non-coding RNA genes. K-mer analysis via GenomeScope revealed a high genome heterozygosity of 2.56% in *S. hypsipeta*, which is significantly higher than that of most reported Asteraceae relatives. This high heterozygosity may be attributed to the species' "sky island" distribution pattern across scattered alpine screes on the Qinghai-Tibet Plateau, combined with wind-mediated pollen and seed dispersal, which is likely to facilitate frequent inter-population gene flow.

This study successfully assembled the chromosome-level genome of *S. hypsipeta*, not only filling the gap in genomic resources for this species but also providing a critical data foundation for species conservation and in-depth analysis of the adaptive evolutionary mechanisms of high-altitude plants. Meanwhile, this high-quality genome provides scientific evidence for further utilization, particularly in terms of exploring the potential of this genus of plants in the synthesis and regulation mechanisms of medicinal active components such as anti-inflammatory and antioxidant compounds, laying a theoretical foundation for the development and sustainable utilization of its medicinal resources.

Methods

Sample collection and sequencing. Fresh leaves and different tissues of *S. hypsipeta* were collected from the gravel mountain at an altitude of 3749.9 m in the Qilian Mountain Spur (E101°23', N37°20') in the northeastern part of the Qinghai-Tibet Plateau (Fig. 1A). A voucher specimen of the collected sample has been formally deposited in the College of Eco-Environmental Engineering, Qinghai University, with the accession number STDLW202402, following the sampling guidelines of the Earth BioGenome Project (EBP). To ensure the integrity and purity of the DNA, the collected material was quickly frozen in liquid nitrogen and stored at −80 °C. High-molecular-weight genomic DNA was extracted using the SDS method and purified with the QIAGEN® Genomic DNA Extraction Kit (Cat. No. 13343, QIAGEN, Germany). Sample integrity and purity were further assessed by

Category	Illumina	PacBio HiFi
Total bases (bp)	216,740,459,100	100,176,029,551
Clean bases(bp)	1,444,936,394	5,714,895
GC content (%)	39.43	—
N50 length (nt)	—	17,695
Genome size estimation	**Genome size (bp)**	**Heterozygosity(%)**
findGSE	4,648,297,270	—
GenomeScope	2,712,944,575	2.559

Table 1. *S. hypsipeta* second- and third-generation sequencing data and genomic assessment results.

Genome assembly		annotation	
Assembled genome size(Gb)	3.18	Percentage of duplicate sequences	87.15%
Contig N50(Mb)	4.33	Number of genes	41,655
Total number of contigs	3177.27	Annotated ncRNA	28,719
Longest contigs(Mb)	25.67	Average exon length(bp)	246.14
GC content(%)	38.65	Average CDS length(bp)	1,076.38
BUSCO completeness of assembly(%)	97.46	Average intron length(bp)	725.26
Coverage depth(X)	30.13	Percentage of functional annotation	94.13%
Scafford N50(Mb)	182.59	Complete BUSCOs	97.51%
Scafford N90(Mb)	113.7	Complete and single-copy BUSCOs	85%

Table 2. Genome assembly and annotation statistics for *S. hypsipeta*.

Agarose Gel Electrophoresis. DNA quality and concentration were examined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA).

For genome survey sequencing, short-read libraries were constructed with an insert average of 200–400 bp using an Agencourt AMPure XP-Medium kit (Beckman Coulter, USA) and then sequenced on the MGISEQ. 2000 (MGI Tech, China) platform, which finally generated a total of 216.74 Gb of raw data, covering 68.15 × of the genome. To perform de novo genome assembly, a single-molecule real-time (SMRT) PacBio library was prepared using the SMRTbell Express Template Preparation Kit 2.0 (Pacific Biosciences, USA) following the manufacturer's instructions, and sequenced on the PacBio Sequel II platform (Nextomics, China) in Circular Consensus Sequence (CCS) mode to produce PacBio HiFi long reads, yielding 100.18 Gb of data (31.5 × coverage).

To prepare the library for high-throughput chromosome conformation capture (Hi-C) sequencing, Young fresh leaves (~1 g) of *S. hypsipeta* were fixed with 1% formaldehyde to coagulate the conformation of DNA. Subsequently, the Hi-C library was constructed based on the standard protocol and sequenced on an illumina HiSeq platform (Illumina, USA), obtaining 231.69 Gb of raw data (72.86 × coverage).

For genome annotation, total RNA was extracted from *S. hypsipeta* for transcriptome sequencing. The RNA library was prepared using the Ultima Dual-mode mRNA Library Prep Kit (Yeason, China) according to the manufacturer's instructions, and PE150 sequencing was conducted on a DNBSEQ-T7 platform (MGI Tech, China).

Genome assembly. Prior to genome assembly, we performed a survey analysis (K-mer method) on the genome of *S. hypsipeta* to assess key features such as genome size, heterozygosity, and repetitive sequence content. Due to the small amount of sequencing data (101.99 Gb) in the initial stage, the survey peak was abnormal. Subsequently, we generated 216.74 Gb of Illumina short-read sequencing data. The raw data were processed with Fastp v0.22.0[11] to remove low-quality reads and adapters, resulting in 195.19 Gb of clean data with a Q20 of 98.84%. The high Q score confirmed that the data were of sufficient quality for subsequent K-mer analysis. We used Jellyfish v2.3.0[12] to compute the frequency of 21-mers (K = 21) (Fig. 1B). This K-mer spectrum was then analyzed by both findGSE v1.0[13] and GenomeScope v2.0[14] to estimate genome characteristics. The results indicate that the estimated genome size was about 2.71 to 4.64 Gb, with a genome heterozygosity was 2.56%, and the GC content was 39.43% (Table 1). This heterozygosity is significantly higher than that of most published Asteraceae relatives with available genome sequences (e.g., *Artemisia annua*[15]: ~1.0–1.5%; *Lactuca sativa*[16]: 0.21%). High heterozygosity not only leads to the misassembly of divergent haplotype regions into chimeric contigs but also tends to generate extensive sequence redundancy during the independent assembly of both haplotypes. To address this challenge, we prioritized the assembly of a single haplotype with superior contiguity and completeness, followed by a stringent redundancy reduction step to eliminate spurious duplications arising from haplotype variations. This strategy ultimately yielded a high-quality, non-redundant reference genome, laying a solid foundation for subsequent analyses.

In PacBio's sequencing platform, basecalling was performed using SMRTLink (https://github.com/WenchaoLin/SMRT-Link) to obtain junction-containing sequencing sequences, and low-quality sequences

Type	Number	Percent (%)
Complete BUSCOs (C)	2,267	97.46
Complete and single-copy (S)	1,976	84.95
Complete and duplicated (D)	291	12.51
Fragmented BUSCOs (F)	11	0.47
Missing BUSCOs (M)	48	2.06
Total BUSCO groups searched	2,326	100
Number of scaffolds	1,325	56.96
Number of contigs	1,325	56.96
Total length (bp)	3,177,274,469	136,598,000

Table 3. BUSCO assessment statistics for *S. hypsipeta* genome assembly.

Sample id	Total reads	Total bases	Clean reads	Clean bases	Q20 rate(%)	Q30 rate(%)	GC(%)
Saussurea hypsipeta	1,544,611,928	231,691,789,200	1,530,485,410	214,198,010,878	98.19	96.18	40.79

Table 4. *S. hypsipeta* Hi-C sequencing data statistics.

Type	Count	Percent (%)
Unique Mapped Paired-end Reads	176,677,137	100
Dangling End Paired-end Reads	17,589,704	9.96
Self Circle Paired-end Reads	66,776	0.04
Dumped Paired-end Reads	4,642	0
Religation Paired-end Reads	3077841	1.74
Valid Paired-end Reads	155,938,174	88.26

Table 5. Hi-C read mapping statistics for *S. hypsipeta* genome assembly.

and junction sequences were removed for proofreading, and finally highly accurate HiFi reads were obtained (100 Gb) (Table 1). Given the high genome heterozygosity (2.56%) previously estimated by K-mer analysis, we employed Hifiasm v0.15.5[17] for de novo genome assembly under default parameters. To address the potential sequence redundancy and chimeric artifacts associated with high heterozygosity, we prioritized the primary haplotype assembly (hap1) from the Hifiasm output, which exhibited superior contiguity and completeness. Following genome assembly, a stringent redundancy reduction process was conducted using the software Purge Dups v1.2.6[18], aiming to eliminate spurious duplications caused by haplotype divergence. The final draft genome assembly size was 3.18 Gb, containing 1,325 contigs with an N50 length of 4.33 Mb and a GC content of 38.65% (Table 2).

To comprehensively evaluate the assembly quality, we performed multi-dimensional QC analyses: Completeness was assessed via BUSCOs v5.4.7[19] based on the eudicots_odb10 dataset from the OrthoDB database, identifying 2,267 (97.46%) complete BUSCOs (1,976 single-copy [84.95%], 291 duplicated [12.51%]), 11 (0.47%) fragmented, and 48 (2.06%) missing; for per-base accuracy, Illumina short reads were mapped to the assembly using BWA v0.7.17[20], achieving a 98.82% mapping rate and 99.96% 1 × coverage, with variant calling via samtools v1.13[21] and bcftools v1.13[22] yielding a 99.9987% per-base accuracy at $\geq 5\times$ depth. GC-Depth analysis was performed using mapped HiFi reads via minimap2 v2.17[23], revealing uniform clustering without separated clades and excluding exogenous contamination with distinct GC characteristics. Merqury was used to evaluate base-level consensus accuracy and sequence redundancy, with a consensus quality value (QV) of around 50, indicating that the genome has achieved high standards of base accuracy and sequence consistency. These results confirm the assembly's high quality, completeness, accuracy, and purity (Table 3).

After completing the initial assembly of the genome, Hi-C sequencing was further utilized to construct a chromosome-level, high-quality genome. A total of 231.69 Gb of data were obtained from Hi-C sequencing, and 214.20 Gb of clean data were obtained after filtration (Q30 $\geq$ 96.18%)(Table 4), of which 88.26% were Valid Pairs-end Reads (155,938,174 pairs), providing sufficient interaction signals for chromosome mounting. With the help of Juicer[24] and 3D-DNA[25], 92.12% of genomic sequences (2.93 Gb) were anchored to 16 pairs of chromosomes (chr01-chr16) (Fig. 3), which elevated the scaffold N50 to 182.59 Mb (Table 2). Among them, the length of the longest chromosome amounted to 274.52 Mb (chr01), with 7.88% of unanchored sequences (Tables 5 and 6). Intrachromosomal interaction heatmaps showed linear distance-dependent interaction patterns (Fig. 2), verifying the topological accuracy of chromosome-level assembly. At the same time, a circos plot visualizing multiple genomic features of *S. hypsipeta* was generated using Circos v0.69-9[26], revealing syntenic

Pseudomolecule	Length (bp)	Percent (%)
chr01	274,516,590	8.64
chr02	270,926,681	8.53
chr03	261,946,219	8.24
chr04	255,873,290	8.05
chr05	204,966,693	6.45
chr06	199,023,945	6.26
chr07	182,597,789	5.75
chr08	174,880,151	5.5
chr09	163,993,996	5.16
chr10	160,171,520	5.04
chr11	150,465,946	4.74
chr12	138,337,934	4.35
chr13	132,234,334	4.16
chr14	128,401,757	4.04
chr15	114,929,137	3.62
chr16	113,707,395	3.58
Total anchored	2,926,973,377	92.12
Unanchored	250,413,092	7.88
Total	3,177,386,469	100.00

Table 6. Chromosome-level assembly statistics of *S. hypsipeta*.

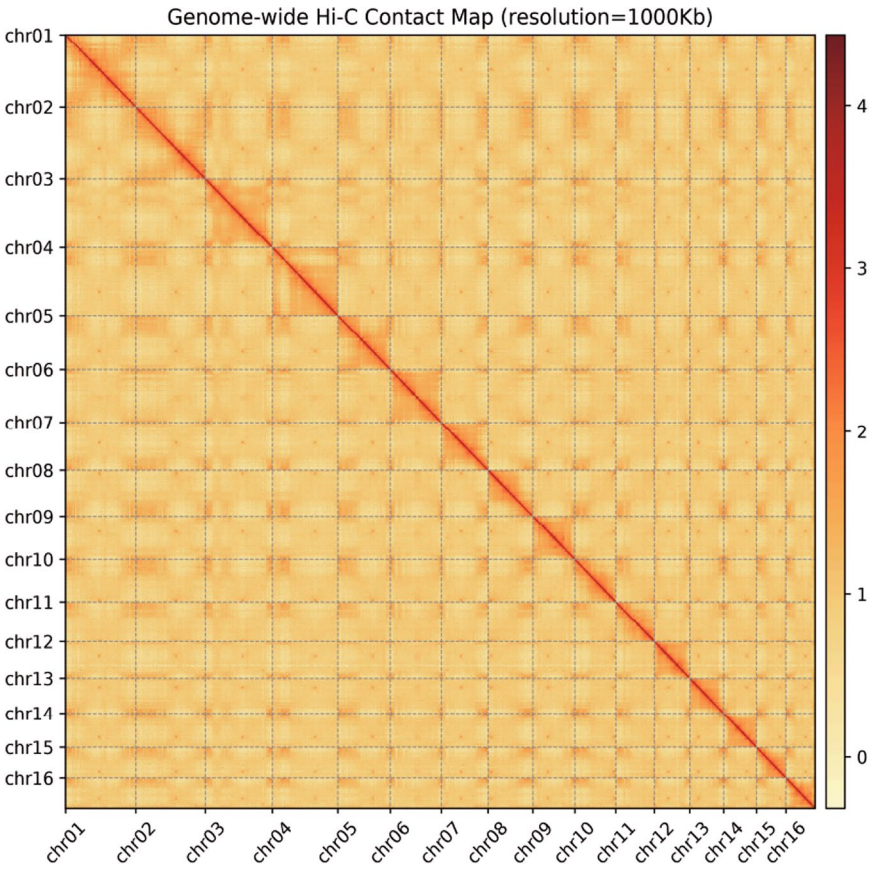

Fig. 2 Hi-C interaction heatmap for the genome assembly of *S. hypsipeta*.

relationships among different chromosomes (Fig. 3), and significant differences between the chromosomes of the alpine plant were revealed. These differences reflect the rearrangements and structural changes that the genome has undergone during evolution.

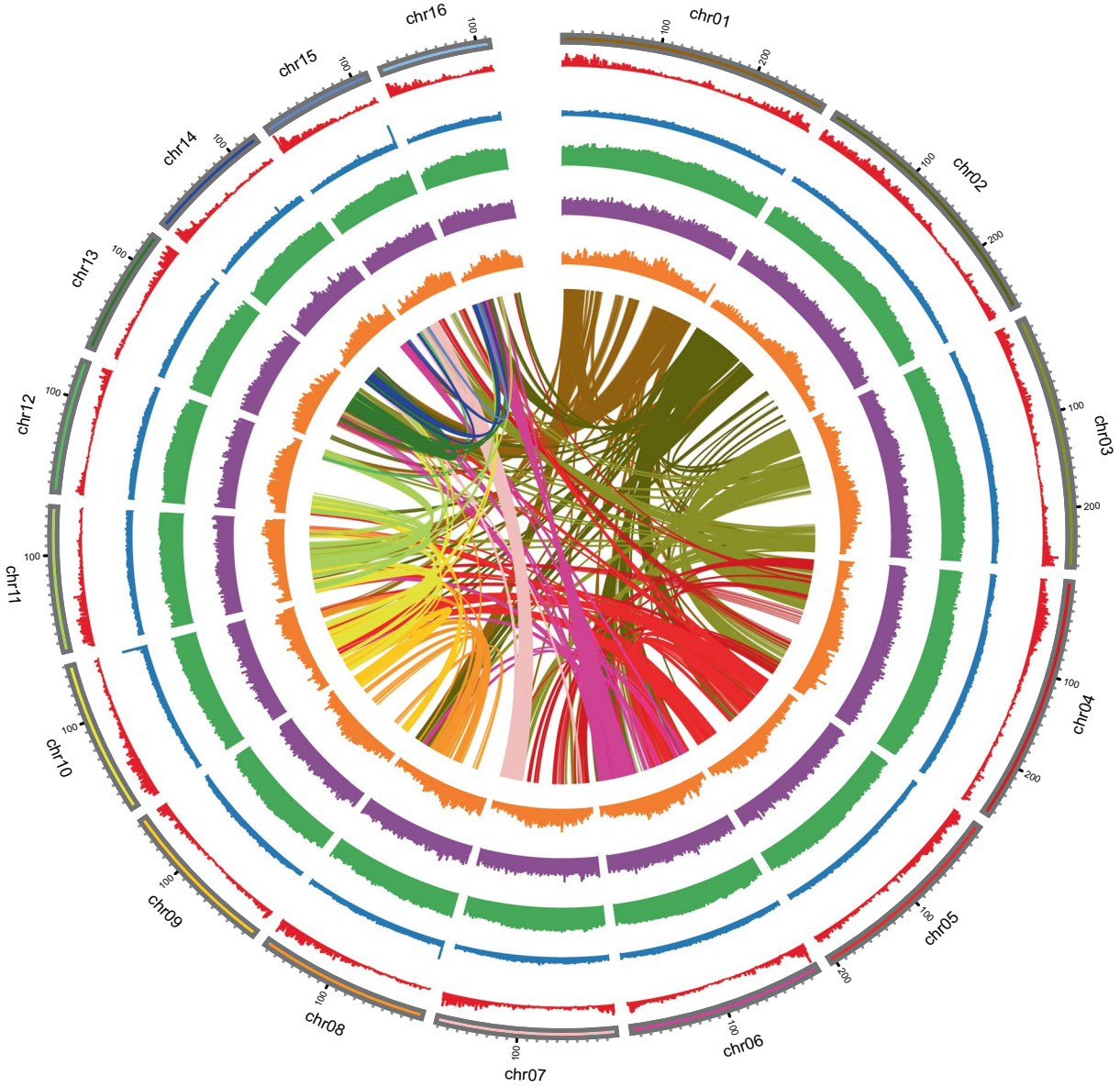

Fig. 3 Genomic features of *S. hypsipeta*. These features are organized in the order of 16 chromosomes, GC content, gene density, repeat density, LTR/Gypsy density, LTR/Copia density, and co-linear regions from outside to inside.

Genome annotation

Repeat sequence annotation. In this study, Simple sequence repeats (SSRs) were identified using Krait v1.0.3[27], yielding 524,213 SSRs (11.66 Mb, 0.37% of the 3.18 Gb genome). Tandem repeats were annotated with Tandem Repeats Finder (TRF) v4.09[28] under default parameters, resulting in 2,073,240 tandem repeats (257.36 Mb, 8.10% of the genome). For transposable element (TE) annotation, a combined de novo and homology-based strategy was employed: (1) A de novo repeat library was constructed using MITE-Tracker v1.0[29], LTR_Finder v1.0[30], LTR_harvest v1.0[31], and RepeatModeler v2.0.2a[32] (default parameters); (2) The de novo library was classified via TEsorter v1.4.6[33] against Repbase v20181026[34]; (3) RepeatMasker v4.1.5[35] was run with default parameters to map sequences against the classified de novo library and Repbase/Dfam v3.5[36] TE library, with overlapping TEs of the same class merged using a 50 bp overlap threshold. In total, 2.77 Gb of repetitive sequences were identified (87.15% of the genome), including 2.75 Gb of TEs (86.48%) and 22.0 Mb of non-overlapping tandem repeats (0.67%)—tandem repeats and SSRs were merged with removal of overlaps to avoid double-counting. Long terminal repeats (LTRs) accounted for 58.45% of TEs (1.86 Gb), with LTR/Gypsy (28.47% of the genome) and LTR/Copia (26.83% of the genome) as the two major superfamilies, and other LTR subfamilies contributing the remaining 3.15% (Fig. 3)—this classification result was derived from TEsorter analysis. DNA transposons, LINEs, and SINEs accounted for 7.41%, 1.5%, and 0.2% of the genome, respectively (Table 7).

Class	Repbase/Dfam TEs		de novo TEs		Combined TEs	
Type	Length(bp)	% in genome	Length(bp)	% in genome	Length(bp)	% in genome
DNA transposons	25,651,980	0.81	229,045,129	7.21	235,415,511	7.41
LINEs	7,679,671	0.24	46,504,157	1.46	47,712,491	1.5
SINEs	256,943	0.01	6,420,035	0.2	6,470,693	0.2
LTRs	477,461,473	15.03	1,839,539,213	57.89	1,857,244,422	58.45
Unknown	43,088	0	746,447,755	23.49	746,484,801	23.49
Others	7,175,288	0.23	25,103,400	0.79	28,785,755	0.91
Total repeats	515,195,184	16.21	2,731,539,104	85.97	2,747,930,050	86.48

Table 7. Statistics of repetitive elements in *S. hypsipeta* genome.

Type		Number	Average Length(bp)	Total Length(bp)	Base Ratio(%)
rRNA	28S rRNA	1,143	4,004	4,577,242	0.1441
	18S rRNA	1,078	1,540	1,660,860	0.0523
	5.8S rRNA	444	155	69,162	0.0022
	5S rRNA	10,433	115	1,209,461	0.0381
tRNA		1,390	75	104,880	0.0033
miRNA		178	122	21,849	0.0007
snoRNA		12,781	106	1,354,935	0.0426
catalytic RNA		3	197	592	0
spliceosomal RNA		857	161	138,633	0.0044
other sRNA		2	123	246	0
cis regulator		71	52	3,749	0.0001
other ncRNA		339	146	49,549	0.0016

Table 8. Non-coding RNA annotation statistics for *S. hypsipeta*.

Gene structure prediction and annotation. A consensus approach integrating transcriptomic evidence, homology-based comparisons, and *ab initio* prediction was employed for gene prediction. HISAT2 v2.2.1[37]-mapped RNA-seq reads were assembled with StringTie v2.2.1[38], and 3,000 of the resulting transcripts with the highest read support (FPKM $\geq$ 10 and intact open reading frames) were selected to train Augustus v3.4.0[39]. Homology-based comparisons were conducted in a second, independent step: ProtHint v2.6.0[40], assisted by Spaln v2.4.0[41], aligned with non-redundant proteins (length $\geq$ 100 aa) from seven Asteraceae species (*Artemisia annua*[42], *Carthamus tinctorius*[43], *Erigeron canadensis*, *Helianthus annuus*[44], *Lactuca sativa*, *Saussurea medusa* Maxim, *Taraxacum mongolicum*[45]) retrieved from NCBI RefSeq and UniProt databases. Finally, BRAKER3 v3.0.3[46] incorporated the Augustus training set, the ProtHint hints, and the original StringTie assemblies. On this basis, it internally re-trains GeneMark-ETP to optimize the prediction parameters, and then drives Augustus to execute a further evidence-integrated prediction run. The prediction results from BRAKER3 (integrating transcriptomic, homology evidence, and *ab initio* prediction) were reconciled for gene model filtering by TSEBRA v1.1.1[47] with default weighting coefficients (StringTie: 0.6, GeneMark-ETP: 0.3, *ab initio*: 0.1), reducing initial models from 48,320 to 41,655 high-quality ones. The results revealed that 41,655 protein-coding genes were successfully annotated in the *S. hypsipeta* genome, including the average gene length, coding sequence (CDS) length, average number of exons per gene, exon length, and intron length were 6,953.64 bp, 1,076.38 bp, 4.37, 246.14 bp, and 725.26 bp (Table 2).

Non-coding RNA prediction and annotation. Based on the assembled genome, non-coding RNAs (ncRNAs) were annotated by integrating results from multiple specialized tools, with final ncRNA sets determined as the union of non-redundant predictions. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE v1.23[48] with eukaryotic parameters (score $\geq$ 20, filtering low-confidence hits with score $<$ 20). For the identification of microRNAs (miRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs), the Infernal cmscan v1.1.4[49] tool was used to search against the Rfam v14.7[50] database with a strict E-value threshold ($\leq$1e-5). Ribosomal RNA subunits were additionally predicted with RNAmmer v1.2[51], with the union of consistent and high-confidence results retained. The results showed that a total of 28,719 non-coding RNAs were annotated in the genome of *S. hypsipeta*, including 178 miRNAs, 12,781 snoRNAs, 13,098 rRNAs, and 1,390 tRNAs (Table 8).

Gene functional annotation. The 41,655 protein-coding genes were functionally annotated using diamond v2.0.15[52] (NR, Swiss-Prot, TrEMBL) and eggNOG-mapper v2.1.0[53] (KEGG, KOG, GO, Pfam) against seven databases (NCBI-NR, KEGG[54], KOG[55], GO[56], Swiss-Prot[57], and TrEMBL[58] (Table 9). In total, 39,211 genes (94.13% of all protein-coding genes) were annotated in at least one database, and the annotation rates for individual databases ranged from 37.28% (KEGG) to 94.05% (NR), with the highest rates of about 94% for NR, reflecting

Database	Number	Ratio(%)
COG/KOG	34,384	82.54
KEGG	15,530	37.28
GO	16,162	38.8
Pfam	32,047	76.93
NR	39,178	94.05
Swiss-Prot	23,763	57.05
TrEMBL	38,097	91.46
Overall	39,211	94.13

Table 9. Gene function annotation statistics for *S. hypsipeta*.

Type	Number	Percent(%)
Complete BUSCOs (C)	2,268	97.51
Complete and single-copy BUSCOs (S)	1,977	85
Complete and duplicated BUSCOs (D)	291	12.51
Fragmented BUSCOs (F)	10	0.43
Missing BUSCOs (M)	48	2.06
Total BUSCO groups searched	2,326	100

Table 10. Genome annotation results of *S. hypsipeta* BUSCO assessment statistics.

strong homology between predicted protein models and known sequences. To comprehensively evaluate annotation quality beyond BUSCO, we integrated three lines of evidence from the study: (1) 76.93% of proteins were annotated with conserved domains via Pfam, verifying structural validity; (2) 66.95–68.78% of genes showed transcriptional support (FPKM $\geq$ 0.1) across multiple tissues, confirming functional expression; (3) gene structure features (average length, exons per gene) were consistent with seven Asteraceae relatives, indicating evolutionary conservation. Additionally, BUSCO assessment (OrthoDB eudicots clade)[59] showed 97.51% complete gene elements (Table 10). These multi-dimensional quality control data—including high BUSCO completeness (97.51%), conserved domain annotation (76.93%), transcriptional support (66.95–68.78%), and consistency with Asteraceae relatives—collectively confirm the reliability and completeness of the genome annotation (Table 2).

Data Records

All raw sequencing data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) database under BioProject accession number PRJNA1293189, including genome assembly (GenBank number: JBQAYS000000000.1[60]), genome survey sequencing data (accession number: SRR34717701[61], SRR34717702[62]), PacBio HiFi sequencing data (accession number: SRR34652138[63]), Hi-C sequencing data (accession number: SRR34702517[64]), RNA-seq sequencing data (accession number: SRR34702540[65],etc). The final chromosome-level genome assembly and corresponding annotation files—including the genome assembly (Sahy.genome.fa), coding sequences (Sahy.cds.zip), gene model (Sahy.gene.zip, GFF3 format), and protein sequences (Sahy.pep.zip), non-coding gene sequence files (Sahy.merged.ncRNA.zip), repetitive sequence files (Sahy.genome.repeat.TRF.zip)—have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.30952745[66]). All aforementioned data are publicly available to ensure widespread accessibility and reproducibility of the study.

Technical Validation

Genome assembly quality assessment. To evaluate the completeness of the assembled genome, two independent methods were employed. (1). Assembly integrity was assessed using BUSCO v5.4.7-based on the Benchmarking Universal Single-Copy Orthologs set of the eudicots_odb10 dataset in the OrthoDB database, revealing a 97.46% completeness rate comprising 84.95% single-copy and 12.51% duplicated BUSCO genes, with 0.47% fragmented and 2.06% missing genes. (2). Genomic survey sequencing short reads were mapped to the final assembly using BWA v0.7.17 under default parameters, achieving 98.82% alignment rate with 99.96% coverage. (3). Merqury analysis was performed to evaluate base-level consensus accuracy and sequence redundancy, with a consensus QV of around 50, confirming high base accuracy. (4). GC-Depth analysis excluded exogenous contamination, showing uniform clustering consistent with plant genome characteristics.

Gene annotation assessment. For gene annotation, we similarly utilized BUSCO v5.4.7 to evaluate the predicted proteins. Of the 2,326 BUSCOs, 2,268 (97.51%) were completed, including 1,977 single-copy BUSCOs and 291 duplicate BUSCOs. In addition, 10 BUSCOs were fragmented and 48 BUSCOs were missing. Multi-dimensional validation further confirmed annotation quality: (1). 94.13% of protein-coding genes were annotated in at least one public database, with 76.93% annotated with conserved domains via Pfam. (2). 66.95–68.78% of genes showed transcriptional activity (FPKM $\geq$ 0.1) across multiple tissues. (3). Gene structure features were consistent with seven Asteraceae relatives, indicating evolutionary conservation.

These results indicate that our predicted gene models have high annotation quality.

Data availability

All data generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) under BioProject accession number PRJNA1293189 and BioSample accession number SAMN50015486, GenBank WGS accession number JBQAYS000000000 (version JBQAYS010000000), and in Figshare (https://figshare.com/) under https://doi.org/10.6084/m9.figshare.30952745. All data are publicly available to ensure long-term preservation and access.

Code availability

No custom codes were used in this study. All bioinformatics tools and software applications were performed according to their respective manuals and protocols. All data analyses were performed using published bioinformatics software with default settings unless otherwise stated.

Received: 19 September 2025; Accepted: 17 February 2026;
Published online: 24 February 2026

References

1. Kane, N. C., Barker, M. S., Zhan, S. H. & Rieseberg, L. H. Molecular Evolution across the Asteraceae: Micro- and Macroevolutionary Processes. *Molecular Biology and Evolution* **28**, 3225–3235 (2011).
2. Raab-Straube, E. The genus Saussurea (Compositae, Cardueae) in China: taxonomie and nomenclatural notes. *Willdenowia* **41**, 83–95 (2011).
3. Xu, L. *et al*. New insights into the phylogeny and infrageneric taxonomy of *Saussurea* based on hybrid capture phylogenomics (Hyb-Seq). *Plant Diversity* **47**, 21–33 (2025).
4. Jin, Y.-L. *et al*. A plot-based dataset of plant community on the Qingzang Plateau. *Chinese Journal of Plant Ecology* **46**, 846–854 (2022).
5. Zhang, X. *et al*. Transcriptomes of Saussurea (Asteraceae) Provide Insights into High-Altitude Adaptation. *Plants* **10**, 1715 (2021).
6. Zhang, Y., Chen, J. & Sun, H. Alpine speciation and morphological innovations: revelations from a species-rich genus in the northern hemisphere. *AoB Plants* **13**, plab018 (2021).
7. Yu, R. *et al*. Dynamic Microwave-Assisted Extraction of Arctigenin from *Saussurea medusa* Maxim. *Chroma* **71**, 335–339 (2010).
8. Dai, W., Ju, X., Shi, G., Guo, J. & He, T. *ShCTR1* Interacts with *ShRBOH1* to Positively Regulate Aerenchyma Formation in *Saussurea inversa* through ROS Mediation. *PHYTON* **93**, 1023–1042 (2024).
9. Dai, W., Ju, X., Shi, G. & He, T. Assembly and Comparative Analysis of the Complete Mitochondrial Genome of *Saussurea inversa* (Asteraceae). *Genes* **15**, 1074 (2024).
10. Wang, J. *et al*. Complete chloroplast genome of *Saussurea inversa* (Asteraceae) and phylogenetic analysis. *Mitochondrial DNA B Resour* **6**, 8–9 (2021).
11. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
12. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
13. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* **34**, 550–557 (2018).
14. 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).
15. Shen, Q. *et al*. The Genome of *Artemisia annua* Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis. *Molecular Plant* **11**, 776–788 (2018).
16. Zhang, B. *et al*. A near-complete chromosome-level genome assembly of looseleaf lettuce (*Lactuca sativa* var. crispa). *Sci Data* **11**, 961 (2024).
17. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
18. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
19. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
21. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
22. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
25. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
26. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res*. **19**, 1639–1645 (2009).
27. Du, L., Zhang, C., Liu, Q., Zhang, X. & Yue, B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* **34**, 681–683 (2018).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
29. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* **19**, 348 (2018).
30. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10**, 48 (2019).
31. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
32. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
33. Zhang, R.-G. *et al*. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res* **9**, uhac017 (2022).
34. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
35. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, 4.10.1–4.10.14 (2009).
36. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81–89 (2016).
37. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
38. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).

39. 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 (2008).
40. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**, lqaa026 (2020).
41. Iwata, H. & Gotoh, O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res* **40**, e161 (2012).
42. Liu, Z. *et al.* Manual correction of genome annotation improved alternative splicing identification of Artemisia annua. *Planta* **258** (2023).
43. Dong, Y. *et al.* The Carthamus tinctorius L. genome sequence provides insights into synthesis of unsaturated fatty acids. *BMC Genomics* **25** (2024).
44. Buti, M. *et al.* Temporal dynamics in the evolution of the sunflower genome as revealed by sequencing and annotation of three large genomic regions. *Theor Appl Genet* **123**, 779–791 (2011).
45. Lin, T. *et al.* Extensive sequence divergence between the reference genomes of *Taraxacum koksaghyz* and *Taraxacum mongolicum*. *Sci. China Life Sci.* **65**, 515–528 (2022).
46. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* 2023.06.10.544449 (2024).
47. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22** (2021).
48. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
49. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
50. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192–D200 (2021).
51. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108 (2007).
52. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
53. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115–2122 (2017).
54. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361 (2017).
55. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
56. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2021).
57. McMillan, L. E. & Martin, A. C. Automatically extracting functionally equivalent proteins from SwissProt. *BMC Bioinformatics* **9** (2008).
58. O'Donovan, C. *et al.* High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* **3**, 275–284 (2002).
59. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research* **51**, D445–D451 (2023).
60. Wang, M. X. & Duan, R. J. *Genbank* https://www.ncbi.nlm.nih.gov/nuccore/JBQAYS000000000.1 (2025).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34717701 (*2025*).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34717702 (*2025*).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34652138 (*2025*).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34702517 (*2025*).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34702540 (*2025*).
66. Wang, M. X. Chromosome-level genome assembly and annotation files of *Saussurea hypsipeta* Diels. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.30952745.v1 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China Grant (31960222); Natural Science Foundation Project of Qinghai Provincial Department of Science and Technology (2025-ZJ-993M); 2023 Kunlun Elite·High-End Innovation and Entrepreneurship Talent Program of Qinghai Province.

Author contributions

R.D., T.H. and H.X. designed the article idea and led the study. M.W., G.H. and R.D. participated in sample collection. M.W., YJ.LM., G.H. and S.Z. participation in bioinformatics analysis. M.W. drafted the manuscript. All authors read, revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.H. or R.D.