scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Hippophae salicifolia*

Zhefei Zeng[1,2], Chunmin Mao[3], Jiayin Zhang[4], Zhongqiong Tian[1,2], Min Xu[5], Wei Li[1,2], Junwei Wang[1,2] ✉, Wenju Zhang[4] ✉ & La Qiong[1,2] ✉

***Hippophae salicifolia*, a dioecious small tree species endemic to the Himalayan region, holds great potential in both ecological conservation and industrial applications. In this study, we employed PacBio HiFi long reads, Illumina short reads, and Hi-C technology to construct a high-quality, chromosome-level reference genome. The assembled genome is approximately 1.11 Gb in size, with a scaffold N50 of 95.29 Mb, and 99.94% of the sequences were successfully anchored to 12 pseudo-chromosomes. A total of 42,547 protein-coding genes were predicted, and approximately 85% of these genes obtained functional annotations. Repetitive elements constituted about 45.25% of the genome, with Long Terminal Repeat (LTR) being the most abundant (32.54%). BUSCO analysis indicated that both the assembly and annotation are highly complete. This high-quality genomic resource provides a valuable foundation for investigating sex determination mechanisms, adaptive evolution, and genomic diversity in *H. salicifolia* and related species, as well as for advancing genetic improvement, resource conservation, and utilization efforts.**

Background & Summary

The genus *Hippophae* (family Elaeagnaceae) is broadly distributed across the Qinghai–Tibet Plateau and its adjacent regions. Members of this genus exhibit remarkable adaptability to harsh environments, including drought, cold, salinity, and nutrient-poor soils[1,2]. Furthermore, *Hippophae* species can form nitrogen-fixing root nodules with Frankia bacteria[2,3], thereby improving soil conditions. Through clonal propagation via root suckers, they can establish stable vegetation communities and effectively mitigate soil erosion. Moreover, their fruits provide a vital food source for various wild animals, thereby contributing to the maintenance of ecosystem diversity. Beyond their ecological significance, these species hold substantial potential for applications in the food, pharmaceutical, and cosmetic industries. Their fruits and leaves are particularly rich in vitamin C, flavonoids, essential fatty acids, and various secondary metabolites, which exhibit potential antioxidant, anti-inflammatory, antibacterial, and cardiovascular protective activities[1,2,4].

Hippophae salicifolia D. Don, a dioecious deciduous tree named for its willow-like leaves[5], is primarily distributed along riverbanks, slopes, and shrublands on the southern slopes of the Himalayas, including southeastern Tibet in China, as well as regions in Nepal, Bhutan, and northern India[6]. Previous studies have indicated that *Hippophae* species possess an XY sex determination system ($2n = 24$)[7–10]. Elucidating the mechanisms underlying sex determination and dioecy in *Hippophae* is crucial for gaining deeper insight into its adaptive evolution in the unique environment of the Qinghai–Tibet Plateau. However, research on sex-related differences in dioecious *Hippophae* species remains notably limited. Most previous studies have focused on chromosome-level genomes[11–16], chloroplast genomes[17–19], mitochondrial genomes[20,21], sex-specific molecular markers[7,9,10,22,23], fruit nutrient content[24,25], and transcriptomes[26,27]. With advancements in third-generation sequencing technologies (e.g., PacBio HiFi), which offer higher accuracy and throughput at reduced costs[28,29], it is now feasible to

[1]Key Laboratory of Biodiversity and Environment on the Qinghai-Tibetan Plateau, Ministry of Education, School of Ecology and Environment, Xizang University, Lasa, China. [2]Yani Observation and Research Station for Wetland Ecosystem of the Tibet (Xizang) Autonomous Region, Xizang University, Linzhi, China. [3]Yunnan Key Laboratory for Integrative Conservation of Plant Species With Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China. [4]Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering and State Key Laboratory of Genetic Engineering, Institute of Biodiversity Sciences and Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai, China. [5]Forestry Survey and Planning Research Institute of Tibet Autonomous Region, Lasa, China. ✉e-mail: jwyx12240315@126.com; wjzhang@fudan.edu.cn; lhagchong@163.com

Assembly feature	
Estimated genome size (Mb)	1,050.93
Assembly size (Mb)	1,108.28
Scaffold N50 (Mb)	95.29
Contig N50 (Mb)	38.57
L 50	5
Gaps	210
GC content (%)	29.76
Genome annotation	
Number of protein-coding genes	42,547
Average gene length (bp)	3,663
Average CDS length (bp)	1,034
Average exon length (bp)	291
Average intron length(bp)	682

Table 1. Statistics of *H. salicifolia* genome assembly and annotation.

generate high-quality reference genomes for non-model species such as *H. salicifolia*, thereby facilitating comprehensive investigations into sex determination and adaptive evolution.

To date, several *Hippophae* species have been successfully assembled at the genome level, including *H. rhamnoides*[11–13], *H. tibetana*[15,16], and *H. gyantsensis*[14]. These genomic resources have substantially advanced our understanding of gene diversity, phylogenetics, adaptive evolution, and sex determination mechanisms within the genus. Nonetheless, genomic research on *H. salicifolia* remains scarce, hindering systematic insights into its genetic background, evolutionary status, and sex determination processes.

In the present study, we employed PacBio HiFi long-read sequencing, Illumina short-read sequencing, and Hi-C data to generate a high-quality, chromosome-level reference genome of *H. salicifolia*. The assembled genome is approximately 1.11 Gb in size, with a scaffold N50 of 95.29 Mb (Table 1), and 99.4% of the sequences were successfully anchored to 12 putative chromosomes. A total of 42,547 genes were predicted, with repetitive elements accounting for approximately 45.25% of the genome. This is the first high-quality reference genome for *H. salicifolia*, thereby providing crucial data for elucidating sex determination mechanisms, genome evolution, and adaptive evolution within the genus. Moreover, it lays a robust scientific foundation for future research in genetic improvement, resource conservation, and sustainable utilization of *Hippophae* species.

Methods

Sample collection and sequencing. In June 2024, samples of a female individual (XX) of *H. salicifolia* were identified and collected by La Qiong and Junwei Wang in the Lebu Valley Scenic Area of Cuona County, Shannan City, Tibet (27°55′25.98″ N, 91°48′49.62″ E) (Fig. 1a). A voucher specimen (No. LQ20240672) was deposited in the Herbarium of the College of Ecology and Environment, Xizang University. To ensure the acquisition of high-quality DNA and RNA, fresh young leaves and bark tissues were collected and promptly flash-frozen in liquid nitrogen on-site, followed by storage at –80 °C in the laboratory. Genomic DNA was extracted using a modified CTAB method[30], and its purity and concentration were measured using a NanoDrop 2000 (Thermo Fisher Scientific, USA) and Qubit 2.0 (Invitrogen, USA). Integrity was assessed by 1% agarose gel electrophoresis. Only DNA samples that satisfied the quality criteria were selected for subsequent library construction and sequencing.

High-quality genomic DNA was sheared into fragments of approximately 350 bp to construct Illumina sequencing libraries used for an initial genome survey. Library quality was verified using an Agilent 2100 Bioanalyzer and quantitative PCR. Paired-end sequencing (150 bp) was conducted on the Illumina NovaSeq. 6000 platform, yielding approximately 62.78 Gb of high-quality short-read data. These short reads were primarily utilized to estimate genome size, GC content, and heterozygosity (Table 2). To obtain a high-accuracy de novo genome assembly, PacBio Sequel II was employed for HiFi sequencing. DNA was sheared into large fragments of 15–20 kb, and the SMRTbell Express Template Prep Kit 2.0 was applied to construct HiFi libraries. Approximately 31.80 Gb of HiFi data were obtained, with an average read length of 15.94 kb and a sequencing depth of about 28.69× (Table 2).

To achieve a chromosome-level assembly, high-throughput chromosome conformation capture (Hi-C) technology was employed. Fresh leaf tissues were fixed in formaldehyde and digested with the restriction enzyme MboI, then subjected to ligation and purification to construct Hi-C libraries. Libraries that passed quality checks underwent sequencing on the Illumina NovaSeq. 6000 platform (150 bp paired-end), generating approximately 102.43 Gb of Hi-C data (Table 2).

In addition, to support gene structure annotation and functional analysis, total RNA was extracted from leaves and bark of *H. salicifolia* to construct transcriptome libraries. The resulting libraries were sequenced on the Illumina NovaSeq. 6000 platform using 150 bp paired-end reads, generating approximately 8 Gb of transcriptome data per sample (Table 3). These transcriptomic resources provide valuable data for subsequent gene prediction and functional annotation.

Genome size estimation and survey analysis. Using high-quality Illumina short-read data, a k-mer analysis was performed to estimate the genome size, heterozygosity, and proportion of repetitive sequences in

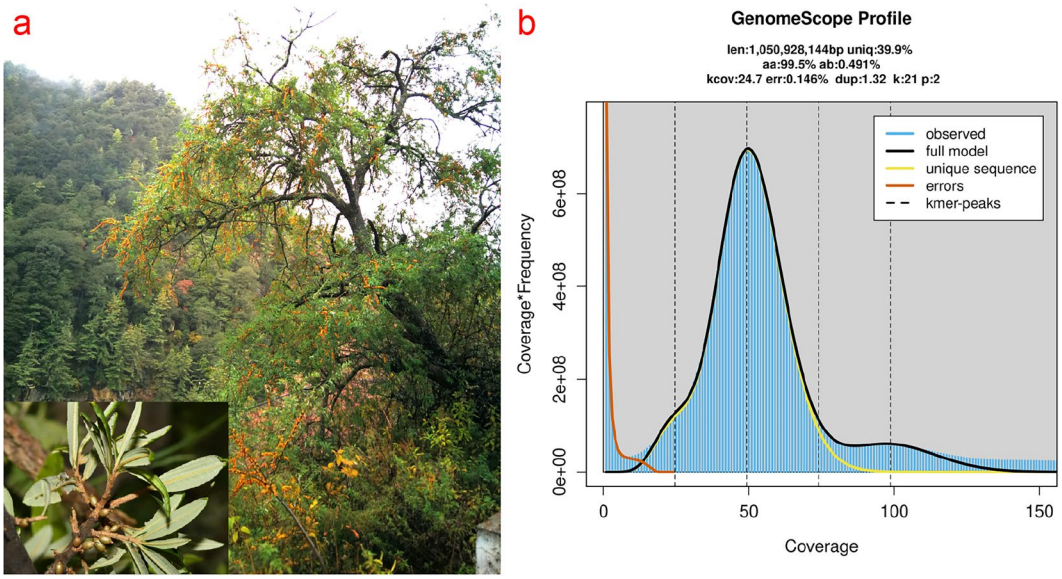

Fig. 1 Genome survey of *H. salicifolia*. (**a**) Photograph of the *H. salicifolia* plant. (**b**) K-mer (k = 21)-based genome size estimation. The blue region depicts the observed 21-mer frequency distribution; the black curve illustrates the fitted model, and yellow and red portions correspond to unique and erroneous K-mer distributions, respectively.

Read_type	Read_base (GB)	Depth (×)
HiFi reads	31.80	28.69
Illumina reads	62.78	56.65
Hi-C reads	102.43	92.42

Table 2. DNA sequencing statistics.

Sample	Sequencing platform	Raw data		Clean data	
		Total number of reads	Total number of bases (G)	Total number of reads	Total number of bases(G)
RNA leaf	Illumina	50,711,832	7.61	49,973,692	7.50
RNA stem	Illumina	58,344,950	8.75	57,367,492	8.61

Table 3. RNA sequencing statistics.

H. salicifolia. First, raw reads were processed with fastp v0.20.0[31] to remove adapter contamination and low-quality reads. Next, Jellyfish v2.2.10[32] was utilized to count the distribution of 21-mer frequencies, and GenomeScope v2.0[33] was employed to estimate genome characteristics. These analyses indicated that the *H. salicifolia* genome size is approximately 1.05 Gb, with a heterozygosity of about 0.49% (Fig. 1b).

Genome assembly. The *H. salicifolia* genome was de novo assembled using high-quality HiFi long-read data with hifiasm v0.19.9[34]. Hifiasm was executed with default parameters to fully exploit the high accuracy and long fragment lengths of the HiFi data, resulting in a highly continuous initial assembly. Subsequently, Illumina short-read data were integrated for assembly polishing. Specifically, BWA-MEM v2.21[35] was applied to map quality-controlled Illumina reads back to the initial assembly, followed by two rounds of error correction with Pilon v1.24[36] ("–fix all") to address potential base errors and fill small gaps, thereby enhancing the assembly's accuracy and completeness.

To achieve a chromosome-level genome assembly, Hi-C data were integrated to further optimize the polished assembly. First, Juicer v2.0[37] was employed to preprocess and map the Hi-C reads to the corrected genome. Then, 3D-DNA[38] was executed with default parameters to scaffold the assembly, followed by manual inspection and adjustment in Juicebox v2.15.07[39] to ensure accuracy and integrity at the chromosome scale. Based on previously published *H. rhamnoides* genome information[10], the chromosomes were designated as Chr01 through Chr12. The final chromosome-level assembly has a total length of 1.11 Gb, with a scaffold N50 of 95.29 Mb, a contig N50 of 38.57 Mb, and an L50 of 5, containing a total of 210 gaps and exhibiting a GC content of 29.76% (Table 1). Approximately 99.94% of the assembled sequences were successfully anchored onto 12 putative chromosomes (Table 4). The Hi-C interaction matrix revealed clear intra-chromosomal signals along the diagonal (Fig. 2),

Chromosome	Length (bp)	GC (%)
Chr01	162,210,453	29.06
Chr02	132,700,124	29.2
Chr03	117,474,828	30.21
Chr04	138,499,311	29.93
Chr05	92,498,625	30.11
Chr06	72,684,416	29.96
Chr07	69,720,992	30.48
Chr08	95,293,244	28.34
Chr09	58,792,619	30.72
Chr10	52,774,575	30.42
Chr11	58,169,684	30.39
Chr12	56,826,425	30.1
ChrUn	631,722	36.99

Table 4. Summary of the 12 pseudochromosomes.

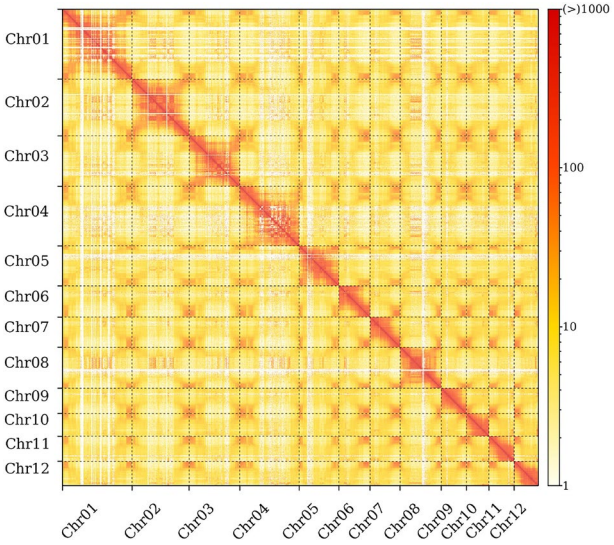

Fig. 2 Heatmap of genome-wide Hi-C data for *H. salicifolia*. Hi-C interaction frequencies are depicted by colors ranging from orange (low frequency) to dark red (high frequency).

indicating a high level of continuity and accuracy, thus providing a robust foundation for subsequent gene annotation and functional analyses.

Repeat annotation. To comprehensively characterize the repetitive elements in the *H. salicifolia* genome, de novo repeat prediction was conducted using RepeatModeler v2.0.1[40] (http://www.repeatmasker.org/RepeatModeler/) to establish a species-specific repeat library. This customized library was subsequently merged with the RepBase database[41] (v20181026, http://www.girinst.org/repbase). The combined library served as input to RepeatMasker v4.1.0[42] (http://www.repeatmasker.org) for the identification and masking of repetitive sequences. RepeatMasker identifies and annotates repetitive elements by comparing sequences to a curated transposable element (TE) library, which defines the classification based on sequence similarity and structural features, including major types such as LTR, LINE, SINE, and DNA transposons.

The annotation indicated that the total length of repetitive sequences was approximately 501.59 Mb, representing 45.25% of the entire genome (Table 5). Among these, LTR retrotransposons comprised the highest proportion (32.54%, ~360.72 Mb), while DNA transposons (4.01%, ~44.50 Mb) and LINE elements (0.45%, ~4.97 Mb) were also present (Table 5). This comprehensive repetitive element profile provides a valuable foundation for investigating genome evolution, structural variation, and gene regulatory mechanisms.

Protein-coding gene prediction and functional annotation. To obtain a high-quality set of protein-coding genes, three complementary strategies were integrated: homology-based prediction, transcriptome-based prediction, and ab initio prediction. First, for homology-based prediction, protein sequences from sequenced *Hippophae* species (e.g., *H. rhamnoides*, *H. tibetana*, and *H. gyantsensis*) were aligned to the *H. salicifolia* genome using GeMoMa v1.9[43] to identify potential orthologous genes. Second, for

Repeat type	Number of fragments	Length (bp)	Percentage of sequence
Retrotransposons	352,617	365,702,691	32.99
SINEs	146	7,408	0
Penelope	1,689	73,168	0.01
LINEs	18,094	4,978,496	0.45
LTR elements	334,377	360,716,787	32.54
BEL/Pao	691	51,919	0
Ty1/Copia	115,952	143,695,980	12.96
Gypsy/DIRS1	202,040	195,355,391	17.62
Retroviral	2088	92,602	0.01
DNA transpsons	108,992	44,502,209	4.01
hobo-Activator	30,329	12,654,528	1.14
Tc1-IS630-Pogo	2,490	111,623	0.01
PiggyBac	330	15,709	0
Tourist/Harbinger	3,187	708,159	0.06
Other	2,334	111,124	0.01
Rolling-circles	4,673	347,737	0.03
unclassified	286,437	91,387,643	8.24
Total interspersed repeats	—	501,592,543	45.25
Satellites	7,374	1,252,346	0.11

Table 5. Summary of the repetitive sequences in *H. salicifolia* genome assembly.

Methods	Gene set	Gene number	Average gene length (bp)	Average CDS length (bp)	Average exon number per gene	Average exon length (bp)	Average intron length (bp)
Ab initio annotation	AUGUSTUS	39,152	3,773	1,105	5.19	213	637
	GlimmerHMM	76,857	12,664	568	3.8	149	4323
Homologous annotation	*H. rhamnoides*	37,287	3,670	1,081	5.16	209	622
	H. tibetana	32,242	4,096	1,146	5.27	217	690
	H. gyantsensis	44,511	3,030	9,73	4.07	238	669
Transcriptome annotation	TransDecoder	24,020	5,948	1,546	7.57	340	695
Integration	EVM	42,599	3,239	1,009	4.34	232	668
Final set	PASA	42,547	3,663	1,034	4.72	291	682

Table 6. Summary of predicted protein-coding genes in *H. salicifolia* genome assembly.

transcriptome-based prediction, RNA-seq data from leaf and bark tissues were assembled into transcripts using stringtie v2.1.3[44], and coding regions were predicted using TransDecoder v5.1.0 (https://github.com/TransDecoder/TransDecoder). Lastly, ab initio predictions were performed with AUGUSTUS v3.3.3[45] (https://github.com/Gaius-Augustus/Augustus), GlimmerHMM v3.0.4[46], and GeneMark-ES v4.38[47]. Species-specific parameters were applied to improve prediction accuracy. All these predictions were integrated using EVidenceModeler (EVM) v1.1.1[48] to generate a high-confidence gene set, and PASA v2.5.2[49] (https://github.com/PASApipeline/PASApipeline) was subsequently employed for further refinement, adding UTR information and identifying novel transcripts. In total, 42,547 protein-coding genes were predicted, with an average gene length of 3,663 bp and an average of 4.72 exons per gene (Table 6). TBtools v2.126[50] was utilized to visualize gene density, GC content, Gypsy and Copia element densities, and chromosomal synteny of the 12 chromosomes (Fig. 3).

To gain comprehensive insights into gene functions, functional annotation was performed by conducting BLASTp[51] searches (E-value $\leq$ 1e−5) against multiple databases, including the Kyoto Encyclopedia of Genes and Genomes (KEGG)[52], euKaryotic Orthologous Groups (KOG), the National Center for Biotechnology Information Non-Redundant database (NCBI-NR, https://www.ncbi.nlm.nih.gov/), Gene Ontology (GO)[53], Clusters of Orthologous Groups (COG)[54], and SwissProt[55]. Additionally, HMMER v3.2.1 was employed, along with the Pfam database, to predict protein domains. Approximately 85.06% of the genes (36,189 genes) were functionally annotated in at least one database, thus providing a solid basis for subsequent functional genomics research (Table 7).

Non-Coding RNA annotation. To identify non-coding RNAs (ncRNAs) in the *H. salicifolia* genome, multiple tools and databases were integrated. A total of 748 tRNA genes were detected using tRNAscan-SE v2.0.0 (http://lowelab.ucsc.edu/tRNAscan-SE/)[56] (Table 8). Using the Rfam v14.2 database[57] and Infernal v1.1.3[58], 5,924 rRNA genes, 196 miRNA genes, and 5,950 snRNA genes were identified. These ncRNA annotations will facilitate studies on transcriptional regulation and the functional mechanisms underlying the *H. salicifolia* genome.

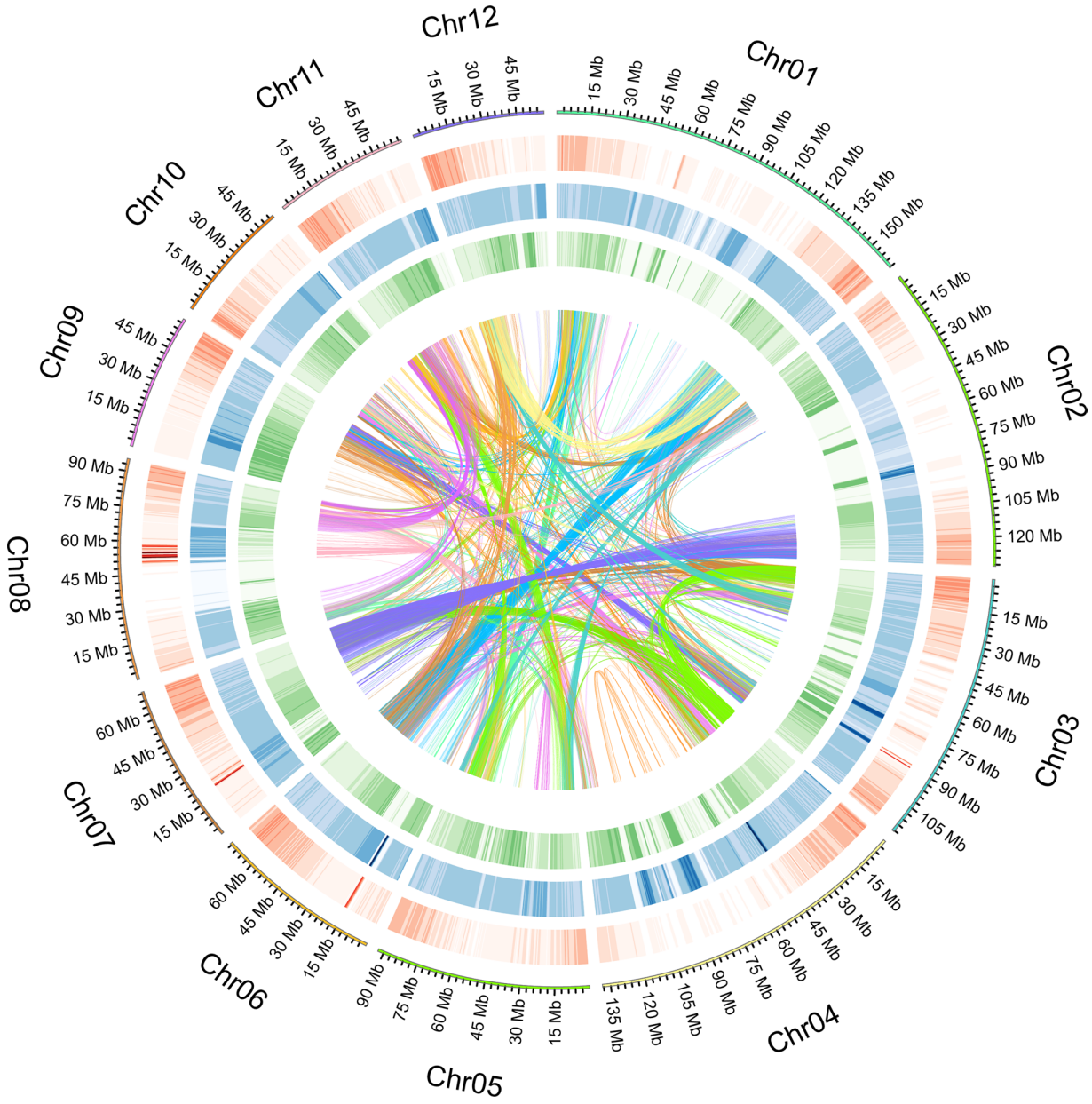

Fig. 3 Genomic features of *H. salicifolia*. From outer to inner circles: 12 chromosomes (Chr01–Chr12), GC content, gene positions, gene density, and syntenic gene blocks represented by connecting lines within the genome.

Anno_Database	Annotated_Number	300 < = length < 1000	length > = 1000
COG_Annotation	11,779	4,118	7,361
GO_Annotation	24,067	9,436	12,880
KEGG_Annotation	35,689	15,958	16,445
KOG_Annotation	20,365	8,252	10,503
Pfam_Annotation	27,375	11,047	15,152
Swissprot_Annotation	26,550	10,704	13,891
nr_Annotation	35,991	16,087	16,444
All_Annotated	36,189	16,181	16,468

Table 7. Annotation results of functional genes in *H. salicifolia*.

Family	Number	Average length(bp)	Total length(bp)	(%) in genome
tRNA	748	74	56,081	0.0051
miRNA	196	127	25,012	0.0023
sRNA	1	374	374	0
antisense	19	166	3,156	0.0003
ribozyme	1	245	245	0
Intron	117	128	15,014	0.0014
Cis-reg:leader	1	118	118	0
Cis-reg:riboswitch	1	132	132	0
Cis-reg:other	7	45	320	0
Cis-reg:total	9	63	570	0.0001
snRNA:CD-box	5,804	106	616,955	0.0557
snRNA:HACA-box	56	126	7079	0.0006
snRNA:splicing	90	139	12,596	0.0011
snRNA:total	5,950	106	636,630	0.0574
18S_rRNA	117	1,804	211,156	0.0191
28S_rRNA	138	3,726	514,273	0.0464
5.8S_rRNA	143	153	21,961	0.002
5S_rRNA	5,526	111	615,476	0.0555
rRNA:Total	5,924	230	1,362,866	0.123
other	6,882	233	16,06,088	0.1449
Total	19,847	173	2,287,089	0.2063

Table 8. Annotation of Non-Coding RNAs in *H. salicifolia*.

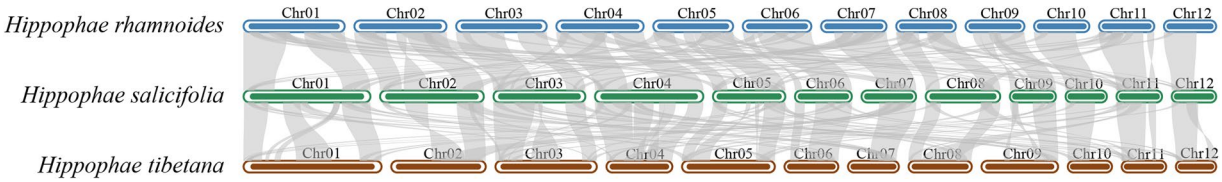

Fig. 4 Genomic synteny relationships between *H. salicifolia* and its closely related species (*H. rhamnoides* and *H. tibetana*). Chromosomes of each species are highlighted in different colors, and the gray lines represent syntenic relationships between genomes.

Genome-wide synteny analysis. The Python version of MCScan implemented in JCVI v1.2.7.52[59] (default parameters) was employed to examine genomic synteny between *H. salicifolia* and its close relatives. The resulting synteny maps support the assessment of structural accuracy and completeness of the assembled genome through comparative analysis (Fig. 4).

Data Records

The raw sequencing data have been deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under accession number SRP548553[60], including PacBio HiFi reads, Illumina PE150 reads, Hi-C reads, and RNA-seq data from various tissues. The final chromosome-scale assembled genome has been deposited in the NCBI GenBank under accession number JBJWFA000000000[61]. In addition, the genome assembly and annotation files have been stored in the Figshare database[62].

Technical Validation

Several strategies were employed to assess the quality of the genome. The completeness of the non-redundant draft genome was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO v5.4.5)[63] with the embryophyta odb10 dataset (1,614 single-copy genes) under default parameters. At the assembly level, BUSCO analysis showed that 98.8% of BUSCO genes were complete (89.8% single-copy, 9.0% duplicated), with only 0.4% fragmented and 0.8% missing (Table 9). At the annotation level, 99.3% of BUSCO genes in the predicted protein-coding gene set were complete (70.6% single-copy, 28.7% duplicated), with only 0.1% fragmented and 0.6% missing. These findings indicate that both the assembly and annotation are highly complete, meeting the standards of a high-quality reference genome.

To validate the accuracy and completeness of the assembly, both Illumina short reads and PacBio HiFi long reads were mapped back to the assembled reference genome. Using BWA-MEM v2.21[32], 97.82% of Illumina reads aligned successfully. With Minimap2 v2.17[64], the mapping rate of PacBio HiFi reads reached 98.53%. These high mapping rates reinforce the accuracy and completeness of the assembled genome.

BUSCOs	assembly (%prop)	annotation (%prop)
Complete (C)	1,595 (98.8)	1,603 (99.3)
Complete and single-copy (S)	1,450 (89.8)	1,139 (70.6)
Complete and duplicated (D)	145 (9.0)	4,64 (28.7)
Fragmented (F)	6 (0.4)	2 (0.1)
Missing (M)	13 (0.8)	9 (0.6)
Total	1,614 (100.0)	1,614 (100.0)

Table 9. BUSCO assessment results for *H. salicifolia* genome assembly and annotation.

Chromosome	No. of Error k-mers	Assembly Length (bp)	QV	Error Rate
Chr01	469,392	162,205,873	38.60	1.38×10^{-4}
Chr02	366,496	132,698,784	38.80	1.32×10^{-4}
Chr03	444,957	117,470,248	37.43	1.81×10^{-4}
Chr04	395,132	138,496,531	38.66	1.36×10^{-4}
Chr05	316,570	92,497,165	37.87	1.63×10^{-4}
Chr06	363,272	72,683,796	36.22	2.39×10^{-4}
Chr07	280,233	69,720,852	37.17	1.92×10^{-4}
Chr08	265,506	95,287,104	38.77	1.33×10^{-4}
Chr09	244,115	58,791,759	37.03	1.98×10^{-4}
Chr10	212,423	52,774,315	37.17	1.92×10^{-4}
Chr11	261,390	58,169,544	36.69	2.14×10^{-4}
Chr12	238,247	56,823,885	36.99	2.00×10^{-4}

Table 10. Chromosome-level QV and estimated error rates of the *H. salicifolia* genome assembly.

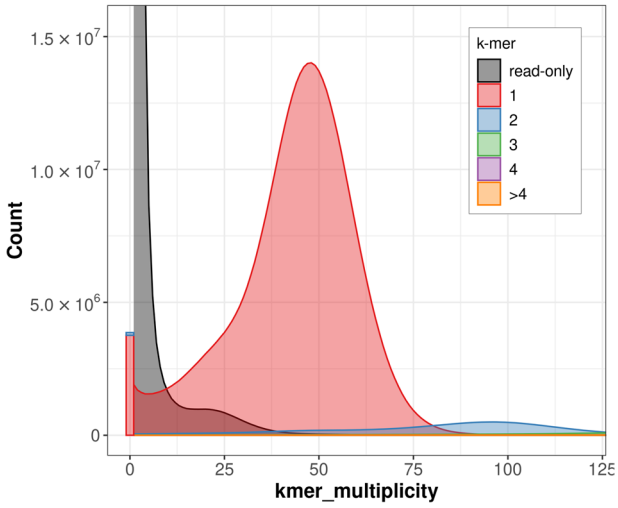

Fig. 5 K-mer spectra-cn plot generated by Merqury showing the distribution of 21-mers from Illumina reads according to their copy number in the assembly. The dominant red peak (~multiplicity 50) represents k-mers that appear once in the assembly, indicating well-represented unique sequences. Gray areas (read-only) correspond to k-mers present in the reads but absent in the assembly, reflecting sequencing errors or unassembled regions. The low abundance of blue (copy = 2) and other multi-copy k-mers suggests low redundancy and high assembly accuracy.

We further assessed the base-level consensus accuracy of the assembled genome using Merqury v1.3[65], based on 21-mers derived from Illumina short reads. Merqury calculated the consensus quality value (QV) for each chromosome individually. The QV scores ranged from 36.22 to 38.80 across all 12 chromosomes, corresponding to base-level error rates between 1.32×10^{-4} and 2.39×10^{-4}, or approximately one error per 750 to 1,500 bases (Table 10). These results demonstrate high base-level accuracy in the genome assembly.

In addition, a k-mer spectra-cn plot was generated (Fig. 5) to visualize the multiplicity distribution of 21-mers from Illumina reads compared to their presence in the assembly. The plot exhibited a dominant peak centered around multiplicity 50, with very few low-frequency or missing k-mers. This indicates that the vast majority of high-quality k-mers from the raw reads were accurately incorporated into the assembly, further supporting its high completeness and low redundancy.

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: 25 December 2024; Accepted: 18 August 2025;
Published online: 28 August 2025

References

1. Ciesarová, Z. *et al*. Why is sea buckthorn (Hippophae rhamnoides L.) so exceptional? A review. *Food Res. Int.* **133**, 109170 (2020).
2. Nybom, H., Ruan, C. & Rumpunen, K. The Systematics, Reproductive Biology, Biochemistry, and Breeding of Sea Buckthorn—A Review. *Genes* **14**, 2120 (2023).
3. Huss-Danell, K., Roelofsen, W., Akkermans, A. D. & Meijer, P. Carbon metabolism of Frankia spp. in root nodules of Alnus glutinosa and Hippophaë rhamnoides. *Physiol. Plant.* **54**, 461–466 (1982).
4. Suryakumar, G. & Gupta, A. Medicinal and therapeutic potential of Sea buckthorn (Hippophae rhamnoides L.). *J. Ethnopharmacol.* **138**, 268–278 (2011).
5. Pant, M., Lal, A. & Rani, A. Hippophae salicifolia D. Don–A plant with multifarious benefits. *Int. J. Pharm. Pharm. Sci.* **6**, 37–40 (2014).
6. Jia, D. R. & Bartish, I. V. Climatic changes and orogeneses in the Late Miocene of Eurasia: The main triggers of an expansion at a continental scale? *Front. Plant Sci.* **9**, 1400 (2018).
7. Zeng, Z. *et al*. Development and validation of sex-linked molecular markers for rapid and accurate identification of male and female Hippophae tibetana plants. *Sci. Rep.* **14**, 19243 (2024).
8. Puterova, J. *et al*. Satellite DNA and transposable elements in seabuckthorn (Hippophae rhamnoides)., a dioecious plant with small Y and large X chromosomes. *Genome Biol. Evol.* **9**, 197–212 (2017).
9. Zeng, Z. *et al*. Development of sex-specific molecular markers for early sex identification in Hippophae gyantsensis based on whole-genome resequencing. *BMC Plant Biol.* **24**, 1187 (2024).
10. Zeng, Z. *et al*. Development and application of sex-specific indel markers for Hippophae salicifolia based on third-generation sequencing. *BMC Plant Biol.* **25**, 692 (2025).
11. Yu, L. *et al*. Genome sequence and population genomics provide insights into chromosomal evolution and phytochemical innovation of Hippophae rhamnoides. *Plant Biotechnol. J.* **20**, 1257–1273 (2022).
12. Wu, Z. *et al*. Genome of Hippophae rhamnoides provides insights into a conserved molecular mechanism in actinorhizal and rhizobial symbioses. *New Phytol.* **235**, 276–291 (2022).
13. Yang, X. *et al*. Chromosome-level genome assembly of Hippophae rhamnoides variety. *Sci. Data.* **11**, 776 (2024).
14. Chen, M. *et al*. Chromosome-level genome assembly of Hippophae gyantsensis. *Sci. Data.* **11**, 126 (2024).
15. Wang, R. *et al*. How to survive in the world's third poplar: Insights from the genome of the highest altitude woody plant, Hippophae tibetana (Elaeagnaceae). *Front. Plant Sci.* **13**, 1051587 (2022).
16. Zhang, G. *et al*. Chromosome-level genome assembly of Hippophae tibetana provides insights into high-altitude adaptation and flavonoid biosynthesis. *BMC biol.* **22**, 82 (2024).
17. Zhou, W., Dong, Q., Wang, H. & Hu, N. Characterization of the complete chloroplast genome of Hippophae tibetana. *Mitochondrial DNA B* **5**, 593–594 (2020).
18. Wang, L., Wang, J., He, C., Zhang, J. & Zeng, Y. Characterization and comparison of chloroplast genomes from two sympatric Hippophae species (Elaeagnaceae). *J. For. Res.* **32**, 307–318 (2021).
19. Zhang, Q., Li, J., Yu, Y. & Xu, H. A new map of the chloroplast genome of Hippophae based on inter-and intraspecific variation analyses of 13 accessions belonging to eight Hippophae species. *Braz. J. Bot.* **46**, 367–382 (2023).
20. Zeng, Z. *et al*. Complete mitochondrial genome of Hippophae tibetana: insights into adaptation to high-altitude environments. *Front. Plant Sci.* **15**, 1449606 (2024).
21. Zeng, Z. *et al*. Assembly and Comparative Analysis of the Complete Mitochondrial Genome of Hippophae salicifolia. *Biology* **14**, 448 (2025).
22. Sharma, A., Zinta, G., Rana, S. & Shirko, P. Molecular identification of sex in Hippophae rhamnoides L. using isozyme and RAPD markers. *Forestry Studies in China* **12**, 62–66 (2010).
23. Zhou, W. *et al*. Molecular sex identification in dioecious Hippophae rhamnoides L. via RAPD and SCAR markers. *Molecules* **23**, 1048 (2018).
24. Wang, Z. *et al*. Phytochemistry, health benefits, and food applications of sea buckthorn (Hippophae rhamnoides L.): A comprehensive review. *Front. Nutr.* **9**, 1036295 (2022).
25. Ma, Q. G. *et al*. Hippophae rhamnoides L.: A comprehensive review on the botany, traditional uses, phytonutrients, health benefits, quality markers, and applications. *J. Agric. Food. Chem.* **71**, 4769–4788 (2023).
26. Liang, J., Zhang, G., Song, Y., He, C. & Zhang, J. Targeted metabolome and transcriptome analyses reveal the pigmentation mechanism of Hippophae (sea buckthorn). fruit. *Foods* **11**, 3278 (2022).
27. Cui, M., Hu, P., Wang, T., Tao, J. & Zong, S. Differential transcriptome analysis reveals genes related to cold tolerance in seabuckthorn carpenter moth, Eogystia hippophaecolus. *PLoS One* **12**, e0187105 (2017).
28. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2014).
29. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024).
30. Sahu, S. K., Thangaraj, M. & Kathiresan, K. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *Int. Sch. Res. Not.* **2012**, 1–6 (2012).
31. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
32. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
33. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
34. 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).
35. Jung, Y. & Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **38**, 2404–2413 (2022).
36. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
37. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
38. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
39. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
40. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457 (2020).
41. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 1–6 (2015).

42. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **25**, 4–10 (2009).
43. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinf.* **19**, 1–12 (2018).
44. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 1–13 (2019).
45. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinf.* **7**, 1–11 (2006).
46. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
47. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
48. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
49. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
50. Chen, C. *et al*. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol Plant.* **16**, 1733–1742 (2023).
51. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Res.* **43**, 7762–7768 (2015).
52. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
54. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (2000).
55. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
56. 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).
57. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
58. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
59. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP548553 (2024).
61. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046201785.1 (2024).
62. Zeng, Z. *et al*. Chromosome-level genome assembly of Hippophae salicifolia. *figshare*. https://doi.org/10.6084/m9.figshare.27987659.v3 (2024).
63. 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).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. 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).

Acknowledgements

This work was supported by the Science and Technology Projects of Xizang Autonomous Region, China (Project Nos. XZ202402ZD0005, XZ202402ZY0023, and XZ202402JX0003).

Author contributions

L.Q., W.L., W.Z., J.W. and Z.Z. conceived and designed the study. C.M., J.Z., M.X., and Z.T. prepared the materials, conducted the experiments, and contributed to the data collection. Z.Z. performed the data analysis, prepared the initial draft of the manuscript, and managed the figures and tables. J.W., W.Z., and L.Q. reviewed and revised the manuscript for intellectual content. All authors participated in improving the manuscript and unanimously approved the final version for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.W., W.Z. or L.Q.