scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A near telomere-to-telomere chromosome-level genome assembly of *Rhodiola yunnanensis* (Crassulaceae)

Mingcheng Wang[1,2], Panyue Du[3], Chengheng Tong[4,5], Jinyi Zhao[5,6], Xinrui Song[4,5] & Ning Miao[4,5] ✉

Rhodiola yunnanensis **(Crassulaceae) is an endemic medicinal plant species distributed in the Hengduan Mountains of southwestern China. Here, we present a high-quality and near telomere-to-telomere chromosome-level genome assembly of *R. yunnanensis* generated using Oxford Nanopore long-read sequencing and Hi-C scaffolding. The assembled genome spans 643.28 Mb with a contig N50 of 51.92 Mb, and 98.46% of the sequence was anchored to 11 pseudo-chromosomes, eight of which contained no internal gaps in the current assembly. Benchmark analyses revealed a BUSCO completeness of 97.46%, while the LTR Assembly Index (LAI) score achieved 20.50, both suggesting a high level of assembly quality. Telomeric repeat motifs were detected at both ends of 10 pseudo-chromosomes, and centromeric regions were identified on all 11 chromosomes. We predicted 36,495 protein-coding genes, 94.16% of which were functionally annotated, together with 1,643 transcription factors and 6,429 non-coding RNAs. This high-quality genome provides a valuable resource for studying alpine adaptation, specialized metabolite biosynthesis, and evolutionary diversification in *Rhodiola* and other high-altitude plants.**

Background & Summary

The genus *Rhodiola* L. (Crassulaceae) currently comprises more than 70 accepted species according to the GBIF Backbone Taxonomy (https://www.gbif.org/), most of which (approximately 90%) are distributed in alpine and subalpine regions of China[1]. *Rhodiola* species are perennial herbaceous or subshrub plants, typically characterized by thick, fleshy rhizomes and succulent leaves, which enable them to survive and adapt to harsh alpine environments[2,3]. Their rhizomes accumulate a wide range of bioactive compounds, among which salidroside and rosavin are the most characteristic and pharmacologically significant constituents[4]. Due to these bioactive compounds, *Rhodiola* species have been widely utilized in both traditional and modern medicine, particularly *R. rosea*[5], which is commonly known as "Tibetan ginseng." The biosynthetic pathways of salidroside and rosavin have been extensively investigated and applied in the field of synthetic biology[6–9]. However, the contents of these compounds vary greatly among *Rhodiola* species[10], and the molecular mechanisms underlying these variations remain largely unresolved.

Ecologically, *Rhodiola* species have evolved strong tolerance to cold, ultraviolet radiation, and hypoxia, making them excellent models for studying the genetic basis of high-altitude adaptation. In addition, the rapid radiation and diversification of *Rhodiola* have been largely driven by the uplift of the Qinghai-Tibet Plateau[11]. The overlapping geographic distributions among species make this genus a model system for investigating speciation and adaptive

[1]Institute for Advanced Study, Chengdu University, No. 2025 Chengluo Road, Chengdu, 610106, China. [2]Sichuan-Xizang Medicinal Resource Breeding and Standardization Team, Engineering Research Center of Sichuan-Xizang Traditional Medicinal Plant, Chengdu University, Chengdu, 610106, China. [3]School of Food and Biological Engineering, Chengdu University, Chengdu, 610106, China. [4]Chengdu Botanical Garden, Chengdu, 610503, Sichuan, China. [5]Southwest Bio-resources R&D Key Laboratory of Sichuan Province, Chengdu Botanical Garden-Sichuan University Joint Laboratory for Ex Situ Conservation and Resource Utilization of Mountain Plants, College of Life Sciences, Sichuan University, Chengdu, 610064, Sichuan, China. [6]Environmental Engineering Assessment Center of Xizang Autonomous Region, Lasa, 850000, Xizang, China. ✉e-mail: miaoning@scu.edu.cn

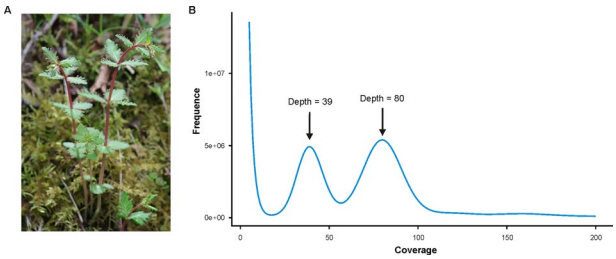

Fig. 1 Genome survey of *R. yunnanensis*. (**A**) Photograph of the *R. yunnanensis* individual sequenced in this study. (**B**) The 21-mer frequency distribution of the DNBSEQ-T7 reads, showing the major homozygous peak at depth 80 and the heterozygous peak at depth 39. The genome size was estimated by dividing the total number of *k*-mers by the depth of the homozygous peak, while the heterozygosity rate was calculated based on the proportion of *k*-mers under the heterozygous peak.

Data type	Usage	Tissue	Number of reads	Total data (Gb)	Genome coverage (×)
DNBSEQ-T7	Genome survey	Leaf	189,317,334	56.80	92.74
Oxford Nanopore	Genome assembly	Leaf	2,904,025	42.84	69.95
DNBSEQ-T7	Hi-C scaffolding	Young leaf	553,118,396	82.97	135.47
RNA-seq (Total)	Gene prediction	—	126,270,106	18.92	—
RNA-seq		Leaf	42,946,846	6.43	—
RNA-seq		Stem	38,323,222	5.74	—
RNA-seq		Fruit	45,000,038	6.74	—

Table 1. Overview of DNA and RNA sequencing data generated in this study.

evolution[12,13]. Moreover, due to excessive harvesting, habitat degradation, and threats from climate change, several *Rhodiola* species have been listed as Class II nationally protected plants in China[14]. Genome sequencing not only provides an opportunity to explore the molecular mechanisms underlying their adaptive traits and evolutionary history[15] but is also of great importance for their conservation and sustainable utilization[16]. However, to date, only a draft genome of *R. crenulata*[17] and a few chromosome-level assemblies from other *Rhodiola* species[18,19] have been published, which remain insufficient to address these biological and conservation-related questions.

R. yunnanensis, an endemic and medicinally essential species native to the Hengduan Mountains of southwestern China, is a crucial component of the alpine flora in this biodiversity hotspot. In this study, we generated a high-quality and near telomere-to-telomere genome assembly of *R. yunnanensis* using Oxford Nanopore Technologies (ONT) long-read sequencing combined with Hi-C scaffolding. We further performed comprehensive and high-quality genome annotation, providing a solid foundation for downstream functional and comparative analyses. This chromosome-scale genome offers a valuable resource for elucidating the molecular mechanisms underlying alpine adaptation and specialized metabolite biosynthesis in the genus *Rhodiola*. Moreover, it enables the identification of key functional genes, the discovery of novel biosynthetic gene clusters, and the exploration of potentially new bioactive compounds. Additionally, this high-quality genome provides a crucial genomic resource for investigating speciation, diversification, and comparative genomics in *Rhodiola* plants.

Methods

Plant sample preparation. Fresh leaves of *R. yunnanensis* were collected near the Dengsheng Protection Station, Wenchuan County, Sichuan Province, southwest China (30.8843° N, 102.9744° E; altitude 3,428 m; Fig. 1A). The leaves were rinsed with distilled water, immediately frozen in liquid nitrogen, and stored at −80 °C for subsequent DNA extraction. For RNA sequencing (RNA-seq), additional leaf, stem, and fruit tissues from the same individual were also collected, flash-frozen in liquid nitrogen, and stored at −80 °C.

Genome survey. High-quality genomic DNA was extracted using the CTAB method[20], assessed for purity and integrity, fragmented, end-repaired, and adapter-ligated for library construction. The libraries were sequenced on the DNBSEQ-T7 platform (MGI Tech Co., Ltd., Shenzhen, China), yielding 56.80 Gb of paired-end reads (Table 1). The 21-mer frequency distribution of the DNBSEQ-T7 clean reads was analyzed using Jellyfish v1.1.12[21], producing a total of 48,994,969,975 *k*-mers with a major peak at a depth of 80 and a heterozygous peak at 39 (Fig. 1B). The haploid genome size was estimated by dividing the total number of *k*-mers by the depth of the major peak, yielding an estimated genome size of approximately 612.44 Mb. The heterozygosity rate, calculated based on the proportion of *k*-mers under the heterozygous peak relative to the total distinct *k*-mers and adjusted for *k*-mer length, resulted in an overall rate of approximately 1.65%. Informed by these survey results, we designed an assembly strategy integrating ONT reads for contig generation, followed by redundant haplotype removal and high-coverage Hi-C-based chromosome anchoring.

Species identification confirmation. To confirm the species identity of the sequenced individual, the complete chloroplast genome was assembled from the DNBSEQ-T7 reads using GetOrganelle v1.7.7.1[22] with

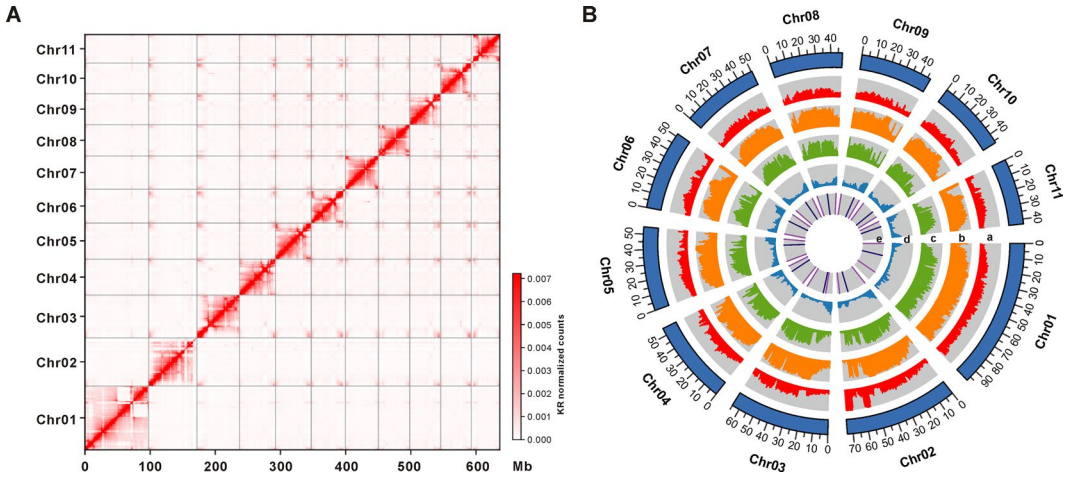

Fig. 2 Chromosome-level genome features of *R. yunnanensis*. (**A**) Hi-C interaction heatmap of the 11 chromosomes of *R. yunnanensis* at a 500-kb resolution. (**B**) Circos plot showing genome features from outer to inner circles: GC content, repeat sequence density, LTR-RT density, gene density, and distribution of telomeres (purple) and centromeres (dark blue).

default parameters. The assembled chloroplast genome was then compared with the previously published *R. yunnanensis* chloroplast genome (GenBank accession number: MN794332.1) and those of other representative *Rhodiola* species available in GenBank. Phylogenetic analysis was performed based on the protein sequences of single-copy genes using MAFFT-LINSI v7.313[23] for multiple sequence alignment and RAxML v8.2.11[24] for maximum likelihood phylogeny reconstruction under the GTRGAMMA model with 1000 bootstrap replicates. The resulting phylogenetic tree placed our sample in the same clade as the reference *R. yunnanensis* with 100% bootstrap support, confirming the species identification.

Genome assembly. For ONT sequencing, high-quality genomic DNA was extracted using a modified CTAB method[20]. DNA purity and concentration were assessed using a Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), and DNA integrity was evaluated by agarose gel electrophoresis. Library preparation was performed using the SQK-LSK114 ligation sequencing kit (Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's instructions. Briefly, high-molecular-weight DNA fragments were size-selected using magnetic beads, followed by damage repair, end repair, and adapter ligation. The final library was purified using magnetic beads and quantified with a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA). The library was then loaded onto an R10.4.1 flow cell and sequenced on the PromethION platform (Oxford Nanopore Technologies, Oxford, UK).

Nanopore raw reads were filtered to remove adapters and low-quality sequences, resulting in 42.84 Gb of clean reads (69.95 × coverage, N50 = 23.46 kb; Table 1). The ONT reads were *de novo* assembled using hifiasm v0.25.0[25] with the parameters "-l 3–hom-cov 70" under ONT mode. To generate a non-redundant haploid representation of the genome, potential duplicated haplotypes were identified and removed from the initial assembly using Purge Haplotigs v1.1.3[26] with the following parameters: "-l 30 -m 50 -h 100". Subsequently, the chloroplast and mitochondrial genomes assembled from the DNBSEQ-T7 reads using GetOrganelle were aligned to the haplotype-purged assembly using BLASTN v2.16.0[27] to identify and remove organelle-derived contigs. The ONT reads were then aligned to the remaining contigs using minimap2 v2.29[28], followed by two iterative rounds of polishing with Racon v1.5.0[29] to improve the assembly accuracy. Finally, the DNBSEQ-T7 reads were aligned to the polished assembly and used for further error correction with NextPolish v1.4.1[30]. After these steps, the polished draft genome assembly represents a non-redundant haploid primary assembly, with alternate haplotypes removed, and consisted of 144 contigs, with a total length of 643.28 Mb, a contig N50 of 51.92 Mb, and a maximum contig length of 74.02 Mb.

To perform Hi-C scaffolding analysis, Hi-C libraries were prepared from approximately 2 g of young leaves through chromatin fixation, restriction digestion, biotin labeling, ligation, and DNA purification. They were subsequently sequenced on the DNBSEQ-T7 platform. A total of 82.97 Gb of Hi-C reads (135.47 × coverage; Table 1) were generated and aligned to the draft assembly using BWA v0.7.18[31] with the parameters "bwa mem -5SP". Duplicate reads were identified and marked using Samblaster v0.1.26[32], and duplicate, low-quality, secondary, and supplementary alignments were filtered out using Samtools v1.3[33] with the parameter "view -F 3340". The filtered BAM files were subsequently used for chromosome-level scaffolding with HapHiC v1.0.6[34] to anchor and orient contigs into 11 pseudo-chromosomes, which includes algorithms to identify and correct potential misassemblies. Hi-C contact maps were then visualized using Juicebox v2.12.00[35] to manually inspect the assembly and confirm the accuracy of contig placements and orientations. In total, 98.46% (633.37 Mb) of the genome sequences were anchored to 11 pseudo-chromosomes, ranging from 43.19 Mb (Chr11) to 97.95 Mb (Chr01) (Fig. 2A; Table 2). Eight pseudo-chromosomes contained no internal gaps, while Chr01 and Chr08 contained two gaps each, and Chr02 harbored ten gaps. The final assembly demonstrated outstanding contiguity, with contig and scaffold N50 values of 51.92 Mb and 54.63 Mb, respectively (Fig. 2B; Table 3). Although no genome size or chromosome number has been previously reported for *R. yunnanensis*, the assembled genome size (643.28 Mb) falls within the diploid *Rhodiola* range (345–654 Mb)[17,18]. Notably, both the chromosome

Chromosome	Length (bp)	Genes	Centromere (bp)	Left telomere (bp)	Right telomere (bp)
Chr01	97,946,844	4,027	25,285,061–32,568,642	3–7,128	74,004,206–74,018,478
Chr02	73,796,112	4,807	43,963,240–50,305,690	1,088–10,082	68,084,218–68,095,466
Chr03	65,166,032	3,039	17,326,407–26,722,491	16,024,913–16,025,010	65,152,444–65,155,838
Chr04	54,881,975	3,032	36,311,684–43,257,855	2,757–5,787	54,866,608–54,875,917
Chr05	54,634,982	3,436	36,532,963–42,993,357	5,490–11,873	54,624,529–54,634,979
Chr06	51,923,441	3,292	31,200,640–35,055,158	893–5,225	51,902,078–51,909,721
Chr07	50,517,994	2,689	30,400,581–32,592,782	3,765–7,999	50,507,642–50,513,633
Chr08	47,631,670	2,660	24,102,057–27,714,362	7–10,198	47,625,529–47,631,667
Chr09	47,153,315	3,101	29,894,086–36,331,335	1,420–22,986	47,130,985–47,148,218
Chr10	46,534,287	2,964	29,075,855–33,267,312	7–3,702	46,521,171–46,527,673
Chr11	43,185,641	2,749	11,326,735–15,917,995	226–6,000	—
Total	633,372,293	35,796	—	—	—

Table 2. Detailed statistics of the 11 pseudo-chromosomes in the *R. yunnanensis* genome.

Assembly	
Estimated genome size (Mb)	612.44
Total length (Mb)	643.28
Number of pseudo-chromosomes	11
Total length of pseudo-chromosomes (Mb)	633.37
Number of contigs	144
Number of scaffolds	130
Number of gaps	14
Contig N50 (Mb)	51.92
Scaffold N50 (Mb)	54.63
Longest contig (Mb)	74.02
Shortest contig (bp)	35,038
LTR assembly index	20.50
BUSCO completeness (%)	97.46
Annotation	
GC content (%)	41.16
Repeat content (%)	69.89
Number of protein-coding genes	36,495
Average gene length (bp)	2,097
Average coding sequence length (bp)	1,110
Average exon length (bp)	209
Average intron length (bp)	228
Functionally annotated genes	34,365
BUSCO completeness (%)	94.93

Table 3. Global statistics for the *R. yunnanensis* genome assembly and annotation.

number (2n = 22) and genome size closely match those of *R. kirilowii* (654 Mb, 2n = 22)[18], a close relative supported by phylogenetic evidence[12], further validating the consistency of our assembly.

Genome annotation. Repetitive elements within the final assembly of *R. yunnanensis* were identified using RepeatModeler v1.0.8[36] and RepeatMasker v4.0.5[37]. A comprehensive repeat library was first constructed by combining the green plant repeat dataset from the Repbase database v22.11[38] with a *de novo* library generated by RepeatModeler. This custom library was then used by RepeatMasker to perform genome-wide identification and classification of repetitive sequences. A total of 449.57 Mb, accounting for 69.89% of the *R. yunnanensis* genome, was identified as repetitive sequences (Table 4). Among them, long terminal repeat retrotransposons (LTR-RTs) were the most abundant, occupying 297.55 Mb (46.25%) of the genome. The remaining repetitive elements comprised unclassified sequences (102.29 Mb), DNA transposons (33.14 Mb), long interspersed nuclear elements (LINEs; 16.72 Mb), short interspersed nuclear elements (SINEs; 0.15 Mb), and other minor repeat types totaling 2.30 Mb (Table 4).

Then, we performed annotation of protein-coding genes (PCGs) based on the repeat-masked genome using three complementary approaches. First, total RNA was extracted from all fresh tissues. High-quality RNA was used to construct mRNA-seq libraries according to the standard protocol for oligo(dT)-based mRNA enrichment. Sequencing was performed on the DNBSEQ-T7 platform, yielding a total of 18.92 Gb of RNA-seq data (Table 1). The RNA-seq reads were *de novo* assembled into transcripts using Trinity v2.8.4[39], and the assembled

Type	Count	Total length (bp)	% of genome
DNA transposons	51,415	33,140,428	5.15
LINEs	27,033	16,717,124	2.60
SINEs	645	145,599	0.02
LTR-RTs	242,388	297,546,415	46.25
Satellites	1,572	477,461	0.07
Simple repeats	8,360	1,819,041	0.28
Low complexity	46	5,887	0.00
Unclassified	170,647	102,287,843	15.90
Total	**502,106**	**449,566,676**	**69.89**

Table 4. Composition and distribution of repetitive elements in the *R. yunnanensis* genome.

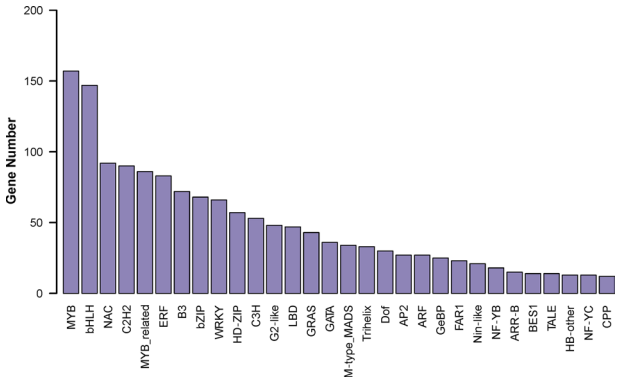

Fig. 3 Distribution of the top 30 transcription factor families identified in the *R. yunnanensis* genome.

transcripts were aligned back to the genome using PASA v2.3.3[40] to predict gene structures based on transcript alignments. Second, protein sequences from several closely related or representative species, including *Rhodiola crenulata*[17], *R. kirilowii*[19], *Sedum album*[41], *Kalanchoe fedtschenkoi*[42], *Malus domestica*[43], and *Arabidopsis thaliana*[44], were aligned to the *R. yunnanensis* genome using TBLASTN v2.16.0[25]. GeneWise v2.4.1[45] was subsequently applied to generate accurate gene models based on these protein alignments. Third, high-quality gene models containing at least three exons and with total lengths greater than 1,500 bp were selected from the PASA predictions and used to train species-specific parameters for *R. yunnanensis* in AUGUSTUS v3.2.3[46]. A *de novo* gene prediction was then performed using AUGUSTUS with the pre-trained parameters. Finally, gene models derived from the three approaches were integrated into a consensus, non-redundant gene set using EvidenceModeler v1.1.1[47], with all resulting gene models retained without additional filtering based on gene length or exon number.

We predicted a total of 36,495 PCGs in the *R. yunnanensis* genome, of which 35,796 (98.08%) were anchored to the 11 pseudo-chromosomes (Table 2), corresponding to an average gene density of 56.5 genes per Mb. The predicted genes exhibited average transcript, coding sequence, exon, and intron lengths of 2,097 bp, 1,110 bp, 209 bp, and 228 bp, respectively, with an average of 5.32 exons per gene (Table 3). To validate the predicted gene structures, RNA-seq reads from leaf tissue were aligned to the genome using HISAT2 v2.2.1[48]. Splice junctions were extracted using RegTools v1.0.0[49] with the parameters "-a 8 -m 50 -M 500000". Of the 157,134 predicted introns, 143,948 (91.60%) were supported by at least one junction read, indicating high reliability of the predicted exon-intron boundaries. To assess potential contamination by transposable elements (TEs), we examined the overlap between predicted genes and annotated repeat regions. A total of 6,175 genes (16.92%) overlapped with repeat regions. Of these, 3,938 genes (63.77%) were supported by at least five RNA-seq reads, and 5,808 genes (94.06%) were functionally annotated(see below), indicating that the majority of repeat-overlapping genes are actively transcribed and encode functional protein domains. Furthermore, 6,799 genes (18.63% of all PCGs) contained TE-related domains (e.g., reverse transcriptase, integrase, transposase). Among these TE-domain-containing genes, 5,937 (87.32%) were supported by at least five RNA-seq reads, and 6,262 genes (92.10%) contained non-TE functional domains. Collectively, these analyses demonstrate the robustness of the gene annotation against TE contamination, with the majority of repeat-overlapping and TE-related genes showing evidence of expression and functional significance. A total of 1,643 transcription factor genes belonging to 57 subfamilies were annotated using PlantTFDB v5.0[50], with members of the MYB (157; 9.6%), bHLH (147; 8.9%), and NAC (92; 5.6%) families representing the largest groups (Fig. 3). Furthermore, functional annotation of all PCGs was performed by aligning their protein sequences against several publicly available databases, including Swiss-Prot, TrEMBL[51], InterPro[52], Gene Ontology (GO)[53], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[54]. Overall, 29,454 (80.71%) genes were functionally annotated by the Swiss-Prot and TrEMBL databases, 33,508 (91.80%) genes contained conserved domains identified by InterProScan, 27,576 (75.56%)

	Number of genes	Percentage (%)
Total	36,495	—
Annotated	34,365	94.16
InterPro	33,504	91.80
KEGG	9,923	27.19
SwissProt	21,700	59.46
TrEMBL	29,413	80.59
GO	27,576	75.56
Unannotated	2,130	5.84

Table 5. Functional annotation statistics of *R. yunnanensis* genes.

Type	Number	Average length (bp)	Total length (bp)
tRNA	494	74.07	36,590
miRNA	87	120.92	10,520
rRNA	4,500	1,402.47	6,311,115
28S	1,302	3,037.28	3,954,538
18S	1,180	1,759.82	2,076,590
5.8S	1135	154.96	175,881
5S	883	117.90	104,106
snRNA	1,348	110.90	149,497
CD-box	1,094	105.28	115,174
HACA-box	38	126.76	4,817
Splicing	216	136.60	29,506

Table 6. Non-coding RNA annotation statistics of *R. yunnanensis*.

genes were assigned to GO terms, and 9,923 (27.19%) genes were mapped to KEGG pathways (Table 5). In addition, we identified a total of 494 tRNAs using tRNAscan-SE v2.0.12[55], and 87 miRNAs, 4,500 rRNAs, and 1,348 snRNAs using the cmscan program of Infernal v1.1.5[56] by searching against the Rfam database[57] (Table 6).

Data Records

The genome assembly and raw sequencing data of *R. yunnanensis* have been deposited in the NCBI database under BioProject accession PRJNA1353568[58]. The genome assembly is available in GenBank under accession number JBVKEA000000000[59]. The DNBSEQ-T7, ONT, and Hi-C data are accessible in the Sequence Read Archive (SRA) under accession numbers SRR35889594[60], SRR35888830[61], and SRR35896369[62], respectively. The RNA-seq data have been deposited in the SRA under accession numbers SRR35896530–SRR35896532[63–65]. Additionally, the final genome assembly and annotations are available in Figshare[66].

Technical Validation

The quality of the final *R. yunnanensis* genome assembly was assessed using four complementary approaches. First, the DNBSEQ-T7 reads generated for genome survey analysis were mapped to the assembly using BWA-MEM[31], achieving a mapping rate of 96.22%. The assembly showed a genome coverage of 98.46%, with 97.60% of the genome covered at a depth of at least 20×. Second, BUSCO analysis (v5.8.2)[67] was performed against the Embryophyta odb10 database. The genome assembly showed 97.46% complete BUSCOs (88.29% single-copy and 9.17% duplicated), with 1.05% fragmented and 1.49% missing. The annotated gene set showed 94.93% complete BUSCOs (87.31% single-copy and 7.61% duplicated), with 2.35% fragmented and 2.72% missing. Third, an LTR Assembly Index (LAI) score of 20.50 was obtained using LTR_retriever v3.0.1[68], suggesting a high-quality assembly (Table 3). In addition, we examined the distribution of *Arabidopsis*-type telomeric repeats (TTTAGGG)_n and found that 10 out of the 11 pseudo-chromosomes possessed telomeric sequences at both ends, whereas Chr11 lacked telomeric repeats at one end (Fig. 2B; Table 2). Furthermore, quarTeT analysis (v1.2.5)[69], which identifies centromeres based on the enrichment of tandem repeats and LTR-RTs while integrating repeat and gene annotations, predicted putative centromeric regions on all 11 pseudo-chromosomes (Fig. 2B; Table 2). Finally, Merqury analysis (v1.3)[70] showed a QV of 31.60 and an error rate of 0.07%, indicating high accuracy in the genome assembly. The *k*-mer completeness was 84.76%, likely due to the high repeat content (69.89%) and heterozygosity (1.65%) of the *R. yunnanensis* genome, which can reduce *k*-mer representation in short-read data. Overall, these evaluations support the high-quality and near-complete nature of the *R. yunnanensis* genome assembly.

Data availability

The genome assembly of *Rhodiola yunnanensis* is available in GenBank under accession number JBVKEA000000000. All associated sequencing data are available under NCBI BioProject PRJNA1353568. The final genome assembly and annotation files are available on Figshare at https://doi.org/10.6084/m9.figshare.30556088.

Code availability

No specific script was generated in this study.

Received: 30 November 2025; Accepted: 9 March 2026;
Published online: 18 March 2026

References

1. Fu, K. T. & Ohba, H. in *Flora of China* Vol. 8 (eds Wu, C. Y. & Raven, P. H.) 202–268 Science Press, Beijing & Missouri Botanical Garden Press, St. Louis (2001).
2. Lubbe, F. C., Rees, M. & Klimešová, J. Below-ground plant functional ecology: towards an integrated perspective on plant ecology and evolution. *Funct. Ecol.* **35**, 728–739 (2021).
3. Males, J. Secrets of succulence. *J. Exp. Bot.* **68**, 2121–2132 (2017).
4. Chiang, H. M., Chen, H. C., Wu, C. S., Wu, P. Y. & Wen, K. C. *Rhodiola* plants: chemistry and biological activity. *J. Food Drug Anal.* **23**, 359–369 (2015).
5. Bernatonienė, J., Jakštas, V. & Kopustinskienė, D. M. Phenolic compounds of *Rhodiola rosea* L. as the potential alternative therapy in the treatment of chronic diseases. *Int. J. Mol. Sci.* **24**, 12293 (2023).
6. Torrens-Spence, M. P., Pluskal, T., Li, F. S., Carballo, V. & Weng, J. K. Complete pathway elucidation and heterologous reconstitution of *Rhodiola* salidroside biosynthesis. *Mol. Plant* **11**, 205–217 (2018).
7. Zhou, X. *et al*. Efficient biosynthesis of salidroside via artificial *in vivo* enhanced UDP-glucose system using cheap sucrose as substrate. *ACS Omega* **9**, 22386–22397 (2024).
8. Bi, H. *et al*. Biosynthesis of a rosavin natural product in *Escherichia coli* by glycosyltransferase rational design and artificial pathway construction. *Metab. Eng.* **69**, 15–25 (2022).
9. Li, L., Liu, M., Bi, H. & Liu, T. High-level production of *Rhodiola rosea* characteristic component rosavin from D-glucose and L-arabinose in engineered *Escherichia coli*. *Metab. Eng.* **82**, 274–285 (2024).
10. Li, X. *et al*. Metabolic discrimination of different *Rhodiola* species using ^{1}H-NMR and GEP combinational chemometrics. *Chem. Pharm. Bull.* **67**, 81–87 (2019).
11. Zhang, J. Q., Meng, S. Y., Allen, G. A., Wen, J. & Rao, G. Y. Rapid radiation and dispersal out of the Qinghai-Tibetan Plateau of an alpine plant lineage *Rhodiola* (Crassulaceae). *Mol. Phylogenet. Evol.* **77**, 147–158 (2014).
12. Ren, C. Q., Zhang, D. Q., Liu, X. Y. & Zhang, J. Q. Genomic data provide a robust phylogeny backbone for *Rhodiola* L. (Crassulaceae) and reveal extensive reticulate evolution during its rapid radiation. *Mol. Phylogenet. Evol.* **186**, 107863 (2023).
13. Rieseberg, L. H. & Willis, J. H. Plant speciation. *Science* **317**, 910–914 (2007).
14. Yang, M. *et al*. Predicting the potential geographical distribution of *Rhodiola* L. in China under climate change scenarios. *Plants* **12**, 3735 (2023).
15. Wang, M., Zhang, S., Li, R. & Zhao, Q. Unraveling the specialized metabolic pathways in medicinal plant genomes: A review. *Front. Plant Sci.* **15**, 1459533 (2024).
16. Jackson, S. A., Iwata, A., Lee, S. H., Schmutz, J. & Shoemaker, R. Sequencing crop genomes: Approaches and applications. *New Phytol.* **191**, 915–925 (2011).
17. Fu, Y. *et al*. Draft genome sequence of the Tibetan medicinal herb *Rhodiola crenulata*. *Gigascience* **6**, gix033 (2017).
18. Zhang, D. Q. *et al*. Two chromosome-level genome assemblies of *Rhodiola* shed new light on genome evolution in rapid radiation and evolution of the biosynthetic pathway of salidroside. *The Plant J.* **117**, 464–482 (2024).
19. Zhang, J. *et al*. Chromosome-level genome and annotation of the tetraploid *Rhodiola kirilowii*. *Sci. Data* **12**, 620 (2025).
20. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytoch. Bull.* **19**, 11–15 (1987).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
22. Jin, J. J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
23. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2018).
24. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
25. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
26. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
27. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
29. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
30. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Faust, G. G. & Hall, I. M. *SAMBLASTER*: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
33. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
35. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
36. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinf.* **5**, 4–10 (2004).
38. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* **110**, 462–467 (2005).
39. Haas, B. J. *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
40. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
41. Wai, C. M. *et al*. Time of day and network reprogramming during drought-induced CAM photosynthesis in *Sedum album*. *PLoS Genet.* **15**, e1008209 (2019).
42. Yang, X. *et al*. The *Kalanchoë* genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nat. Commun.* **8**, 1899 (2017).

43. Daccord, N. *et al*. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
44. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
45. Birney, E., Clamp, M. & Durbin, R. GeneWise and GenomeWise. *Genome Res.* **14**, 988–995 (2004).
46. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
47. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
48. 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).
49. Cotto, K. C. *et al*. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. *Nat Commun.* **14**, 1589 (2023).
50. Jin, J. *et al*. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**, D1040–D1045 (2017).
51. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
52. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
53. Gene Ontology Consortium The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
54. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
55. 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).
56. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
57. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
58. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1353568 (2025).
59. Wang, M. *et al*. *GenBank* https://identifiers.org/ncbi/insdc:JBVKEA000000000 (2025).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35889594 (2025).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35888830 (2025).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35896369 (2025).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35896530 (2025).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35896531 (2025).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35896532 (2025).
66. Wang, M. A near telomere-to-telomere chromosome-level genome assembly of *Rhodiola yunnanensis* (Crassulaceae). *Figshare* https://doi.org/10.6084/m9.figshare.30556088 (2025).
67. 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).
68. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
69. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res.* **10**, uhad127 (2023).
70. 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 research was jointly funded by China's National Natural Science Foundation (U24A20355), National Key R&D Program of China (Grant No. 2022YFD2200504), and the National Key Research & Development Program of China (2016YFD0600203).

Author contributions

N.M. and M.W. designed the study. M.W., P.D., C.T., J.Z. and X.S. performed the data analyses and drafted the manuscript. M.W. and N.M. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.