



OPEN

DATA DESCRIPTOR

Chromosome-level Genome Assembly of the Tibetan Medicinal Plant Duiyiwei (*Phlomoides rotata*)

Zhongqiong Tian^{1,2}, Junwei Wang^{1,2}, Zhefei Zeng^{1,2}, Jinfang Chen^{1,2}, Ticao Zhang³ & Qiong La^{1,2,3}

Phlomoides rotata (Duiyiwei), a threatened Tibetan medicinal plant with anti-inflammatory and analgesic properties, has an understudied genome. We generated a chromosome-level assembly of its 4.12 Gb genome using Oxford Nanopore, PacBio HiFi, Illumina, and Hi-C sequencing, anchoring 22 pseudochromosomes (contig N50 = 173.16 Mb, scaffold N50 = 191.90 Mb). Annotation identified 70,881 protein-coding genes (99.3% complete via BUSCO). Comparative genomic analyses provide insights into the phylogenetic relationships of *P. rotata*. These data provide insights into high-altitude adaptation, bioactive compound biosynthesis, and evolutionary history, supporting conservation and drug development efforts.

Background & Summary

Phlomoides rotata (*P. rotata*), a perennial herbaceous plant within the Lamiaceae family, occupies a highly specialized ecological niche. It is primarily distributed in meadows and along riverbanks at elevations ranging from 3,100 to 5,000 meters, with its range concentrated in the provinces of Gansu, Qinghai, Sichuan, Yunnan, and the Tibet Autonomous Region¹. In traditional Tibetan medicine, *P. rotata* is widely recognized as ‘Daba’ and ‘Dabubba’. However, its restricted wild growth and narrow distribution in high-altitude regions render it particularly vulnerable^{1,2}. The increasing demand for *P. rotata*, driven by its significant pharmacological properties, has resulted in overharvesting. This anthropogenic pressure has critically endangered its survival, qualifying it as a candidate for inclusion in the list of diploid endangered species ($2n = 2X = 22$) on the Qinghai-Tibet Plateau (QTP)³. Understanding the genetic basis of *P. rotata* is essential not only for advancing our knowledge of its unique biological traits but also for developing effective conservation strategies to protect this valuable resource.

In recent years, genomic research within the Lamiaceae family has experienced significant growth. The genomes of several economically important species, widely used in medicine, cuisine, and the fragrance industry, have been sequenced and published. Notable examples include *Salvia officinalis*, *Salvia bowleyana*, *Salvia rosmarinus*, *Lavandula angustifolia*, *Leonurus japonicus* (*L. japonicus*), *Prunella vulgaris*, and *Pogostemon cablin* (*P. cablin*). These studies have provided valuable insights into the genetic mechanisms underlying secondary metabolite production, physiological adaptations, and evolutionary relationships^{4–10}. Despite advances in family-wide genomic studies, *P. rotata* remains relatively underexplored at the genomic level. Current research on this species has predominantly focused on its chemical composition, pharmacological properties, clinical applications, and chloroplast genome^{1,2,11–17}. The lack of a complete genome has impeded a thorough understanding of its genetic architecture, including genes associated with bioactive compound biosynthesis and the genetic mechanisms facilitating adaptation to high-altitude environments. According to the 2015 edition of the Pharmacopoeia of the People’s Republic of China, *P. rotata* demonstrates a broad spectrum of pharmacological activities, including hemostatic, analgesic, anti-inflammatory, anti-tumor, and immune-enhancing effects. The plant’s natural products comprise a rich array of bioactive compounds, including flavonoids, cycloartane glycosides, phenylethanoid glycosides, and volatile oils^{18,19}. Flavonoids, ubiquitous secondary metabolites in the plant kingdom, have been extensively studied for their pharmacological properties. Modern research has

¹Key Laboratory of Biodiversity and Environment on the Qinghai-Tibetan Plateau, Ministry of Education, School of Ecology and Environment, Tibet University, Lhasa, China. ²Yani Wetland Ecosystem Positioning Observation and Research Station, Tibet, China. ³State Key Laboratory of Phytochemistry and Natural Medicines, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China. e-mail: zhangticao@mail.kib.ac.cn; lhagchong@163.com

Statistics	ONT	PacBio	NGS	Hi-C	RNA-seq
Raw data (Gb)	85.00	88.62	0.80	3.00	0.47
N50 (bp)	101,522	16,472	150	150	150
Longest reads (bp)	536,411	58,297	150	150	150
Mean read length (bp)	8,267	16,276	150	150	150

Table 1. Statistics of the sequencing data used for genome assembly of *P. rotata*.

demonstrated their anti-inflammatory, antioxidant, antiviral, antitumor, and antibacterial activities^{20,21}. These effects are attributed to their capacity to modulate cellular signaling pathways, scavenge free radicals, and interact with critical molecular targets. Phenylethanoid glycosides, another significant class of compounds in *P. rotata*, have also demonstrated notable pharmacological potential. Numerous studies have documented their antibacterial, memory-enhancing, immune-boosting, antiviral, antitumor, antioxidant, hepatoprotective, and cardioprotective properties^{21,22}. The complex structures and diverse biological activities of phenylethanoid glycosides render them promising candidates for drug development. The volatile oils in *P. rotata* contain active components, including long-chain fatty acids (LCFAs) and their esters, such as palmitic acid, linoleic acid, and linoleic acid ethyl ester. Extensive research has demonstrated that these volatile oils exhibit anticancer effects through multiple mechanisms, including apoptosis induction, angiogenesis inhibition, and immune response modulation²³.

In this study, we report the chromosome-level genome assembly of *P. rotata*, achieved by integrating Oxford Nanopore Technology (ONT) and PacBio long-read sequencing, Illumina short-read sequencing, and high-throughput chromatin conformation capture (Hi-C) sequencing data. A total of 795.69 Mb of 150 bp paired-end reads, 85.00 Gb of ONT reads, 88.62 Gb of PacBio HiFi reads, 470.53 Mb of RNA-seq reads, and 2.95 Gb of Hi-C reads were generated for *P. rotata*. K-mer frequency analysis estimated the genome size at 4.62 Gb, with a heterozygosity rate of 2.1%. These findings indicate that *P. rotata* exhibits a high level of heterozygosity²⁴. The *P. rotata* genome was assembled from ONT reads using NextDenovo and polished with NGS reads via NextPolish²⁵, yielding a draft assembly of 30 contigs with a contig N50 of 173.20 Mb. Subsequently, Hi-C data were used to scaffold and correct the contigs, enabling the unambiguous anchoring of contigs onto the 22 chromosomes of *P. rotata* and achieving a scaffold N50 value of 192.00 Mb. The final assembled genome is 4.12 Gb, closely matching the estimated genome size. Genome annotation of *P. rotata* predicted 70,881 protein-coding genes (PCGs) using the MAKER2 pipeline²⁶, with 66,592 annotated on chromosomes.

The high-quality genome assembly of *P. rotata* advances Tibetan medicinal plant research by providing a genetic foundation for decoding secondary metabolite biosynthesis (flavonoids and phenylethanoid glycosides), enabling metabolic engineering and synthetic biology applications to enhance pharmaceutical compound production. Comparative genomic analyses within Lamiaceae reveal evolutionary processes, gene family dynamics, and genomic rearrangements driving family diversification. The genome also sheds light on genetic mechanisms of high-altitude adaptation to extreme environments (low oxygen, cold, and UV radiation), offering insights for plant evolution research and conservation strategies of endangered alpine species.

Methods

Plant material collection, DNA and RNA extraction, and sequencing. Fresh young leaves were collected from a mature *P. rotata* plant in Mozhugongka County (N 29.705°, E 92.083°, altitude 4,135 m) on the QTP, China. The samples were transported to Wuhan Benagen Technology Company Limited for genome sequencing. Genomic DNA was extracted using the Qiagen DNeasy Plant Mini Kit following the manufacturer's protocol. For short-read sequencing, 150 bp short-insert libraries were prepared and sequenced in 300 bp paired-end mode using the Illumina HiSeq X Ten platform. The raw sequencing data were pre-processed by trimming adapters and filtering out low-quality reads with FASTP²⁷. For long-read sequencing, a 20 kb insert Circular Consensus Sequence (CCS) library was prepared and sequenced on the PacBio Revio platform to generate High-Fidelity (HiFi) reads. Additionally, ONT long-read sequencing was performed. The Hi-C library was constructed in accordance with an established protocol. Specifically, 2.0 g of young leaves were cross-linked using a 1.0% formaldehyde solution. Chromatin was then extracted, digested with MboI (New England Biolabs), and the resulting DNA ends were labeled, biotinylated, diluted, and randomly ligated. Total RNA was extracted from the leaves, stems, flowers, and roots of each species using the Qiagen RNeasy Plant Mini Kit and pooled to create a mixed sample. RNA-sequencing libraries were constructed using the Illumina TruSeq RNA Library Preparation Kit, followed by paired-end sequencing with a read length of 150 bp on the HiSeq X Ten platform. mRNA was enriched using anti-polyA magnetic beads, and paired-end cDNA libraries were subsequently constructed. The mRNA was fragmented, circularized, and then used to prepare libraries, which were sequenced in PE150 mode on the Illumina HiSeq X Ten platform.

K-mer based genome size estimation. The genome size of *P. rotata* was estimated through k-mer analysis using GCE v1.0²⁴. The analysis, performed with findGSE software²⁸, suggested a genome size ranging from 4.62 to 5.40 Gb (Table S1, Fig. S1). To determine the ploidy level of the sequenced material, GenomeScope 2.0 and Smudgeplot v0.2.5 were employed based on 19-mer analysis²⁹. The results revealed that the AB ratio (0.28) and AAB ratio (0.26) were predominant, confirming that *P. rotata* is diploid (Fig. S2).

Genome assembly and quality control. The long-read sequences generated by the PacBio sequencing platform were assembled using *hifiasm* v0.16.1-r375³⁰. This process involved assembling PacBio HiFi reads

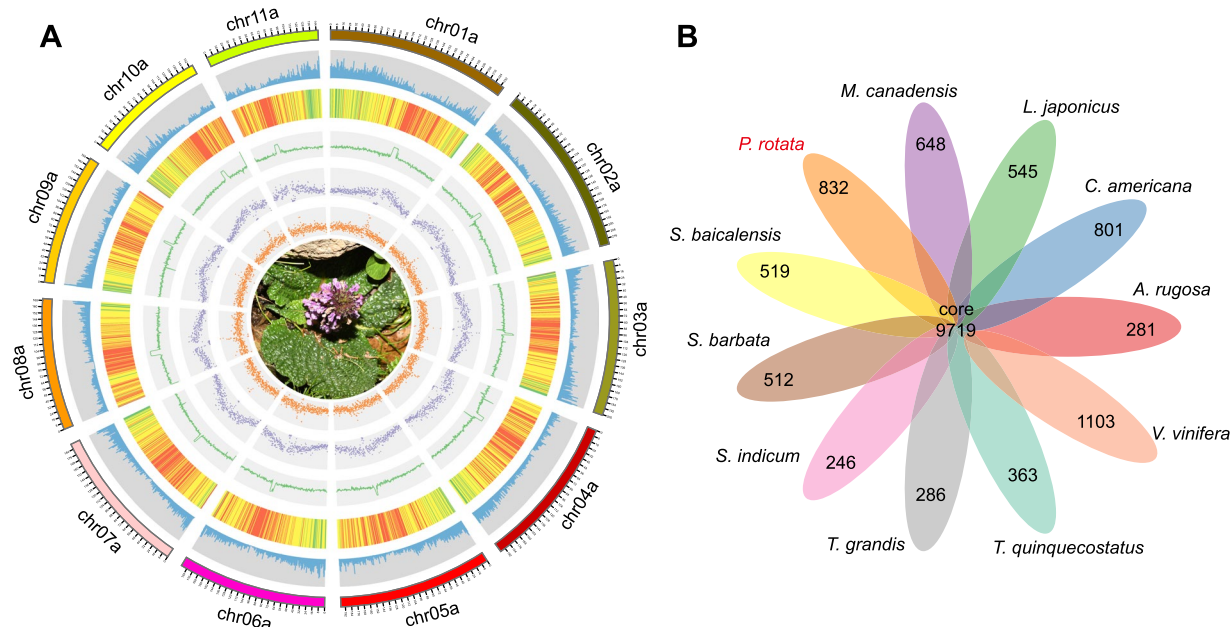


Fig. 1 Phylogenomics and genome characterizations of *P. rotata*. **(A)** Genomic characteristics of *P. rotata*. Tracks show, from outside to inside, the densities of genes (blue), repeat content (green-yellow-red), GC content (green), *Copia* content (purple), *Gypsy* content (orange), and the links between each chromosome, respectively. **(B)** Flower diagram showing the shared and unique gene families among 11 species. Flower plot displaying the shared core orthogroups (in the center) and the species-specific orthogroups (in the petals) for the relatives.

into contigs. The resulting haplotype assemblies were then used for subsequent in-depth analysis. Hi-C reads were aligned to the draft genome using Juicer³¹, followed by scaffolding with the 3D-DNA pipeline³² and manual refinement in Juicebox³³. Gaps in the assembly were filled using the quarTeT software³⁴ with PacBio HiFi reads, applying a fixed gap length of 100 bp. The telomere-specific sequence (TTTAGGG)n was successfully assembled at the ends of most chromosomes. To resolve incomplete or missing telomere sequences, PacBio HiFi reads were mapped back to the chromosomes, with a focus on reads near the telomeric regions, and subsequently assembled into contigs using *hifiasm*³⁰. The contigs were aligned to the chromosomes, and the chromosomal sequences were extended outward to assemble the telomere regions as comprehensively as possible. The chloroplast and mitochondrial genomes were assembled using GetOrganelle³⁵. Ultimately, a high-quality genome assembly was achieved. The assembly quality was evaluated using BUSCO v2.0.1 (Benchmarking Universal Single-Copy Orthologs)³⁶ with the embryophyta_odb10 database ($n = 1,614$). To further evaluate the quality of the genome assembly, Illumina short-reads were mapped to the assembled genome sequence using BWA v0.7.17²⁴ with default parameters. Additionally, PacBio HiFi reads were aligned to the genome using minimap2²⁷. Non-primary alignments were excluded, and the mapping ratio and coverage percentage were calculated for subsequent quality assessment.

Genome Annotation

Repetitive elements identification. Repetitive elements were identified using a combination of homology-based and de novo approaches. For homology-based annotation, RepeatMasker (available at <http://www.repeatmasker.org/>) was employed to align the genome sequences against the Repbase database, enabling the detection and classification of diverse repetitive elements. For de novo-based prediction, the Extensive de novo TE Annotator (EDTA) pipeline³⁸ was applied with the parameters (--sensitive 1 --anno 1). Initially, a TE library was constructed using EDTA's default settings. Subsequently, RepeatMasker was employed to identify repetitive elements within the generated library.

Gene structure annotation. Gene structure annotation was performed using a combination of three approaches: de novo gene prediction, homology-based prediction, and RNA-seq-based gene prediction. For de novo transcript assembly, transcripts were reconstructed using Trinity³⁹ based on transcriptome reads. The reads were then aligned to the genome using HISAT2⁴⁰, and the transcripts were subsequently assembled with StringTie⁴¹. For homology-based prediction, gene annotation was performed using 438,926 non-redundant protein sequences from multiple plant species (*P. cablin*, *Salvia miltiorrhiza*) as homologous protein evidence. Using transcript evidence, gene structures were annotated through the PASA pipeline⁴². Subsequently, full-length genes were identified by aligning them with reference proteins. The full-length gene set was then used to train and optimize AUGUSTUS⁴³ over five iterations, and SNAP⁴⁴ was also trained for gene prediction. Annotation was performed using the MAKER pipeline²⁶. Subsequently, EVidenceModeler (EVM) was employed to integrate annotations from MAKER²⁶ and PASA⁴², producing a consolidated set of gene annotations. To exclude transposable element (TE) coding regions, TESorter⁴⁵ was used to identify TE protein domains within the genome.

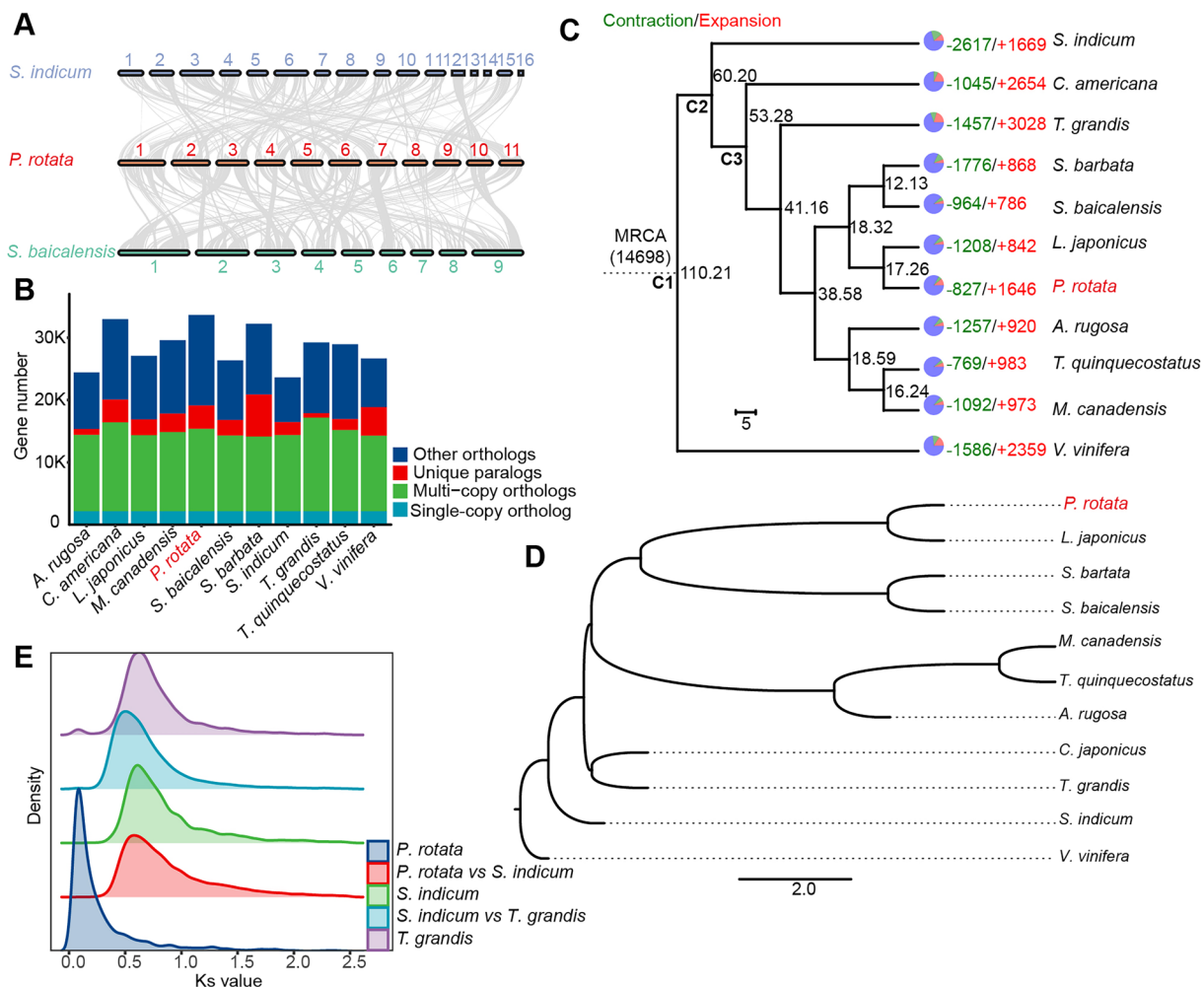


Fig. 2 *P. rotata* genome evolution. (A) Genomic collinearity among *P. rotata* and its other two closely related species. (B) Gene distribution in *P. rotata* and 10 other representative species. (C) Phylogenetic tree based on 11 species identified by Orthofinder2 to show divergence times. (D) Astral coalescence tree. (E) The KaKs analysis of *P. rotata*. Pie charts show the percentage of gene families that underwent expansion or contraction. The numbers above the node indicate the range of species divergence times. C1-C3 mark calibration points used to estimate the divergence times.

Subsequently, EVM was used to mask the identified TE domains. Finally, the EVM annotations were refined using PASA⁴² to incorporate untranslated regions (UTRs) and alternative splicing events. Gene annotations that were abnormal (containing internal stop codons, ambiguous bases, or lacking start/stop codons) or excessively short (<50 amino acids) were filtered out.

Gene function prediction. Gene function prediction was performed using three approaches. (1) EggNOG-Mapper⁴⁶ was employed to annotate gene functions, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms, by aligning sequences against the eggNOG homologous gene database. (2) For sequence similarity analysis, DIAMOND⁴⁷ was used to align protein sequences against databases such as SwissProt, TrEMBL, NR, and Arabidopsis. The goal was to identify the most significant matches, with alignment criteria set at a minimum identity of 30% and an e-value threshold of less than 1e-5. (3) Domain similarity search: InterProScan⁴⁸ was used to query multiple sub-databases within InterPro, including PRINTS, Pfam, SMART, PANTHER, and CDD, to identify conserved protein sequences, motifs, and functional domains.

Non-coding RNA (ncRNA) structures annotation. tRNA annotation was performed using tRNAscan-SE v1.23⁴⁹, while rRNA annotation was conducted with barrnap (<https://github.com/tseemann/barrnap>), with partial results excluded. Non-coding RNAs were annotated by aligning sequences against the Rfam database using RfamScan.

Comparative genomic analyses. Gene family clustering was conducted for 11 plant species, including *Agastache rugosa* (*A. rugosa*), *L. japonicus*, *Mentha suaveolens* (*M. suaveolens*), *P. rotata*, *P. cablin*, *Rehmannia glutinosa* (*R. glutinosa*), *Salvia miltiorrhiza* (*S. miltiorrhiza*), *Scutellaria baicalensis* (*S. baicalensis*), *Scutellaria barbata*

Features	Values
Predicted genome size (Gb)	4.62–5.40
Assembled genome size (Gb)	4.12
Number of contigs	30
N50 of contigs (Mb)	173.16
Number of scaffolds	24
N50 of scaffolds (Mb)	191.90
GC content (%)	38.41
Total length and percent of repeats (Gb)	2.71 (65.71%)
LTRs (Gb)	2.13 (51.72%)
<i>Copia</i> (Mb)	789.33 (19.17%)
<i>Gypsy</i> (Mb)	892.44 (21.67%)
LINEs (Mb)	0.76 (0.02%)
TIRs (Mb)	322.57 (7.83%)
Number of annotated genes	70,881
Mapping rate	92.60–99.70%
BUSCO of assembly	1,601 (99.20%)
BUSCO of annotation	1,603 (99.30%)

Table 2. Statistics of genome assembly and annotation of *P. rotata*. Note: LTRs, Long Terminal Repeats; LINEs, Long Interspersed Nuclear Elements; TIRs, Terminal Inverted Repeats.

(*S. barbata*), *Thymus quinquecostatus* (*T. quinquecostatus*), and *Tectona grandis* (*T. grandis*), using DIAMOND⁴⁷ and OrthoFinder2⁵⁰ (parameter: -M msa) (Table S2). Syntenic blocks among the genomes of *P. rotata*, *S. indicum*, and *S. baicalensis* were identified using MCScanX⁵¹. For the OrthoFinder analysis, the longest predicted protein sequence from each gene was selected as the representative input. Subsequently, Trimal v1.4.1⁵² was used to trim poorly aligned regions from the multiple protein sequence alignments. Astral-Pro v1⁵³ was employed to infer the species tree using 20,071 gene trees derived from orthologous genes present in at least 60.0% of the species, focusing on single-copy genes. We assumed that the four codon sites exhibited distinct substitution rates. For the molecular clock analysis, an independent rates model was applied (clock parameter = 2), and the Generalized Time-Reversible (GTR) model was used for nucleotide substitution. The first 500,000 iterations of the Markov Chain Monte Carlo (MCMC) chain were designated as the burn-in phase. After burn-in, sampling was performed every 100 iterations, totaling 100,000 sampling events. Three calibration points (C1: *Sesamum indicum* vs. *Vitis vinifera*: 111.4–123.9 million years ago (Mya), C2: *Sesamum indicum* vs. *Tectona grandis*: 32.6–63.0 Mya, C3: *Tectona grandis* vs. *Scutellaria baicalensis*: 19.4–39.9 Mya)^{54–58} derived from the TimeTree database (<http://www.timetree.org/>) were applied to constrain the divergence times of the nodes. Divergence times were estimated using the MCMCTree program within PAML v4.9⁵⁹. Gene families exhibiting expansion or contraction events were identified using CAFE5⁶⁰. Subsequently, expanded gene families in *P. rotata* were analyzed for enrichment in GO terms and KEGG pathways. For both GO term and KEGG pathway enrichment analyses, a significance threshold of $p < 0.05$ was applied. Coding sequences (CDS) and protein sequences were aligned with ParaAT v2.0⁶¹ (parameters: -m muscle -f axt). The Nonsynonymous Substitution Rate/Synonymous Substitution Rate (Ka/Ks) ratio was calculated using the 'YN' model in KaKs_Calculator v2.0⁶². Ks and Fourfold Degenerate Transitions (4DTv) values were analyzed to detect potential whole-genome duplication events and to assess divergence between orthologs and paralogs.

Data Records

Raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1284616, with accession numbers SRS25198821–SRS2519883464⁶³. Genome assemblies have been deposited in the NCBI under BioProjects PRJNA1266861 and PRJNA1266862, with accession numbers JBQBLY0000000000⁶⁴ and JBQBLZ0000000000⁶⁵. Genome annotation files are publicly available in the Figshare repository (<https://doi.org/10.6084/m9.figshare.29069453.v1>)⁶⁶.

Technical Validation

Genome assembly validation. *Genome size and ploidy confirmation.* The estimated genome sizes, ranging from 4.62 Gb to 5.40 Gb, were consistent with the final assembled genome size of 4.12 Gb. To further validate the ploidy level, Smudgeplot analysis of k-mer spectra was conducted. The results revealed that 26.0% and 28.0% of k-mer pairs exhibited AB and AAB ratios, respectively, confirming that *P. rotata* is a diploid species with a chromosome number of $2n = 2X = 22$ ⁶⁷.

Sequencing depth and assembly integrity. A substantial amount of sequencing data was generated for genome assembly, including approximately 85.0 Gb of ONT data ($\sim 40 \times$ coverage), 88.6 Gb of PacBio HiFi data ($\sim 42 \times$ coverage), 119.4 Gb of short-read data ($\sim 57 \times$ coverage), and 443.2 Gb of Hi-C read data (Table 1). The PacBio and ONT subreads were independently assembled into contigs, yielding a total of 30 contigs. Following correction and scaffolding with Hi-C data, all corrected reads were successfully assembled into 22 pseudochromosomes. The

contig N50 and scaffold N50 values were 173.2 Mb and 191.9 Mb, respectively, with only six gaps remaining in the assembly, demonstrating a high level of assembly completeness and quality (Table 2).

Completeness and mapping evaluation. To assess the completeness of the genome assembly, we employed BUSCO v2.0.1³⁶ with the embryophyta_odb10 dataset. The analysis revealed a remarkably high level of completeness, with 99.2% of BUSCO genes showing complete matches. For a more comprehensive evaluation of assembly quality, Illumina, ONT, and PacBio reads were mapped to the assembled genome. The mapping rates ranged from 92.6% to 99.7%, indicating that the assembled genome accurately represents the original sequencing data. These results underscore the high quality and reliability of the assembly, making it suitable for further research and analysis (Table 2).

Genome annotation validation. Gene model accuracy. Genome annotation was performed using a combination of transcript-based, de novo, and homology-based prediction methods. A total of 70,881 gene models were predicted, of which 66,592 (93.9%) were functionally annotated by aligning with public protein databases (Table 2). These results indicate a high level of accuracy in gene prediction.

Repetitive elements characterization. A thorough analysis of repetitive elements in the *P. rotata* genome identified 5,151,597 repetitive elements, spanning a total length of 2.70 Gb and representing 65.71% of the genome. Long terminal repeat retrotransposons (LTRs) were the most abundant repetitive elements, comprising 51.7% of the genome. Among these, Ty1/Copia and Ty3/Gypsy elements represented 19.2% and 21.7% of the genome, respectively. Furthermore, terminal inverted repeat retrotransposons (TIRs) accounted for 7.83% of the repetitive elements (Table 2).

Non-coding RNA annotation accuracy. For non-coding RNA (ncRNA) annotation, tRNA was identified using tRNAscan-SE v1.23⁴⁹, rRNA was annotated with barrnap, and other ncRNAs were annotated using RfamScan. A total of 5,294 ncRNA genes were identified, comprising 24 rRNAs, 2,359 tRNAs, and 2,911 other ncRNAs (Table 2).

Annotated genes completeness. The completeness of the annotated protein-coding genes was assessed using BUSCO analysis with the embryophyta_odb10 dataset. The results revealed a 99.3% complete match rate, further validating the high quality of the genome annotation (Table 2). The genome Circos plot is depicted in Fig. 1A.

Syntenic analyses. MCScanX⁵¹ analysis reveals robust syntenic relationships among the three species. Notably, a high degree of synteny is identified between sesame chromosome 12, *S. baicalensis* chromosome 9, and *P. rotata* chromosome 10, all of which exhibit chromosomal inversions. Furthermore, significant synteny was observed between *P. rotata* chromosome 3 and *S. baicalensis* chromosomes 1 and 3, potentially resulting from chromosomal structural variations or gene duplication events (Fig. 2A).

Phylogenomic analyses. Phylogenomic analyses were conducted by comparing the *P. rotata* genome with 10 published plant species (using *V. vinifera* as an outgroup) via OrthoFinder v2.5.4⁵⁰ with default settings (parameter: -M msa). A total of 26,418 orthogroups were identified, including 832 orthogroups (comprising 3,704 unique genes) that were exclusive to *P. rotata* (Fig. 1B, Fig. 2B, Table S3). Using 2,150 single-copy orthogroups from 11 species, phylogenetic analysis showed that *P. rotata* forms a clade with *L. japonicus*, sister to *Scutellaria* (Fig. 2C). Based on 20,071 gene trees, ASTRAL analysis demonstrated that *P. rotata* and *L. japonicus* form a distinct clade, with the genus *Scutellaria* as an outer sister group. Both the phylogenetic tree and ASTRAL results displayed congruent topological structures, further supporting a closer evolutionary relationship between *P. rotata* and the genera *Leonurus* and *Scutellaria* (Figs. 2C and 2D). KEGG functional enrichment analysis of the expanded gene families in *P. rotata* revealed significant enrichment in the pathways of sesquiterpenoid and triterpenoid biosynthesis, as well as oxidative phosphorylation (Fig. S3A). In contrast, the contracted gene families were predominantly enriched in the phenylpropanoid biosynthesis pathway (Fig. S3B). Phylogenetic analysis revealed that *P. rotata* shares a closer genetic relationship with *L. japonicus*. As illustrated in Fig. 2C, their divergence time is estimated to be approximately 17.26 Mya. Furthermore, genomic analyses indicated *P. rotata* underwent core angiosperm hexaploidization and a Lamiaceae-specific tetraploidization event (Fig. 2E, Fig. S4).

Code availability

No custom code was used for this study. All data analyses were conducted using published bioinformatics software with default settings unless otherwise specified.

Received: 20 November 2024; Accepted: 12 August 2025;

Published online: 29 September 2025

References

1. Cui, Z. H. *et al.* Traditional uses, phytochemistry, pharmacology and toxicology of *Lamiophlomis rotata* (Benth.) Kudo: a review. *RSC Advances* **10**, 11463–11474 (2020).
2. Li, Y. *et al.* *Lamiophlomis herba*: A comprehensive overview of its chemical constituents, pharmacology, clinical applications, and quality control. *Biomedicine & Pharmacotherapy* **144**, 112299 (2021).
3. Li, L., Wang, X. Z., Fang, S. X. & Hou, F. H. Reservoir characteristics and controlling factors of the Upper Paleozoic sandstone, Eastern Ordos basin. *Xinan Shiyou Xueyuan. Xuebao/Journal of Southwestern Petroleum Institute* **24**, 4 (2002).

4. Shen, Y. *et al.* Chromosome-level and haplotype-resolved genome provides insight into the tetraploid hybrid origin of *patchouli*. *Nature Communications* **13**, 3511 (2022).
5. Zhang, S. *et al.* Chromosome-level genome assembly of *Prunella vulgaris* L. provides insights into pentacyclic triterpenoid biosynthesis. *The Plant Journal* **118**, 731–752 (2024).
6. Han, D. *et al.* The chromosome-scale assembly of the *Salvia rosmarinus* genome provides insight into carnolic acid biosynthesis. *The Plant Journal* **113**, 819–832 (2023).
7. Hamilton, J. P. *et al.* Chromosome-scale genome assembly of the ‘Munstead’ cultivar of *Lavandula angustifolia*. *BMC Genomic Data* **24**, 75 (2023).
8. Wang, X. *et al.* De novo chromosome-level genome assembly of Chinese motherwort (*Leonurus japonicus*). *Scientific Data* **11** (2024).
9. Li, C. Y. *et al.* The sage genome provides insight into the evolutionary dynamics of diterpene biosynthesis gene cluster in plants. *Cell Reports* **40** (2022).
10. Zheng, X. *et al.* Insights into salvianolic acid B biosynthesis from chromosome-scale assembly of the *Salvia bowleyana* genome. *Journal of Integrative Plant Biology* **63**, 1309–1323 (2021).
11. Wang, J. Mao, X. X. & Ma, Y. Complete chloroplast genome of *Lamiophlomis rotata*: comparative genome analysis and phylogenetic analysis. *Mitochondrial DNA Part A* **33**, 29–39 (2022).
12. Pema, Y. Ma, C., Bautista, M. A. C. & Chen, T. The complete chloroplast genome sequence of the Tibetan herb *Phlomis rotata* (Benth. ex Hook.f.) Mathiesen (Lamiaceae). *Mitochondrial DNA Part B* **6**, 3261–3262 (2021).
13. La, M. *et al.* Constituent analysis and quality control of *Lamiophlomis rotata* by LC-TOF/MS and HPLC-UV. *Journal of Pharmaceutical and Biomedical Analysis* **102**, 366–376 (2015).
14. Li, T. *et al.* Comparative investigation of aerial part and root in *Lamiophlomis rotata* using UPLC-Q-Orbitrap-MS coupled with chemometrics. *Arabian Journal of Chemistry* **15**, 103740 (2022).
15. Jiang, Y. *et al.* Network Pharmacology-Based Prediction of Active Ingredients and Mechanisms of *Lamiophlomis rotata* (Benth.) Kudo Against Rheumatoid Arthritis. *Frontiers in Pharmacology* **10** (2019).
16. Zhan, H. *et al.* Exploring the pharmacological mechanisms and key active ingredients of total flavonoids from *Lamiophlomis rotata* (Benth.) Kudo against rheumatoid arthritis based on multi-technology integrated network pharmacology. *Journal of Ethnopharmacology* **317**, 116850 (2023).
17. Zhu, B. *et al.* *Lamiophlomis rotata*, an Orally Available Tibetan Herbal Painkiller, Specifically Reduces Pain Hypersensitivity States through the Activation of Spinal Glucagon-like Peptide-1 Receptors. *Anesthesiology* **121**, 835–851 (2014).
18. Qiao, F. *et al.* Flavonoid synthesis in *Lamiophlomis rotata* from Qinghai-Tibet Plateau is influenced by soil properties, microbial community, and gene expression. *Journal of Plant Physiology* **287**, 154043 (2023).
19. Wan, G. *et al.* The total polyphenolic glycoside extract of *Lamiophlomis rotata* ameliorates hepatic fibrosis through apoptosis by TGF- β /Smad signaling pathway. *Chinese Medicine* **18**, 20 (2023).
20. Kim, H. P., Park, H., Son, K. H., Chang, H. W. & Kang, S. S. Biochemical pharmacology of biflavonoids: Implications for anti-inflammatory action. *Archives of Pharmacol Research* **31**, 265–273 (2008).
21. Song, B. *et al.* Gossypin: A flavonoid with diverse pharmacological effects. *Chemical Biology & Drug Design* **101**, 131–137 (2023).
22. Wu, L. *et al.* Therapeutic potential of phenylethanoid glycosides: A systematic review. *Medicinal Research Reviews* **40**, 2605–2649 (2020).
23. Roopashree, P. G., Shetty, S. S. & Suchetha Kumari, N. Effect of medium chain fatty acid in human health and disease. *Journal of Functional Foods* **87**, 104724 (2021).
24. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303**, 3997 (2013).
25. Hu, J., Fan, J., Sun, Z., Liu, S. & Berger, B. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
26. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**, 188–196 (2008).
27. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
28. 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 (2017).
29. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**, 1432 (2020).
30. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
31. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
32. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *science* **356**, 92–95 (2017).
33. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101 (2016).
34. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10** (2023).
35. Yu, J. J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 241 (2020).
36. 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).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
38. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275 (2019).
39. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652 (2011).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360 (2015).
41. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**, 290–295 (2015).
42. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
43. 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).
44. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 59 (2004).
45. Zhang, R. G. *et al.* TESorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9** (2022).
46. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution* **34**, 2115–2122 (2017).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).

48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
49. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**, 955–964 (1997).
50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238 (2019).
51. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49–e49 (2012).
52. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
53. Zhang, C., Scornavacca, C., Molloy, E. K. & Mirarab, S. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Molecular Biology and Evolution* **37**, 3292–3307 (2020).
54. Harris, L. W. & Davies, T. J. A Complete Fossil-Calibrated Phylogeny of Seed Plant Families as a Tool for Comparative Analyses: Testing the ‘Time for Speciation’ Hypothesis. *PLOS ONE* **11**, e0162907 (2016).
55. Naumann, J. *et al.* Single-Copy Nuclear Genes Place Haustorial Hydnoraceae within Piperales and Reveal a Cretaceous Origin of Multiple Parasitic Angiosperm Lineages. *PLOS ONE* **8**, e79204 (2013).
56. Verboom, G. A., Stock, W. D. & Cramer, M. D. Specialization to Extremely Low-Nutrient Soils Limits the Nutritional Adaptability of Plant Lineages. *The American Naturalist* **189**, 684–699 (2017).
57. Roalson, E. H. & Roberts, W. R. Distinct Processes Drive Diversification in Different Clades of Gesneriaceae. *Systematic Biology* **65**, 662–684 (2016).
58. Vargas, P. *et al.* Testing the biogeographical congruence of palaeofloras using molecular phylogenetics: snapdragons and the Madrean–Tethyan flora. *Journal of Biogeography* **41**, 932–943 (2014).
59. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
60. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
61. Zhang, Z. *et al.* ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* **419**, 779–781 (2012).
62. Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics & bioinformatics* **8**(1), 77–80 (2010).
63. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP596310> (2025).
64. Tian, Z. *Phlomis rotata* cultivar S1, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBQBL000000000> (2025).
65. Tian, Z. *Phlomis rotata* cultivar S1, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBQBL000000000> (2025).
66. Tian, Z. Chromosome-level genome assembly of *Phlomis rotata*. *figshare* <https://doi.org/10.6084/m9.figshare.29069453.v1> (2025).
67. Wang, J. *et al.* *Lamiochloa rotata* identification via ITS2 barcode and quality evaluation by UPLC-QTOF-MS couple with multivariate analyses. *Molecules* **23**, 3289 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31760127 to L.Q.), a first-class discipline construction project in ecology (00060906-01, 00060835/001 to L.Q.), the Science and Technology Program of Xizang Autonomous Region (XZ202402ZY0023, XZ202402JX0003 to L.Q.), High-level graduate research project, Xizang University (2021-GSP-B019 to Z.T.).

Author contributions

T.Z., and Q.L. conceived and designed the study; Z.T., J.W., Z.Z., and J.C. prepared the materials; Z.T., T.Z., and J.W. conducted the experiments, analyzed data, and prepared the results; Z.T., T.Z., and Q.L. wrote and improved the manuscript. All authors approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-025-05800-4>.

Correspondence and requests for materials should be addressed to T.Z. or Q.L.

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025