BMC Genomics

RESEARCH Open Access

Multi-omics analyses of *Dracocephalum moldavica* L. reveal two flavonoid glycosyltransferases in tilianin biosynthesis

Qian Zhao[1,2†], Xiaoxue Zhang[1†], Wenfang Wang[1], Hui Yu[1,3] and Zhanli Wang[1*]

Abstract

Background To explore reference genome of *Dracocephalum moldavica* L. and provide a basis for the study of the mechanism of tilianin biosynthesis.

Results We present telomere-to-telomere (T2T) genome assembly with the size of 255.61 Mb and a contig N50 length of 58.45 Mb. By integrating multi-omics resources, we identified two UDP- glycosyltransferases, Dm7GlcT1 and Dm7GlcT2, catalyzing the glucosylation of acacetin at 7-OH, which was the final step in tilianin biosynthetic pathway. The findings from the *in vitro* enzyme activity assay and qRT-PCR analysis further indicate that Dm7GlcT1 may be responsible for catalyzing tilianin biosynthesis in stems and flowers, whereas glucosylation mediated by Dm7GlcT2 could occur in all tissues.

Conclusion Our work presents the first T2T genome assembly of *D. moldavica* and identified two key UDP-glycosyltransferases associated with tilianin biosynthesis, which provided a valuable reference for further metabolic network construction of tilianin biosynthesis and genetic improvement of *D. moldavica* with high tilianin product.

Keywords *Dracocephalum moldavica* L., Genome, T2T genome assembly, UDP-glycosyltransferase, Tilianin

Introduction

Dracocephalum moldavica L. is an annual plant of Dracocephalum in Lamiaceae, also called Xiangqinglan in chinese, commonly used as traditional folk medicine for the Mongolian and Uighur [1, 2]. In traditional Chinese medicine, *D. Moldavica* has the effects of purging fire, clearing heat and hemostasis [3]. The effective part of *D. Moldavica* could be used to treat hypertension and hyperlipidemia, prevent atherosclerosis and protect against myocardial ischemia–reperfusion injury, and it also owns broad application prospects in the prevention and treatment of cardiovascular and cerebrovascular diseases [2, 4]. Previously, because of its great medicinal value, the research on *D. Moldavica* predominantly focused on the isolation and identification of compounds as well as the pharmacodynamic studies of total extracts. In recent years, flavonoids as the main bioactive compounds in *D. Moldavica*, have become the focus of research due to its extensive pharmacological effects such as inhibiting oxidation damage and inflammatory reactions in cerebral ischemia reperfusion injury [5], protecting astrocytes against H_2O_2-induced apoptosis [6] and reducing inflammation to improve pulmonary fibrosis

†Qian Zhao and Xiaoxue Zhang contributed equally to this work.

*Correspondence:
Zhanli Wang
wang.zhanli@hotmail.com
[1]Inner Mongolia Key Laboratory of Disease-Related Biomarkers, The Second Affiliated Hospital, Baotou Medical College, Baotou, China
[2]Translational Medicine Center, Baotou Medical College, Baotou 014040, China
[3]School of Basic Medicine, Baotou Medical College, Baotou, China

[7]. With the continuous deepening of pharmacological research, comprehensive and rapid identification of its chemical composition and clarification of the key material for its efficacy have become urgent problems to be solved [8].

The role of tilianin is not only a main active component in total flavonoids of *D. Moldavica*, but also one of the key quality control standards for *D. moldavica* [8]. The functions of tilianin have been studied extensively, including anti-inflammatory [9–11], anti-oxidation [12, 13], diabetes prevention [14, 15], heart protection [16–18] and nerve protection [19, 20]. Recent research paid more attention to the mechanism of the medicinal effect of tilianin. For example, tilianin pretreatment prevents myocardial ischemia–reperfusion injury via preservation of mitochondrial function [21], tilianin post-conditioning attenuates myocardial ischemia/reperfusion injury via mitochondrial protection and inhibition of apoptosis [22–24]. However, tilianin is relatively rare because of the low contents in most plants and no mature method of chemical synthesis [25]. Meanwhile, the pathways of tilianin biosynthesis also have few reports, which severely restricted the wide application of tilianin in medical field.

Glycosylation is a prominent modification reaction in plant secondary metabolites which is often the last step in the biosynthesis of natural compounds [26]. UDP-glycosyltransferases (UGTs), as the largest GT families in the Carbohydrate-Active enZymes database (CAZy) [27], usually catalyze the transfer of UDP-glucose to specific acceptor molecules leading glycoside formation [26]. M. Feng et al. had verified the glucosylation of acacetin at 7-OH to form tilianin in *Nardostachys jatamansi* [25]. Mostly, they identified a flavonoid glycosyltransferase in the biosynthesis of tilianin and linarin via AI-assisted protein structure clustering [25], which provided us critical clues to identify the key enzymes in *D. moldavica*.

Whole-genome sequencing techniques have significant advances in breeding, identification, phylgenetic evolution and disease management for traditional Chinese herbs. Presently, there is no reference genome in telomere-to-telomere (T2T) level in *Dracocephalum*, which restricts the functional research of *D. moldavica* to a certain extent. *D. moldavica* is a main source of tilianin, the lack of genome prevented us to carry out the systematic identification of the key enzymes in biosynthesis of tilianin, hindered the progress in construction of biosynthetic network of secondary metabolites in *D. moldavica*. Thus, a high-quality genome of *D. moldavica* is urgent for the exploration of critical genes during biosynthesis process of secondary metabolites.

Here, we report a first T2T genome of *D. moldavica* with high quality by integrating the data of PacBio HiFi, ONT ultra-long and Hi-C. The size of final genome is approximately 255.61 Mb with a contig N50 length of 58.45 Mb, and 94.28% of the assembled sequences were anchored into 4 pseudochromosomes. Based on sequence similarity and expression pattern, two putative UGTs, Dm7GlcT1 and Dm7GlcT2, were chose to verify the catalytic capability for glucosylation via enzyme activity assay *in vitro*. Both genes encoding 7-O-glucosyltransferases could catalyze acacetin to form tilianin. Our studies firstly identified key genes in the key process of tilianin biosynthesis in *D. moldavica*, which laid a foundation for yield increase of tilianin in *D. moldavica* and provided new clues in functional gene identification for tilianin biosynthesis of plant-based medicines.

Materials and methods

Plant materials, DNA and RNA extraction

The young leaves were collected from a single *D. moldavica* plant cultivated in the medicinal plants' conservation nursery of BaoTou Medical College to extract cell nuclei and genomic DNA for genome survey and sequencing. Total RNA was extracted from roots, stems, young leaves, mature leaves and flowers from a mature *D. moldavica* plant for sequencing, then stems, leaves as well as flowers were also gathered for the analysis of key metabolites via metabonomics. Three biological replicates were performed in three single plants for collection. The genomic DNA and total RNA of *D. moldavica* was extracted using the CTAB method [28] and Trizol-based methods from young leaves and 5 tissues of a plant, respectively. The concentration of DNA and RNA was detected using Qubit fluorometer, NanoDrop and agarose gel electrophoresis, which were also provided evidence for quality control.

Library construction and sequencing

For short-read sequencing, the shotgun library was prepared from qualified genomic DNA with 300 ~ 400 bp inserted fragment following the standard protocol, which was then performed the paired-end sequencing using DNBseq sequencing platform. Finally, 365.85 Gb of sequencing data was obtained for the following analysis.

For PacBio HiFi sequencing, the genomic DNA with higher quality was used to construct the SMRT bell library whose insert size was approximately 20 kb (Pacific Biosciences), then the circular consensus sequencing (CCS) model (Pacific Biosciences) was used for the library sequencing on the PacBio Revio platform. In total, 44.47 Gb of HiFi reads were generated for the de novo assembly.

For ONT ultra-long sequencing, the Ligation Sequencing Kit (Nanopore, SQK-LSK110) with the standard protocol was used for the library construction. The purified library was sequenced on the PromethION Nanopore sequencer and approximately 87.78 Gb of filtered long-reads data was generated.

For full-length RNA sequencing, the high-quality RNA (RIN > 9) was reverse transcribed to cDNA via UMI base PCR cDNA Synthesis Kit (BGI). Then the amplification of double-stranded cDNA was used to construct the single-molecule long-read isoform sequencing (Iso-Seq) library, which was further sequenced via PacBio Sequel platform with 42.20 Gb of data obtained. For transcriptome sequencing, mRNA was firstly enriched from the total RNA for fragmentation and reverse transcription to obtain double-stranded cDNA. After terminal repair and PCR, the cDNA libraries were constructed based on 5 tissues and then used for sequencing via DNBseq platform.

For Hi-C sequencing, the restriction enzyme Mbo I was firstly used to digest the crosslinked DNA which was fixed by formaldehyde. Then the ends of digested DNA were marked with biotin and ligated by DNA ligase. After interrupted to 300 ~ 500 bp fragments using ultrasound, the crosslinked DNA was enriched to form the Hi-C libraries. Finally, the libraries were sequenced on DNBseq sequencing platform and a total of 65.95 Gb of raw data was obtained.

Genome survey

The genome size of *D. moldavica* was firstly estimate by flow cytometry. The young leaves were firstly harvested from mature plants and cleaned with sterile water. After chopped up in dissociation solution [Galbraith's isolation buffer: 45 mM $MgCl_2$, 30 mM Sodium citrate, 20 mM MOPS, 0.1% (v/v) TritonX-100], the cell nucleuses were gradually released during incubated on ice for 30 min. Then the homogenate containing nucleuses was filtered through nylon mesh (48 μm). Before analysis using flow cytometry, the cell nucleuses were stained by 150 μL dye buffer [propidium iodide (1 mg/mL), RNAse A (0.2 mg/mL)] in the dark for 10 min on ice. The genome size of *Lycium barbarum* was used as reference in genome survey and the experiment was performed on BD FACSCanto II.

Another method to evaluate the genome size was K-mer frequency analysis. The next-generation sequencing (NGS) data was firstly removed the adaptor and filtered the low-quality reads using Fastp (v0.23.2) [29] to generate k-mer (k = 21). Then the k-mer frequency was calculated by Jellyfish (v2.3.0) [30] and size of genome was further estimated by GenomeScope (v2.0) [31].

Genome assembly and evaluation

Based on the PacBio HiFi reads and ONT ultra-long reads, Hifiasm (v0.19.6, https://github.com/chhylp123/hifiasm) with default parameters was preliminarily used to generate primary contigs. After removing haplotigs using purge_haplotigs (v1.0.4) [32], the Hi-C data was further used to anchor the contigs to pseudochromosomes. After mapping the data to the primary contigs using HiCUP (v0.7.2) [33] with default parameters, the valid pairs were filtered by the Juicer pipeline (v1.6) [34] to anchor into pseudochromosomes with 3D-DNA pipeline (v180,922, parameters: -r 0, https://github.com/theaidenlab/3d-dna). Finally, the refinement of the Hi-C maps was performed using Juicebox (v1.11.08) [35].

To obtain T2T genome assembly, minimap2 (v 2–2.24) [36] was used to align ONT ultra-long reads to the genome, and the reads which mapped once at the terminal of pseudochromosome within 100 bp were collected. The reads with the most counts were labeled as reference, with others as queries, then both reference and queries were assembled to consensus sequences using medaka_consensu (v1.7.2, https://github.com/nanoporetech/medaka). Blastn (2.11.0 +) [37] was used to align the above consistent sequences to the terminal of pseudochromosome, and the sequences with over 90% coverage were used to replace the terminal sequences. In addition, ONT ultra-long reads were also used to fill the gaps between contigs by TGS-GapCloser (v1.2.0, parameters: –min_nread 10) [38], then the genome assembly underwent an error correction using Pilon (v1.23, parameters: –fix all –changes) [39] with short-read sequences.

To assess genome accuracy, minimap2 (v2.24-r1122) [36] and bwa (v0.7.12-r1039) [40] were further used to map Pacbio HiFi reads and short reads to genome assembly, respectively. Then the genome completeness was evaluated using BUSCO (v5.3.1, parameters: -m genome –limit 10) [41] with embryophyta_odb10 database.

Genome annotation

Repeat elements were annotated via homolog-based and de novo-based methods. According to the RepBase database, homolog repeat elements were identified using RepeatMasker (v4.0.9, https://www.repeatmasker.org) and RepeatProteinMask (v4.0.9) with default parameters, especially DNA transposons, long interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE), long terminal repeats (LTRs) and satellites. As for de novo annotation, a de novo library was firstly built by RepeatModeler (v1.0.11) [42] and LTR_FINDER_parallel (v1.0.7) [43] with default parameters, and the annotation was further performed by RepeatMasker (v4.0.9). In addition, tandem repeats in genome assembly were identified by Tandem Repeats Finder (v4.09) [44].

Three methods were applied for the structural annotation of protein-coding genes, including homology-based, de novo-based and transcriptome-based prediction. Homology-based prediction was performed using Liftoff (v1.6.3) [45] and Exonerate (v2.2.0, parameters: –model protein2genome –showtargetgff1) [46] combined the genomes and annotations of five closely related species (*Arabidopsis thaliana*, *Salvia hispanica*, *Salvia miltiorrhiza*, *Salvia splendens* and *Scutellaria baicalensis*).

Besides, the predicted models were constructed by GlimmerHMM (v3.0.4) [47] and AUGUSTUS (v3.3.2) [48] in de novo prediction. The coding sequence in ISOseq and traditional RNAseq data were identified by TransDecoder (v5.5.0, https://github.com/TransDecoder/TransDecoder) with default parameters in transcriptome-based prediction. Finally, MAKER2 (v2.31.10) [49] was applied to integrate the prediction results from three methods to generate a reliable gene set. As for non-coding RNA, transfer RNA (tRNA) was annotated using tRNAscan-SE (v1.3.1) [50], ribosomal RNA (rRNA) was identified by BLAST (2.11.0 +) [37], small nuclear RNA (snRNA) and microRNA (miRNA) was recognized using Rfam (v14.8, parameters: cmscan –rfam –no hmmonly) [51]. BUSCO analysis was also performed to evaluate the completeness of genome annotation.

For the functional annotation of protein-coding genes, two methods, including sequence similarity and domain/motif conservation, were applied. 9 databases about gene function, including KEGG (105.0, 2023–01-01), SwissProt (2023–03-01), TrEMBL (2023–03-01), InterPro (93.0, 2023–03-02), KOG (2003–03-01), GO (2023–04-01), Pfam, NR (2023–04-01) and PlantTFDB (5.0) were firstly used to performed function annotation based on the gene set of *D. moldavica* using diamond (v2.0.14, parameters: –evalue 1e-05) [52], then the annotation information of KEGG was correlated by KOBAS (v3.0, parameters: -t blastout:tab -s ko) [53]. Besides, the software InterProScan (v5.61–93.0, parameters: –seqtype p –formats TSV –goterms –pathways -dp) [54] was used to identify the motif and domain of *D. moldavica* according to the conserved sequences from InterPro and HMMER3 (v3.3.1) [55].

Comparative genomic analysis

To perform comparative genomic analysis, 14 plant species were firstly gathered for the gene family clustering using Blastp (2.11.0 +) and OrthoMCL (v2.0.9, parameters: -I 1.5) [56], which comprised *D. moldavica*, *D. rupestre*, *Andrographis paniculate*, *Arabidopsis thaliana*, *Dorcoceras hygrometricum*, *Oryza sativa*, *Salvia hispanica*, *Salvia miltiorrhiza*, *Sesamum indicum*, *Solanum lycopersicum*, *Antirrhinum majus*, *Lycium barbarum*, *Scutellaria baicalensis*, *Rosmarinus officinalis*. Then the single-copy orthologous genes were extract from those species for phylogenetic tree. The protein sequences of those genes were aligned using MAFFT (v7.481) [57] with default settings, then the alignment results were combined to create super-alignment matrixs. Finally, a maximum-likelihood phylogenetic tree was constructed using RAxML (v8.2.12, parameters: -f a -N 100 -m GTRGAMMA) [58]. The MCMCtree (v4.9, parameters: clock = 3;model = 0) [59] from PAML was used to estimate the divergence times of those species, and the calibration points were from the website TimeTree. The CAFE5 (v5.0.0) [60] was used to identify the expansions and contractions of gene families, followed by enrichment to GO and KEGG for functional annotation. Intergenomic synteny was inferred using WGDI (v0.5.6) [61] to search syntenic blocks and visualized using JCVI (v1.1.22) [62] among *D. moldavica*, *S. miltiorrhiza* and *S. baicalensis*. Then software WGDI was also applied to calculate the number of synonymous substitutions per synonymous site (Ks) between syntenic gene pairs from syntenic blocks, the distributions of Ks were used to infer whole-genome duplication events that occurred during the evolutionary history.

Identifying candidate functional UGT genes

The plant secondary product glycosyltransferase (PSPG) box is a conserved feature sequence of UGTs with 44 amino acids. Thus, PSPG box was firstly used as a query to identify UGTs via searching against *D. moldavica* genome using Blastp (2.11.0 +), and the proteins with evalue < 1e-5 were predicted as UGTs. Then the amino acid sequences of predicted UGTs were used for phylogenetic analysis with other plant UGTs. All UGT sequences were aligned using R package msa (v1.30.1) [63] with the method of "ClustalW", and a maximum-likelihood tree with 1000 bootstrap replicates was constructed via R package phangorn (v2.11.1) [64], the visitation of the phylogenetic tree was performed via R package ggtree (v3.6.2) [65].

Co-expression analysis

Total 15 transcriptome samples from five different tissues were gathered to construct the weighted correlation network using R package [66] (v1.72–5). The genes with low expression level [total of log2(TPM + 1) in 15 samples < 1] were removed. According to the scale-free topology criterion, the power value of adjacency functions for unsigned networks was determined ($R^2 > 0.85$), and unsigned type of topological overlap matrix (TOM) was calculated for WGCNA network construction and module detection (minModuleSize = 150, mergeCutHeight = 0.25). The Pearson correlation coefficient between metabolite content and module membership was calculated to get the related modules which might contain the functional genes.

Metabolomic analysis

Stems, leaves (mixed young and mature leaves) and flowers from *D. moldavica* with three biological replicates were sampled for metabolite sequencing. After freeze-drying using Scientz-100F vacuum freeze-dryer, the material was pulverized in the mixer mill (MM 400, Retsch) for 1.5 min at a frequency of 30 Hz. Then 50 mg powdered sample was dissolved in 1200 μL of −20 ℃

pre-cooled 70% methanol solution and vortexed for 30 s every 30 min, repeated six times. After centrifuging at 12,000 rpm for 3 min, the supernatant was filtered via microporous filter membrane with 0.22 μm pore size for UPLC-MS/MS analysis.

The data acquisition instrument system mainly included an UPLC-ESI–MS/MS system (UPLC, ExionLC™ AD, https://sciex.com.cn/) and Tandem mass spectrometry system (https://sciex.com.cn/). The Agilent SB-C18 column (1.8 μm, 2.1 mm × 100 mm) was used for chromatographic separation, and solvent A (water containing 0.1% formic acid) and solvent B (acetonitrile containing 0.1% formic acid) were used as mobile phase. A gradient of elution was started with 95% A and 5% B, then the proportion of solvent B was increased linearly to 95% within 9 min, after maintaining for 1 min, the proportion of solvent B was decreased to 5% within 1.1 min and this condition was holding for 2.9 min. The flow velocity was set at 0.35 mL/min, the column oven was set at 40 ℃ and the injection volume was 2 μL. After connecting to an ESI-triple quadrupole-linear ion trap (QTRAP)-MS, ESI source was at the temperature of 500 ℃ with an ion spray voltage (IS) of 5500 V (positive mode)/−4500 V (negative mode). Ion source gas I, gas II, and curtain gas were set at 50, 60, and 25 psi, respectively. Collision-activated dissociation (CAD) was set as high. Multiple reaction monitoring (MRM) transitions were performed in QQQ scans with collision gas (nitrogen) set to medium. Declustering potential and collision energy was done for each individual MRM transitions, and specific MRM transitions were monitored according to the metabolites eluted in each period.

The data of MS was processed using Analyst (v1.6.3) software. Metabolite identification was performed according to the Metware Database from Wuhan Metware Biotechnology Co., Ltd. The isotopic signals and duplicate signals, including Na^+, K^+, NH_4^+ and fragment ions belonging to other larger molecular weight substances, were removed during data analysis. Metabolites quantification was analyzed using MRM mode, and the count per second (CPS) of characteristic ions in all detected substances were obtained. After integration and correction of chromatographic peaks using MultiQuant (v3.0), the areas of peak were represented the relative content of the corresponding substances. Same metabolite across different samples were also used to normalize and correct the peak areas to ensure result accuracy.

As for analysis of differential metabolites, Principal Component Analysis (PCA) was firstly used to show the overall metabolites differences among the samples of each group. Then Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) was used to calculate the value of Variable Importance in Projection (VIP), which was performed via R package MetaboAnalystR (v3.0) [67]. The differentially accumulated metabolites were determined by the thresholds $VIP \geq 1$ and $|Log_2FC| \geq 1$. The data was log transform (log2) and mean centering before OPLS-DA. Finally, the data with Z-score normalization was used for cluster via K-means algorithm.

Enzyme activity assay *in vitro*

To verify the catalyze activity of putative UGTs, the CDS of UGTs were firstly cloned and inserted into the vector pDEST17 (Table S8). Then the recombinant plasmids were transformed into *E. coli* Rosetta (DE3), which grown on LB medium containing ampicillin for overnight. The monoclonal cultures were selected and transferred into 3 ml LB medium until the optical density 600 reached 0.6. After adding 0.2 mM Isopropyl-β-D-1-thiogalactopyranoside (IPTG), the cultures were grown for 18 h at 16℃ with shaking at 180 rpm to induce protein expression. The cells were harvested via centrifugating and resuspended in 30 mL 50 mM Tris–HCl buffer (containing 300 mM NaCl, pH 7.8), then 1 mM phenylmethylsulfonyl fluoride (PMSF) was added for protein extraction. Finally, after high-pressure disruption of cells and centrifugation at 12,000 rpm at 4℃ for 1 h, the supernatant was collected for protein purification. The crude protein supernatant was co-incubated with Ni-nitrilotriacetic acid agarose (Yeasen, 20503ES10), then the mixture was packed into a column and washed three times with 6 mL buffer 1 (50 mM Tris–HCl, 300 mM NaCl, 10 mM imidazole, pH 7.8), 6 mL buffer 2 (50 mM Tris–HCl, 300 mM NaCl, 20 mM imidazole, pH 7.8), 6 mL buffer 3 (50 mM Tris–HCl, pH 7.8, 300 mM NaCl, 50 mM imidazole), respectively. Finally, the protein was eluted by 1 mL buffer 4 (50 mM Tris–HCl, pH 7.8, 300 mM NaCl, 250 mM imidazole), SDS–polyacrylamide gel electrophoresis was used to test protein purification effect. The purified proteins were immediately used for an enzymatic assay. The reaction system (100 μL) consisted of 5 μg protein extract, 2 μL Acacetin (70 mM), 1 μL UDP-G (50 mM), 5 μL $MgCl_2$ (50 mM) and 10μL Tris–HCl (1M, pH 7.5). After incubation at 30℃ for 1 h, the products were tested via LC–MS/MS. The inactivation protein treated by high temperature was used for control group.

qRT-PCR analysis

To measure the expression levels of the 7GlcT genes, qRT-PCR was performed based on the cDNA using gene-specific primers. Total RNA was isolated from stems, young leaves, mature leaves and flowers of a plant with three repeats. The qRT-qPCR was performed on Applied Biosystems 7500 Real-Time PCR system using SGExcel FastSYBR Mixture (Sangon Biotech, B532955). *Actin* was used as the internal control gene. The relative expression

of genes was calculated using the $2^{-\Delta\Delta Ct}$ method. Primer sequences were listed in Table S1.

Results

Genome assembly and annotation of the *D. moldavica* T2T genome

A representative image of *D. moldavica* is shown in Fig. 1A. Based on the result of genome survey, the genome size of *D. moldavica* was predicted to be 266 Mb by K-mer frequency analysis (262.91 ± 2.74 Mb by flow cytometry), with a 0.435% heterozygosity ratio. Different sequencing data including ONT ultra-long reads (87.78 Gb) and PacBio HiFi reads (44.47 Gb) were firstly integrated to generate a primary assembly for *D. moldavica* with a contig N50 of 41.20 Mb, which was successfully anchored into 4 pseudochromosomes using Hi-C data (65.95 Gb) with a 94.28% anchoring rate (Table S2). Then the ONT ultra-long reads were also used for telomere extension and gap filling. By using telomere repeats as queries, 7 telomeres were identified on the ends of the 4 chromosomes (Fig. 1B). The downstream end of Chr3 had no motif repeat on which was not considered a telomere, every other chromosome had two telomeres located on both ends, and a maximum of 2,428 motif repeats was on the downstream end of Chr4 (Table S3). In addition, there was only a gap on Chr4 after gap filling. Short reads data were subsequently used for correction and polishing, a final 255.61 Mb genome of *D. moldavica* was obtained with a contig N50 length of 58.45 Mb, which was the first T2T genome in genus *Dracocephalum*. Then the short reads and HiFi reads were aligned to the final genome to assess correctness. The mapping rates of short and long reads were 96.49% and 99.62%, the coverage rates at least 10 × of genome based on short and long reads were 99.56% and 98.87%, respectively. The graphs of GC content and sequencing depth also showed the uniformity of sequencing (Fig. S1). BUSCO analysis was used to assess completeness, 98.57% and 95.85% of conserved genes were mapping to complete BUSCOs as well as single-copy BUSCOs.

The repeat elements were annotated using the assembly genome via homolog- and de novo-based methods. We found the proportion of repeat elements in the genome was 44.77%. Transposable elements (TEs) were the major repeat sequences made up 25.85%, and long terminal repeat (LTR) were the major TEs accounted for 5.20% of the genome. The protein-coding genes were predicted via multiple strategies, and a total of 27,425 genes were obtained with average gene length and coding sequence length of 3,319 bp and 1,296 bp (Table S4). Meanwhile, 36,254 transcripts were identified with mean length of 3,751.87 bp. BUSCOs analysis showed 99% of conserved gene orthologs were detected completely. Then 26,065 protein-coding genes (95.04%) and 34,717 mRNA (95.76%) were annotated by at least one of 9 databases (Table S5). For noncoding RNAs, 7,997 ribosomal RNAs (rRNAs), 635 transfer RNAs (tRNAs), 387 small nuclear RNAs (snRNAs) and 109 microRNAs (miRNAs) were predicted (Table S6). Figure 1C depicted the integrated distributions, including GC content and density distribution of gene, repeat sequence, LTR, LINE and DNA-TE.

Comparative genomic analysis

To investigate the evolutionary traits, the genome of *D. moldavica* was compared with 13 species with published genomes, including 9 plants in Lamiales, 2 plants in Solanales and 2 model plants (*A. thaliana* and *O. sativa*). After clustering analysis, 16,184 gene families were identified in *D. moldavica*, including 270 unique families and 853 unique genes (Table S7). Among 14 species, a total of 6,500 common gene families were found and 716 single-copy orthologous genes were used to construct the phylogenetic tree (Fig. 2A, Fig. S2). Then we found *D. moldavica* and *D. rupestre* (2 plants in *Dracocephalum*) had closest relationship with divergence time of 8.5 million years ago (Mya). Besides, *S. baicalensis* diverged from Lamiaceae around 47.3 Mya, *D. moldavica* and *D. rupestre* then diverged from Lamiaceae around 21.9 Mya (Fig. 2B).

The analysis of expansion and contraction of gene families was further performed to find special traits in evolution, from which 884 expanded and 997 contracted gene families were identified in *D. moldavica* (Fig. 2B). GO enrichment analysis of expanded gene families showed the genes enriched in many metabolic-related pathways, including "UDP-glycosyltransferase activity", "diterpenoid biosynthetic process", "sulfotransferase activity" and "protein dimerization activity". Meanwhile, KEGG enrichment analysis revealed the expanded gene families enriched in the terms related to secondary metabolism such as "Flavonoid biosynthesis", "Phenylpropanoid biosynthesis" and "Ascorbate and aldarate metabolism" (Table S8).

Intragenomic synteny analysis was used to determine the number of whole-genome duplication (WGD) events in *D. moldavica*. Firstly, 1,014 and 1,075 syntenic blocks were identified in *D. moldavica* vs. *S. baicalensis* and *D. moldavica* vs. *S. miltiorrhiza*, respectively. The graph of intragenomic synteny showed a low frequency of chromosomal fragment rearrangement among *D. moldavica*, *S. baicalensis* and *S. miltiorrhiza* (Fig. 2C). Then only one peak in the distribution of synonymous substitutions per synonymous site (Ks) indicated that *D. moldavica* only experienced a WGD event (Fig. 2D).

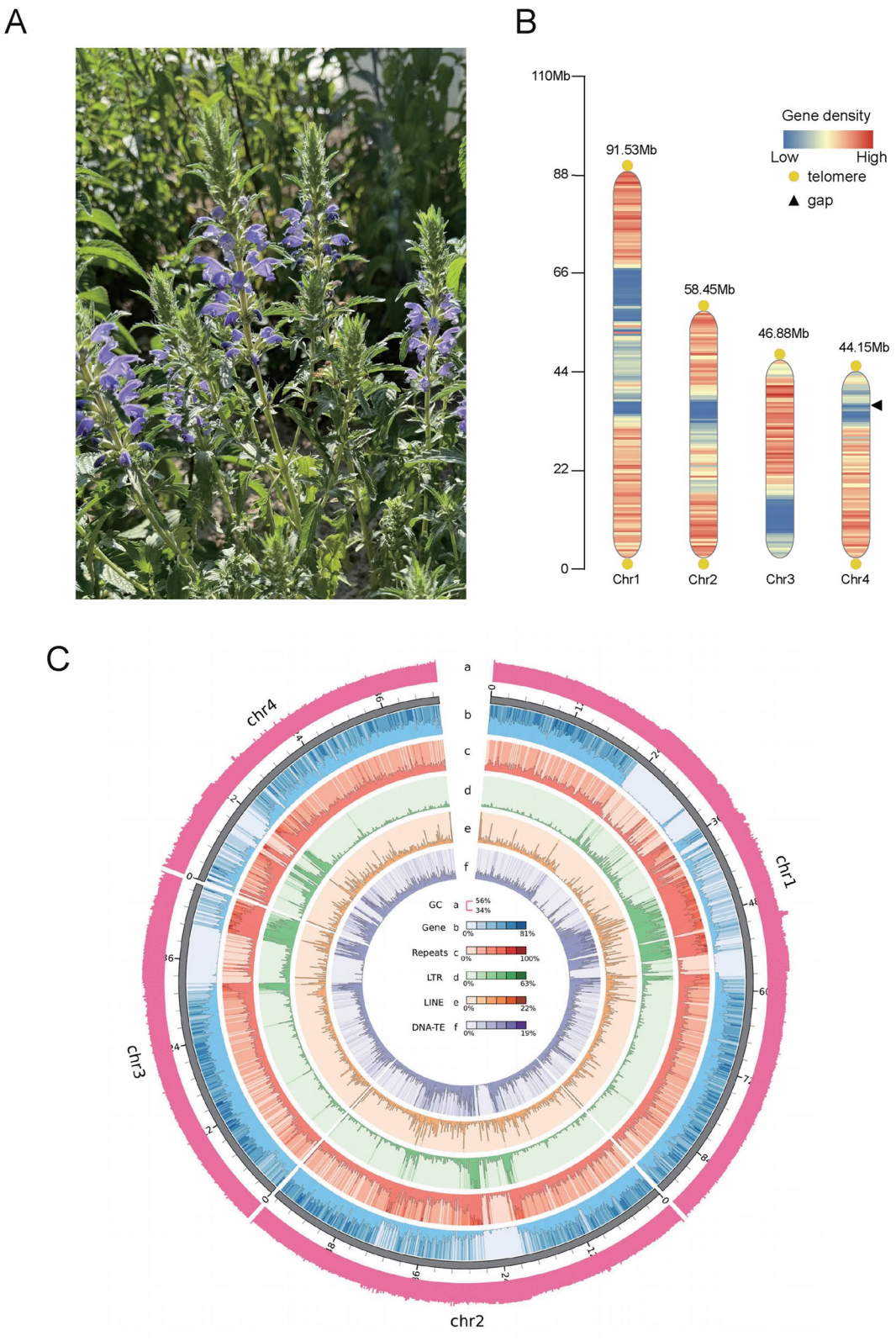

Fig. 1 T2T genome assembly and genomic features of *D. moldavica*. **A** Photograph of the *D. moldavica* plant. **B** Display of gene density and telomere distribution in *D. moldavica*, with the scale indicating chromosomes in 500 kb segments. **C** Genomic overview of the four assembled chromosomes of *D. moldavica*, presented in 80 kb segments

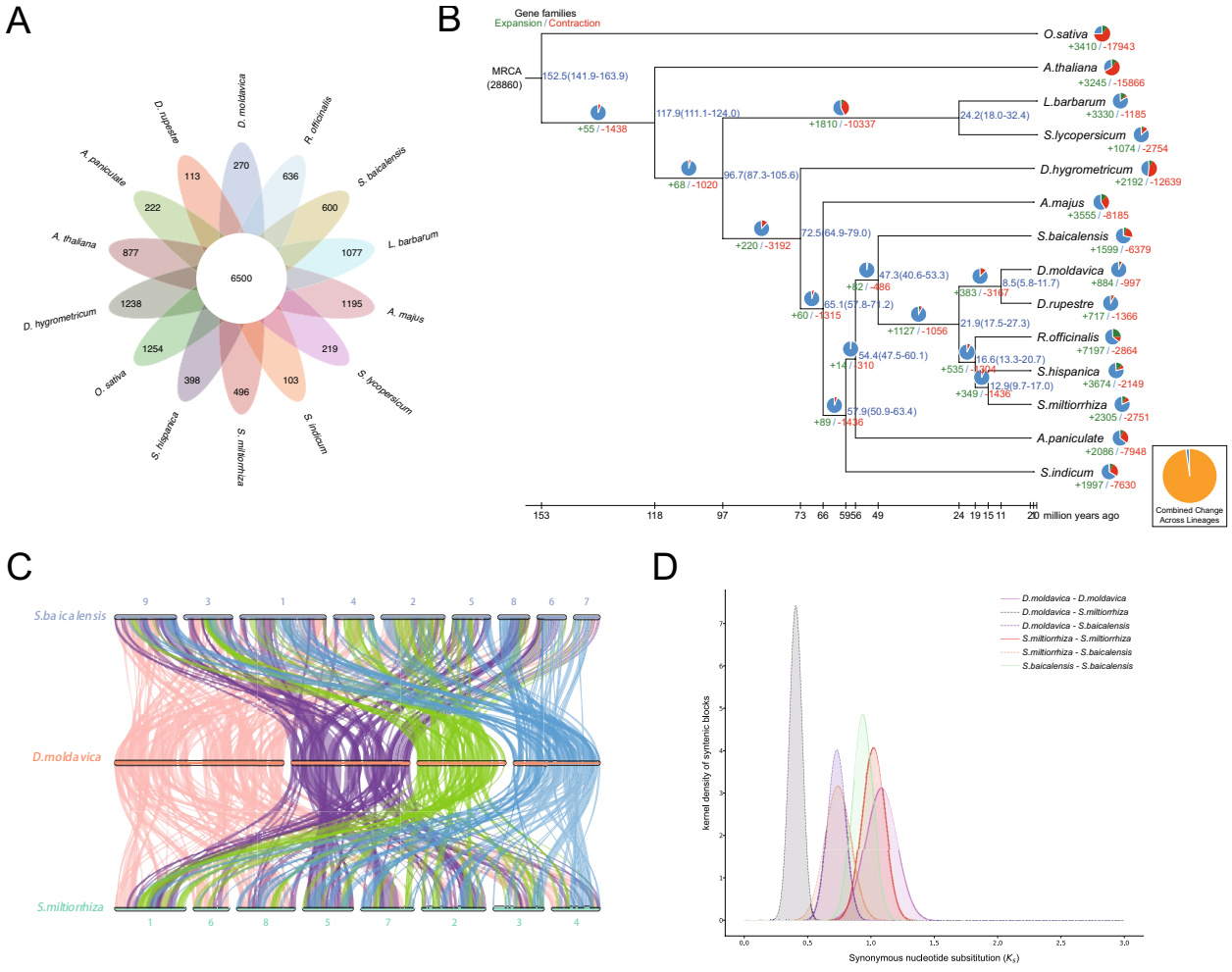

Fig. 2 Comparative genomic analysis of *D. moldavica*. **A** Floral diagram illustrating shared gene families. **B** Phylogenetic tree of 14 plant species. **C** Genome synteny comparison among *D. moldavica*, *S. baicalensis* and *S. miltiorrhiza*. **D** Ks distribution of WGD-derived gene pairs in *D. moldavica*, *S. baicalensis* and *S. miltiorrhiza*

Identification of UDP-glycosyltransferase (UGT) genes in *D. moldavica*

According to previous research, the UGTs, especially 7-O-glycosyltransferases (7GTs) might catalyze the glucosylation of acacetin at 7-OH to form tilianin [25]. To identify the key genes in the biosynthesis of tilianin, we used PSPG box to query the genome of *D. moldavica* and found 157 high-confidence UGTs with the length of amino acids ranging from 275 to 552. The potential UGT genes in *D. moldavica* were evenly dispersed across the four chromosomes, among which 17 expanded genes were mainly on Chr1 and Chr4 while 9 contracted genes were mainly on Chr3. Then the full-length amino acid sequences of 40 known UGTs from other plants in GenBank and all predicted UGTs were used for phylogenetic tree construction. Based on the functionally characterized UGTs, 50 UGTs were clustered into five groups, including 3-O-glycosyltransferases (3GTs), 5-O-glycosyltransferases (5GTs), 7GTs, C-glycosyltransferases (CGTs) and glycoside glycosyltransferase (GGTs). We found 16 UGTs were clustered in the 7GTs group which might comprise the UGTs catalyzed the biosynthesis of tilianin (Fig. 3A). Among 16 possible 7GT genes, there were 7 of them on Chr1, followed by 5 on Chr3, and Chr2 as well as Chr4 had 2 genes, respectively. According to the phylogenetic tree, we found two protein sequences, Dm7GlcT1 and Dmo_1G0004630, were clustered with many known flavonoid 7-O-glucosyltransferase such as Sb7GlcT in *S. baicalensis*, NtIS5a in *Nicotiana tabacum*, NjUGT73B1 in *Nardostachys jatamansi*, especially NjUGT73B1, which could efficiently catalyze the glucosylation of acacetin at 7-OH to produce tilianin. Hence, *Dm7GlcT1* and *Dmo_1G0004630* might be the candidate genes to participate the biosynthesis of tilianin.

The expression pattern of UGT genes were further deciphered using transcriptomic data in five tissues, including roots, stems, young leaves, mature leaves and flowers. Firstly, 29,617 expressed genes were used to

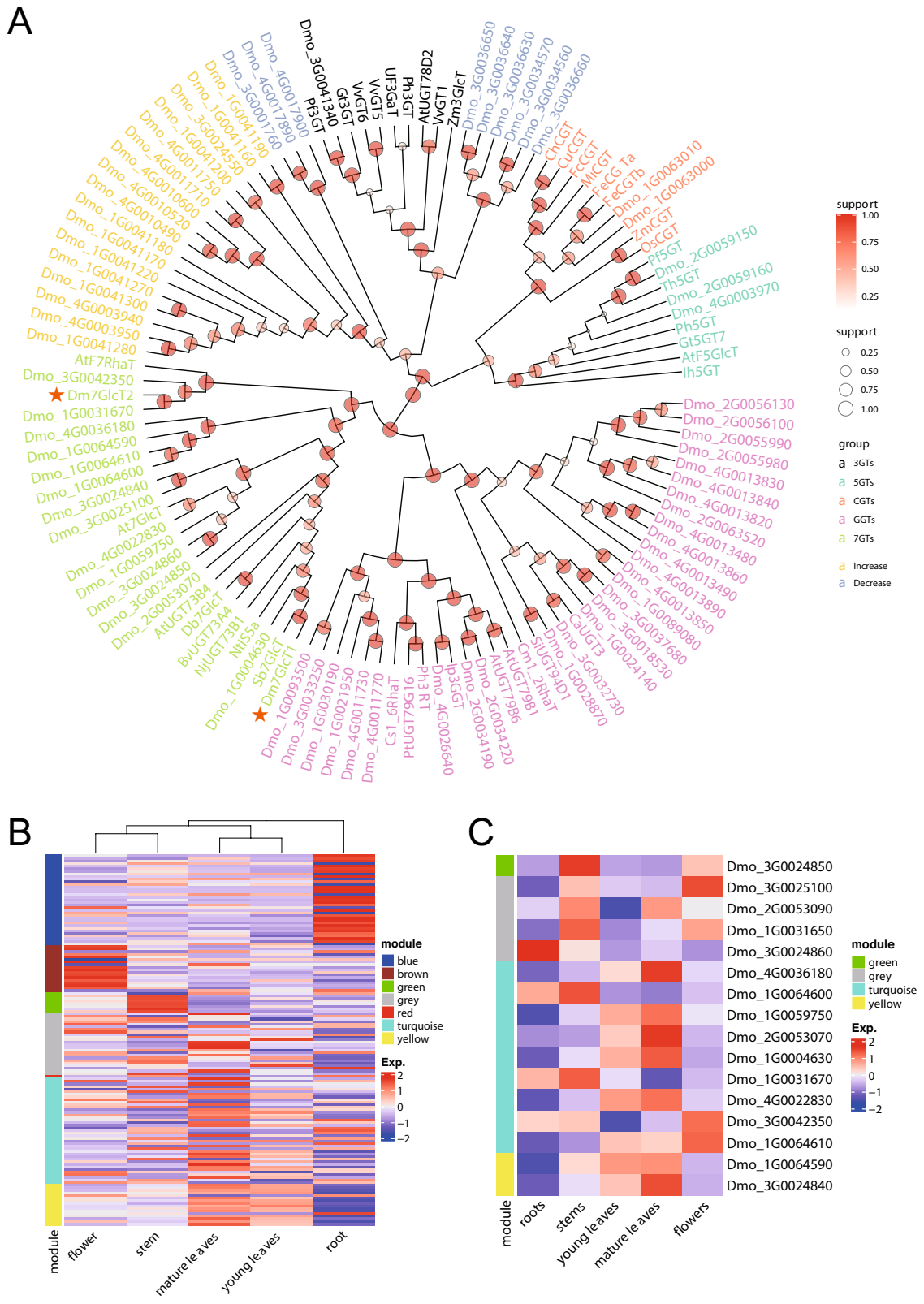

Fig. 3 UTG genes in *D. moldavica*. **A** Phylogenetic tree of annotated UGTs in *D. moldavica* and other plant species, increase and decrease meant the UGTs in the expanded and contracted gene families. **B** Expression patterns of UTG genes across various tissues. **C** Expression patterns of 7GT genes across different tissues

construct the weighted co-expression network with the power of 26, and 7 modules were identified including six meaningful modules and a grey module consisted of genes without consistent expression pattern. The expression heatmap showed the UGT genes in brown, blue, green and yellow module were highly expressed in flowers, roots, stems and leaves, respectively, which suggested the diverse activity of UGTs in different tissues might contribute to the glycosylation of various secondary metabolites in different stages (Fig. 3B). As for the candidate genes from phylogenetic analysis, we found *Dm7GlcT1* and *Dmo_1G0004630* were in grey and turquoise module, respectively. *Dm7GlcT1* was not a tissue specific expression gene and exhibited relatively high expression levels in stems, mature leaves and flowers, while *Dmo_1G0004630* was mainly expressed in leaves (Fig. 3C).

Metabolome profile of secondary metabolites in *D. moldavica* at different tissues

The secondary metabolites were abundant in *D. moldavica*. However, fewer studies focused on the kind of compounds and total content of the key compounds in different tissues of *D. moldavica*. Thus, we used the data of metabolome to fill in the blanks in related research fields. Based on UPLC-MS/MS platform and special database, 1,407 metabolites were identified in three tissues (stems, leaves and flowers). PCA plot showed the high consistency of the repeat sample in each tissue (Fig. S3). Those metabolites could be classified into 9 classes, including 481 Flavonoids, 264 Terpenoids, 231 Phenolic acids, 144 Alkaloids, 102 Lignans and Coumarins, 22 Quinones, 8 Tannins, 4 Steroids and 151 Others (Fig. 4A). Analysis of differentially accumulated metabolites (DMs) was performed between two of three tissues. Compared to stems, 607 and 621 DMs had significantly higher relative content in flowers and leaves, and most of them were flavonoids, suggesting flowers and leaves might be the main tissues for flavonoids storage. Then more terpenoids exhibited high relative content in leaves, while relatively few DMs were abundant in stems (Fig. 4B).

Based on previous reports, naringenin was a common flavonoid precursor derived from naringenin chalcone, which regulated by two key enzymes, chalcone synthase (CHS) and chalcone isomerase (CHI) [68]. Then the flavone synthases catalyzed the synthesis of the intermediate apigenin from naringenin [25]. After methylation of apigenin to form acacetin by O-methyltransferase, the tilianin was biosynthesized from acacetin via adding a glucosyl group [25]. In addition, the rhamnosylation of tilianin could form linarin which was the key step in the biosynthetic pathway of linarin [25] (Fig. 4C). To investigate the biosynthesis process of tilianin, we focused on the relative content of related metabolites. Tilianin exhibited relative higher content in stems and flowers than in leaves, while the accumulation of acacetin was mainly in flowers and leaves, less in stems. Then the content of intermediate apigenin and two precursors including naringenin chalcone and naringenin were mainly accumulated in flowers, suggesting the flowers were the main tissues for acacetin biosynthesis in *D. moldavica* (Fig. 4D). There were two possible reasons for different accumulation of tilianin and acacetin in different tissues: two metabolites were biosynthesized in different tissues or tilianin was transported to stems after biosynthesis. Besides, the relative content of following products linarin showed a similar trend with tilianin which was significantly higher in stems and flowers.

The correlation analysis between the data of transcriptome and metabolome was performed to find the key 7GT genes. The results showed only one 7GT gene, *Dm7GlcT2*, with higher expression level in stems and flowers, whose expression level was positively associated with the relative content of tilianin ($R = 0.9997$, $p = 0.01476$, Fig. 4E). Based on the accumulation of tilianin and acacetin, we supposed that the key 7GT genes showed high expression levels in stems or flowers, which was consistent with the expression pattern of *Dm7GlcT2*. Thus, we suspected *Dm7GlcT2* might be a UTGs with function of tilianin biosynthesis. Furthermore, the expression pattern of Dmo_3G0025100, Dmo_3G0024850 and Dmo_3G0042350 also showed the similar trends with *Dm7GlcT2*, but all of them might be the 7GTs with low activity based on the homologous protein with known function [69–72]. We also focused on the two candidate 7GT genes, *Dmo_1G0004630* and *Dm7GlcT1*, despite their expression showed insignificant correlation with content of any metabolites, *Dm7GlcT1* was highly expressed in stems and moderately expressed in flowers which was consistent with the accumulation pattern of tilianin, suggesting the gene might have catalytic function in those tissues.

Function verification of candidate 7GTs via enzyme activity assay *in vitro*

The integrated results of multi-omics research indicated two 7GT genes, *Dm7GlcT1* and *Dm7GlcT2*, named *Dm7GlcT1* and *Dm7GlcT2*, might catalyze the glucosylation of acacetin at 7-OH to form tilianin. To test the enzyme activity of Dm7GlcT1 and Dm7GlcT2 *in vitro*, the recombinant expression plasmids containing *Dm7GlcT1* and *Dm7GlcT2* coding sequences were firstly transformed into *E. coli* to induce proteins expression. After purification, the single bands of Dm7GlcT1 and Dm7GlcT2 were predicted as 54 kDa and 52.4 kDa via SDS-PAGE analysis, respectively. The purified proteins were then used for enzymatic assay, and the reactions were analyzed by LC–MS/MS. The results showed that a

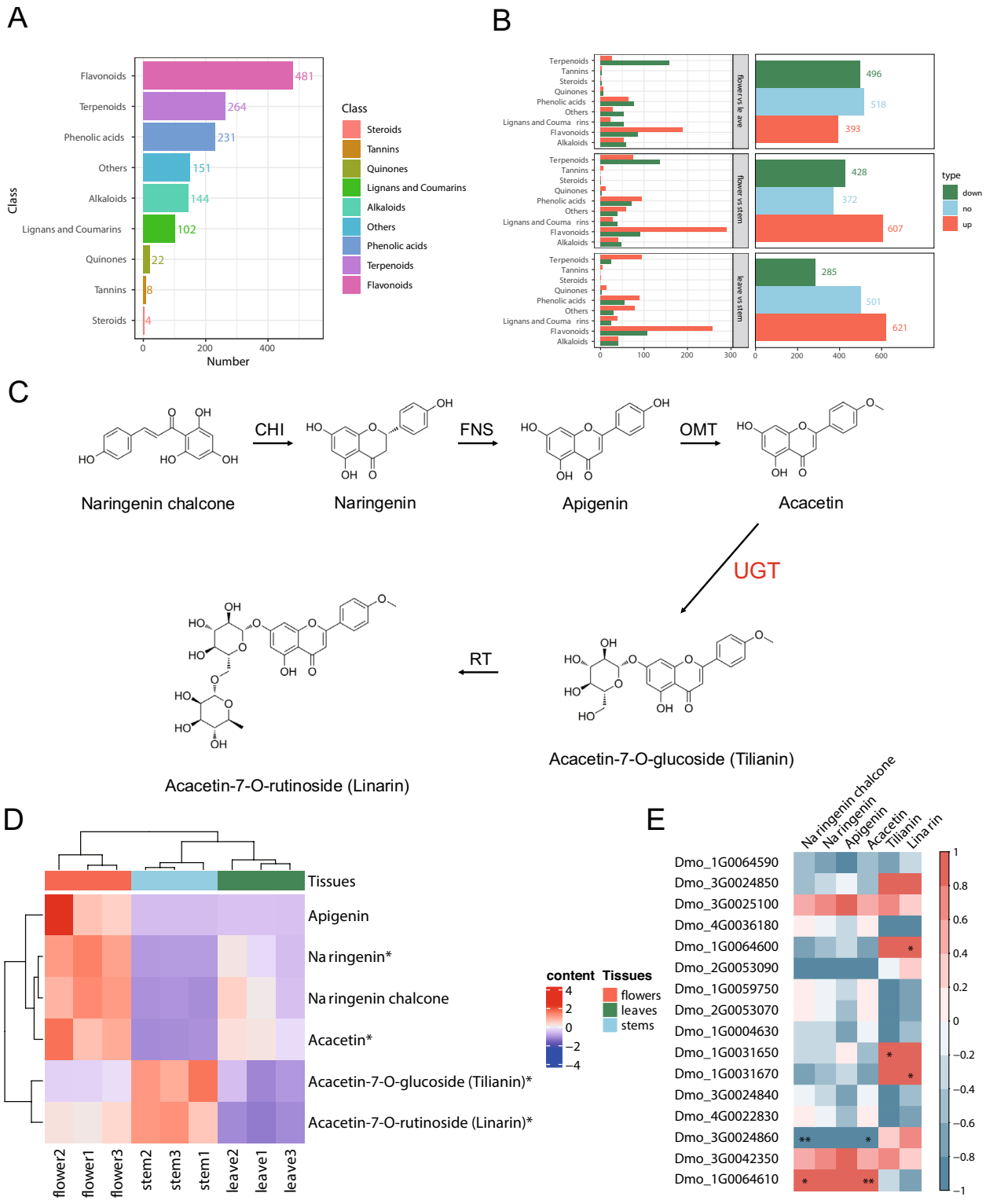

Fig. 4 Metabolome profile of major secondary metabolites in *D. moldavica*. **A** Categories of secondary metabolites identified in the metabolome. **B** Visualization of differentially accumulated metabolites (DMs) across different groups. **C** Possible biosynthetic pathways of tilianin in *D. moldavica*. **D** Relative levels of tilianin and five related metabolites in *D. moldavica*. **E**. Pearson correlation analysis between the relative contents of 6 metabolites and the expression levels of 7GT genes in the transcriptome. * meant $p < 0.05$ and ** meant $p < 0.01$

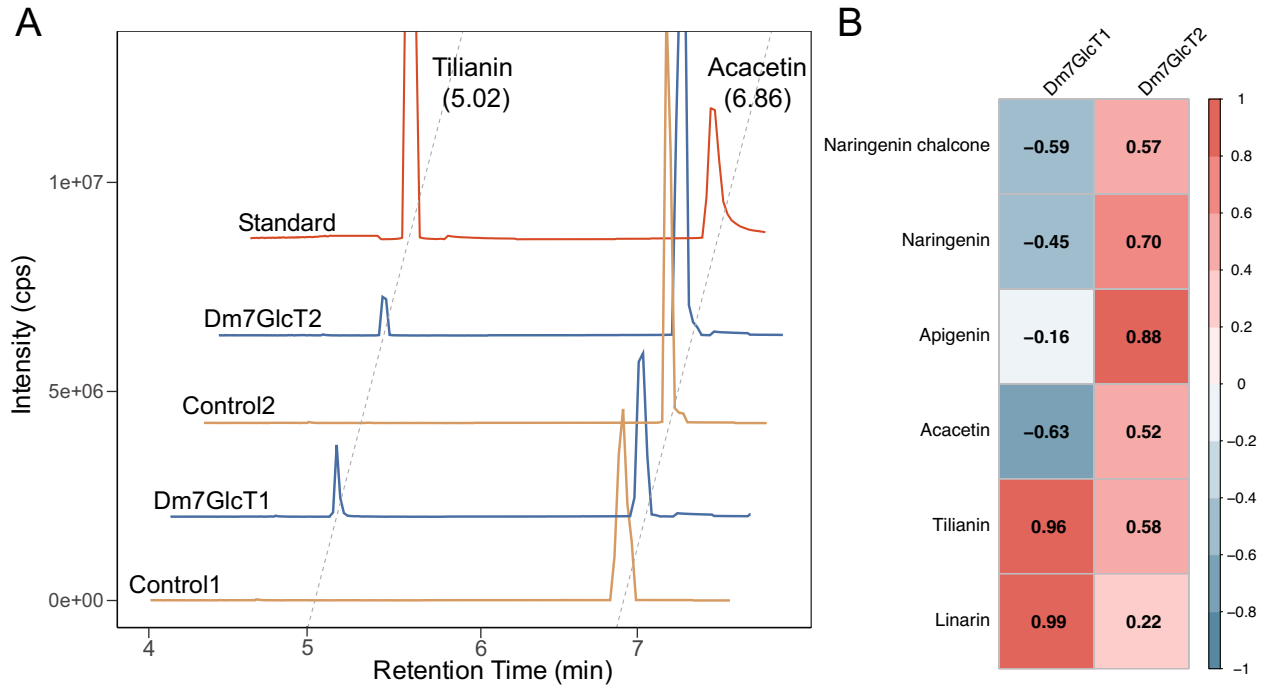

Fig. 5 Enzyme activity assay and qRT-PCR validation. **A** LC–MS/MS analysis of the reactions catalyzed by Dm7GlcT1 and Dm7GlcT2, including controls and standards. **B** Pearson correlation analysis between the relative levels of 6 metabolites and the relative expression levels of 7GT genes measured by qRT-PCR. The values shown in the heatmap represent the correlation coefficients

new peak appeared on the chromatogram compared with control group, and the peaks of reference standards confirmed the new peak was tilianin, indicating Dm7GlcT1 and Dm7GlcT2 could catalyze the 7-OH glucosylation of acacetin. According to the intensity of product tilianin, Dm7GlcT1 might have higher catalytic efficiency than Dm7GlcT2 (Fig. 5A). Then we used qRT-PCR to verify the expression pattern of two 7GT genes. The relative expression level of *Dm7GlcT1* was higher in stems and moderate in flowers. Correlation analysis indicated the expression of *Dm7GlcT1* showed positive correlation with the content of tilianin and linarin in metabolome, while *Dm7GlcT2* did not exhibit a significant expression bias, but with obviously higher expression in flowers. Moreover, the expression of *Dm7GlcT2* also positively correlated with content of tilianin (Fig. 5B, Fig. S4). The results suggested that Dm7GlcT1 might catalyze tilianin biosynthesis in stems and flowers, while glucosylation catalyzed by Dm7GlcT2 might happen in all tissues.

Discussion

D. moldavica is one of the typical plants in genus *Dracocephalum* which is renowned for its diverse natural products, including flavonoids, terpenoids, phenolic acids and alkaloids [8]. As traditional folk medicine, the pharmacological activity of *D. moldavica*, including cardiovascular protection, anti-oxidative, antimicrobial, antifungal, anti-complementary, anti-cancer, anti-inflammatory, has been extensively explored in medical experiments [2]. Previous studies have predominantly focused on the active compounds of *D. moldavica*, especially flavonoids [5–7]. Tilianin as a primary constituent in the total flavonoid extract of *D. moldavica*, has huge medical value including treatment of cardiovascular and cerebrovascular diseases [2] as well as human pharyngeal squamous cell carcinoma [73]. Due to the difficulty in chemical synthesis, the main source of tilianin is still from plant [5], particularly *D. moldavica*. Thus, the importance of the investigation of tilianin biosynthetic pathway in *D. moldavica* is self-evident.

High-quality genome is the foundation of the biosynthesis research of key metabolites [68, 74, 75]. Therefore, we first assembled a T2T genome of *D. moldavica* using multiple sequencing technologies. The size of genome assembly is approximately 255.61Mb and a contig N50 is 58.45 Mb. Then 94.28% of the assembled bases were anchored into 4 pseudochromosomes. Despite benefits from the long-read sequencing technologies, the *D. moldavica* genome assembly still has a gap in Chr4 because of the genomic complexity. Nevertheless, as the first T2T genome in genus *Dracocephalum*, the release of *D. moldavica* genome provides more clues for the functional research of other similar traditional Chinese herbs.

Published study showed the tilianin was biosynthesized from acacetin via adding a glucosyl group which was catalyzed by UGTs [25]. To explore the final step

in tilianin biosynthetic pathway, 157 high-confidence UGTs were further identified based on sequence similarity, among which 16 UGTs were clustered in the 7GTs group. According to phylogenetic tree, Dm7GlcT1 and Dm7GlcT2 were in a main branch with NjUGT73B1, a known flavonoid 7-O-glucosyltransferase. Moreover, the sequence identity of Dm7GlcT1 and Dm7GlcT2 with NjUGT73B1were 52.94% and 31.89%, respectively. Correlation analysis between expression level and the accumulation of metabolites showed two genes, *Dm7GlcT1* and *Dm7GlcT2*, encoded two 7-O-glucosyltransferases which might catalyzed acacetin to form tilianin. However, tilianin and its substrate acacetin was not consistent in the main accumulated tissues, which might cause by three reasons: acacetin was synthesized primarily in flowers but glucosylated in stems; tilianin was synthesized in flowers then transported to stems; different UGTs hold different catalytic efficiency in different tissues or developmental stages. From transcriptome data, we found both *Dm7GlcT1* and *Dm7GlcT2* exhibited higher expression level in stems, conforming with the high accumulation of tilianin in stems, suggested *Dm7GlcT1* and *Dm7GlcT2* could catalyzed tilianin biosynthesis in stems. In addition, the data of metabonomics showed tilianin was mainly accumulated in stems, which could provide a potential strategy for the genetic breeding screen or gene modification of *D. moldavica* with productive tilianin based on the tissue-specific expression patterns.

The enzyme activity assay *in vitro* was used to verify the catalytic function of Dm7GlcT1 and Dm7GlcT2. The result of LC–MS/MS showed the intensity of product tilianin catalyzed by Dm7GlcT1 was slightly higher than that catalyzed by Dm7GlcT2, implied the stronger catalytic ability of Dm7GlcT1. Moreover, qRT-PCR results indicated *Dm7GlcT1* was mainly expressed in stems, moderately expressed in flowers, and *Dm7GlcT2* did not show the significant expression difference in different tissues. Therefore, we supposed Dm7GlcT1 might be a main catalytic enzyme during tilianin biosynthesis in stems, while the catalytic ability of Dm7GlcT2 did not show the tissue specificity. Besides, the non-specific expression implied the multiple catalytic functions of Dm7GlcT2 based on different substrates, which need further research.

Multi-omics approach is a powerful tool to identify the key enzymes catalyzed the biosynthesis of medicinal compounds [68, 74, 75]. In our study, genome assembly with high quality is the basis of functional genomics investigation, then transcriptome and metabolome were used to narrow down the range of candidate genes. This approach helped us to identify two UGTs with function of tilianin biosynthesis, which not only filled the gap in knowledge of the metabolic pathways related with tilianin of *D. moldavica* but also provided insights into *D. moldavica* quality improvement. Moreover, these resources, especially T2T genome, will greatly promote *D. moldavica* breeding, varieties identification and utilization.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12551-y.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.
Supplementary Material 8.
Supplementary Material 9.
Supplementary Material 10.
Supplementary Material 11.
Supplementary Material 12.

Acknowledgements
Not applicable.

Authors' contributions
ZLW is the corresponding author, and he contributes to the conception of the study. HY contributes to design. QZ complete the bioinformatics analysis. XXZ and WFW contribute to the experiment ***in vitro***. QZ and XXZ contribute to data analysis and interpretation. All authors contribute to the writing of the manuscript.

Funding
Zhanli Wang receives the National Natural Science Foundation of China (82170296) and the Science and Technology Program of the Joint Fund of Scientific Research for the Public Hospitals of Inner Mongolia Academy of Medical Sciences (2024GLLH0581). Hui Yu receives funding from National Natural Science Foundation of China (82260092).

Data availability
The T2T genome assembly data of ***D. moldavica*** have been deposited in the Genome Warehouse in National Genomics Data Center (NGDC, https://bigd.big.ac.cn/gwh) with accession number GWHHAGL00000000. The raw sequence data of ***D. moldavica*** are available in the Genome Sequence Archive (GSA) in NGDC (https://ngdc.cncb.ac.cn/gsa, GSA: CRA033009). Besides, the raw data also have been deposited in National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) with the BioProject accession number of PRJNA1357323 and BioSample accession number of SAMN53168731.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 4 November 2025 / Accepted: 12 January 2026
Published online: 23 January 2026

References

1. Sultan A, et al. Flavonoids from *Dracocephalum moldavica*. Chem Nat Compd. 2008;44(3):366–7.
2. Zhan M, et al. *Dracocephalum moldavica* L.: An updated comprehensive review of its botany, traditional uses, phytochemistry, pharmacology, and application aspects. Fitoterapia. 2024;172:105732.
3. Tan M, et al. Development of solid lipid nanoparticles containing total flavonoid extract from *Dracocephalum moldavica* L. and their therapeutic effect against myocardial ischemia-reperfusion injury in rats. Int J Nanomedicine. 2017;12:3253–65.
4. Maimaiti A, et al. Improvement of total flavonoids from *Dracocephalum moldavica* L. in rats with chronic mountain sickness through (1)H-NMR metabonomics. Evid Based Complement Alternat Med. 2021;2021:6695346.
5. Jia J, et al. The inhibitory effects of *Dracocephalum moldavica* L. (DML) on rat cerebral ischemia reperfusion injury. J Toxicol Environ Health A. 2017;80(22):1206–11.
6. Zheng RF, et al. Total flavones of *Dracocephalum moldavica* L. protect astrocytes against H_2O_2-induced apoptosis through a mitochondria-dependent pathway. BMC Complement Med Ther. 2020;20(1):78.
7. He C, et al. Total flavonoid extract from *Dracocephalum moldavica L*. improves pulmonary fibrosis by reducing inflammation and inhibiting the hedgehog signaling pathway. Phytother Res. 2023;37(7):2745–58.
8. Xu M, et al. Qualitative and quantitative analysis of chemical components of *Dracocephalum moldavica* based on UPLC-Q-TOF-MS/MS and UPLC. Zhongguo Zhong Yao Za Zhi. 2024;49(23):6352–67.
9. Sun T, et al. Tilianin improves cognition in a vascular dementia rodent model by targeting miR-193b-3p/CaM- and miR-152-3p/ CaMKIIα-mediated inflammatory and apoptotic pathways. Front Immunol. 2023;14:1118808.
10. Li Y, et al. Evaluation of anti-inflammatory and antioxidant effects of chrysanthemum stem and leaf extract on zebrafish inflammatory bowel disease model. Molecules. 2022;27(7):2114.
11. Kim S, et al. Herbal combinational medication of *glycyrrhiza glabra*, *agastache rugosa* containing glycyrrhizic acid, tilianin inhibits neutrophilic lung inflammation by affecting CXCL2, interleukin-17/STAT3 signal pathways in a murine model of COPD. Nutrients. 2020;12(4):926.
12. Xu S, et al. Tilianin protects against nonalcoholic fatty liver disease in early obesity mice. Biol Pharm Bull. 2023;46(3):419–26.
13. Li J, Xu S. Tilianin attenuates MPP(+)-induced oxidative stress and apoptosis of dopaminergic neurons in a cellular model of Parkinson's disease. Exp Ther Med. 2022;23(4):293.
14. García-Díaz J, et al. Antidiabetic, antihyperlipidemic and anti-inflammatory effects of tilianin in streptozotocin- nicotinamide diabetic rats. Biomed Pharmacother. 2016;83:667–75.
15. Zhang R, et al. Renoprotective effects of tilianin in diabetic rats through modulation of oxidative stress via Nrf2-Keap1 pathway and inflammation via TLR4/MAPK/NF-κB pathways. Int Immunopharmacol. 2020;88:106967.
16. Yao J, et al. Synergistic cardioptotection by tilianin and syringin in diabetic cardiomyopathy involves interaction of TLR4/NF-κB/NLRP3 and PGC1a/SIRT3 pathways. Int Immunopharmacol. 2021;96:107728.
17. Tian L, et al. Pretreatment with tilianin improves mitochondrial energy metabolism and oxidative stress in rats with myocardial ischemia/reperfusion injury via AMPK/SIRT1/PGC-1 alpha signaling pathway. J Pharmacol Sci. 2019;139(4):352–60.
18. Jiang H, et al. Tilianin protects against ischemia/reperfusion- induced myocardial injury through the inhibition of the Ca(2+)/calmodulin- dependent protein kinase II-dependent apoptotic and inflammatory signaling pathways. Biomed Res Int. 2020;2020:5939715.
19. Khattulanuar F, et al. Tilianin: a potential natural lead molecule for new drug design and development for the treatment of cardiovascular disorders. Molecules. 2022;27(3):673.
20. Jiang H, et al. Tilianin ameliorates cognitive dysfunction and neuronal damage in rats with vascular dementia via p-CaMKII/ERK/CREB and ox-CaMKII- dependent MAPK/NF-κB pathways. Oxid Med Cell Longev. 2021;2021:6673967.
21. Yuan Y, et al. Tilianin pretreatment prevents myocardial ischemia-reperfusion injury via preservation of mitochondrial function in rat heart. Phytomedicine. 2017;34:106–14.
22. Syed Abd Halim S, et al. Natural products targeting PI3K/AKT in myocardial ischemic reperfusion injury: a scoping review. Pharmaceuticals (Basel). 2023;16(5):739.
23. Zeng C, et al. Cardioprotection of tilianin ameliorates myocardial ischemia-reperfusion injury: role of the apoptotic signaling pathway. PLoS ONE. 2018;13(3):e0193845.
24. Wen J, et al. Tilianin extracted from *Xiangqinglan* inhibits apoptosis induced by mitochondrial pathway and endoplasmic reticulum stress in H9c2 cells after oxygen-glucose deprivation/reoxygenation. J Tradit Chin Med. 2023;43(1):42–50.
25. Feng M, et al. An efficient flavonoid glycosyltransferase NjUGT73B1 from *nardostachys jatamansi* of alpine Himalayas discovered by structure-based protein clustering. Phytochemistry. 2024;227:114228.
26. Ren VC, et al. Identification of UDP-rhamnosyltransferases and UDP-galactosyltransferase involved in flavonol glycosylation in *Morella rubra*. Hortic Res. 2022;9:uhac138.
27. Wu B, et al. Genome-wide identification, expression patterns, and functional analysis of UDP glycosyltransferase family in peach (*Prunus persica* L. Batsch). Front Plant Sci. 2017;8:389.
28. Allen G, et al. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc. 2006;1(5):2320–5.
29. Chen S, et al. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
30. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
31. Ranallo-Benavidez T, et al. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.
32. Roach M, et al. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
33. Wingett S, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 2015;4:1310.
34. Durand N, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
35. Durand N, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101.
36. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
37. Altschul S, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
38. Xu M, et al. TGS-gapcloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience. 2020;9(9):giaa094.
39. Walker B, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9(11):e112963.
40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
41. Simão F, et al. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
42. Price A, et al. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–8.
43. Ou S, Jiang N. Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10:48.
44. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
45. Shumate A, Salzberg S. Liftoff: accurate mapping of gene annotations. Bioinformatics. 2021;37(12):1639–43.
46. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
47. Majoros W, et al. Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20(16):2878–9.
48. Stanke M, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.
49. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
50. Lowe T, Eddy S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–64.
51. Griffiths-Jones S, et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;33:D121–4.
52. Buchfink B, et al. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18(4):366–8.
53. Bu D, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. Nucleic Acids Res. 2021;49(W1):W317–25.

54. Jones P, et al. Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
55. Mistry J, et al. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):e121.
56. Li L, et al. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
57. Nakamura T, et al. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34(14):2490–2.
58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
59. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
60. Mendes F, et al. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2021;36(22–23):5516–8.
61. Sun P, et al. WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant. 2022;15(12):1841–51.
62. Tang H, et al. Synteny and collinearity in plant genomes. Science. 2008;320(5875):486–8.
63. Bodenhofer U, et al. msa: an R package for multiple sequence alignment. Bioinformatics. 2015;31(24):3997–9.
64. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3.
65. Yu G. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinform. 2020;69(1):e96.
66. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
67. Chong J, Xia J. Metaboanalystr: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics. 2018;34(24):4313–4.
68. Wen J, et al. An integrated multi-omics approach reveals polymethoxylated flavonoid biosynthesis in *Citrus reticulata cv. chachiensis*. Nat Commun. 2024;15(1):3991.
69. Lim EK, et al. Arabidopsis glycosyltransferases as biocatalysts in fermentation for regioselective synthesis of diverse quercetin glucosides. Biotechnol Bioeng. 2004;87(5):623–31.
70. Poppenberger B, et al. The UGT73C5 of *arabidopsis thaliana* glucosylates brassinosteroids. Proc Natl Acad Sci U S A. 2005;102(42):15253–8.
71. Hou B, et al. N-glucosylation of cytokinins by glycosyltransferases of *Arabidopsis thaliana*. J Biol Chem. 2004;279(46):47822–32.
72. Poppenberger B, et al. Detoxification of the Fusarium mycotoxin deoxynivalenol by a UDP-glucosyltransferase from Arabidopsis thaliana. J Biol Chem. 2003;278(48):47905–14.
73. Jiang H, et al. Tilianin extracted from *Dracocephalum moldavica* L. induces intrinsic apoptosis and drives inflammatory microenvironment response on pharyngeal squamous carcinoma cells via regulating TLR4 signaling pathways. Front Pharmacol. 2020;11:205.
74. Pei T, et al. Gap-free genome assembly and CYP450 gene family analysis reveal the biosynthesis of anthocyanins in *Scutellaria baicalensis*. Hortic Res. 2023;10(12):uhad235.
75. Qi H, et al. The high-quality genome of lotus reveals tandem duplicate genes involved in stress response and secondary metabolites biosynthesis. Hortic Res. 2023;10(5):uhad040.

Publisher's Note