DNA Research, 2022, 29(2), 1–10
https://doi.org/10.1093/dnares/dsac008
Advance Access Publication Date: 5 April 2022
Resource Article: Genomes Explored

Resource Article: Genomes Explored

The chromosome-level genome assembly of *Gentiana dahurica* (Gentianaceae) provides insights into gentiopicroside biosynthesis

Ting Li, Xi Yu, Yumeng Ren, Minghui Kang, Wenjie Yang, Landi Feng, and Quanjun Hu *

Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China

*To whom correspondence should be addressed. Tel. +86 13258195631. Fax. +86 028 8541 2053.
Email: huquanjun@scu.edu.cn

Received 23 January 2022; Editorial decision 1 April 2022; Accepted 2 April 2022

Abstract

Gentiana dahurica Fisch. is a perennial herb of the family Gentianaceae. This species is used as a traditional Tibetan medicine because of its rich gentiopicroside constituents. Here, we generate a high-quality, chromosome-level genome of *G. dahurica* with a total length of 1,416.54 Mb. Comparative genomic analyses showed that *G. dahurica* shared one whole-genome duplication (WGD) event with *Gelsemium sempervirens* of the family Gelsemiaceaei and had one additional species-specific WGD after the ancient whole-genome triplication with other eudicots. Further transcriptome analyses identified numerous enzyme coding genes and the transcription factors related to gentiopicroside biosynthesis. A set of candidate *cytochrome P450 genes* were identified for being involved in biosynthetic shifts from swertiamarin to gentiopicroside. Both gene expressions and the contents measured by high-performance liquid chromatography indicated that the gentiopicrosides were mainly synthesized in the rhizomes with the highest contents. In addition, we found that two above-mentioned WGDs, contributed greatly to the identified candidate genes involving in gentiopicroside biosynthesis. The first reference genome of Gentianaceae we generated here will definitely accelerate evolutionary, ecological, and pharmaceutical studies of this family.

Key words: medicinal plant, genome assembly, *Gentiana dahurica*, transcriptome, gentiopicroside biosynthesis

1. Introduction

Gentiana dahurica Fisch. (Gentianaceae) ($2n = 26$) of Sect. *Cruciata*[1] is widely distributed in grasslands, mountain meadows and hillside forests at altitude of 800–4500 m in southwest and northwest China.[2] The rhizomes and flowers of this species are widely used as traditional Chinese medicinal 'Xiao-Qin-Jiao' to treat rheumatoid arthritis and cure sore throat and cough for more than 2000 years.[3] The major secondary metabolites of this species comprise multiple different seco-iridoid glycosides (including, gentiopicroside, loganic acid, swertiamarin, and sweroside).[4,5] Among them, gentiopicrosides were demonstrated to have significant anti-inflammatory, analgesic, and antibacterial properties, which are also valuable in the treatment of immune system diseases in both clinical and pharmaceutical practices.[6] More than 1000 species are acknowledged for the Gentianaceae and the genus *Gentiana* contains around 360 species and 15 sections over the world.[2] As the typical section, Sect. *Cruciata* comprise around 22 species and most species (including *G. dahurica*) occurring in China, contain the highly concentrated

© The Author(s) 2022. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

gentiopicroside constituents and have been continuously used as traditional Chinese medicines.[3,5]

Gentiopicrosides comprise as one of terpenoid indole alkaloids (TIAs). TIAs are also found in several families of Gentianales, including Apocynaceae, Loganiaceae, Gentianceae, and Nyssaceae.[7,8] On the basis of annotated metabolites and enzymes identified, biosynthesis of gentiopicroside was suggested to be similar to that of iridoids in *Catharanthus roseus.*[9,10] Such a candidate pathway comprises three major stages. First, the precursor of terpenoids, isopentenyl diphosphate (IPP) is synthesized by either cytosolic mevalonic acid (MVA)[8,11] or plastidial 2-C-methyl-D-erythritol-4-phosphate (MEP).[4,12] Second, dimethylallyl diphosphate reacts with IPP to generate geranyl diphosphate, which is then converted into geraniol.[13,14] Finally, secologanin arises from geraniol by a series of enzymatic reactions.[15,16] However, all likely candidate genes have not been identified or confirmed in the species from Gentianaceae. In addition, whole-genome duplication (WGD) was found to occur in other families with TIAs.[8,9] The WGD-derived genes were found to be involved in biosynthesis of these chemicals.[8] It remains unknown whether such an independent WGD also occurred in the Gentianaceae because up to now, no *de novo* genome was reported for the family.

Here, we assemble a chromosome-level reference genome of *G. dahurica* using a combination of three technologies (Nanopore, Illumina paired-end and Hi-C), it's also the first genome of the whole Gentianaceae family. Using comparative and evolutionary analyses, we explored genomic evolution of this species and recovered one WGD event with *Gelsemium sempervirens* of the family Gelsemiaceaei and had one additional species-specific WGD events. Based on this reference genome, we further employed transcriptome and metabolome data of different tissues to examine expressions of the candidate genes and contents of the related chemicals in order to provide further insights into gentiopicroside biosynthesis. These genomic resources will be highly useful for evolutionary, ecological and pharmaceutical studies of the species in the Gentianaceae in the future.

2. Materials and methods

2.1. Plant materials collection

One wild *G. dahurica* ($2n = 26$) individual (Fig. 1A) was collected from Sunan Yugur Autonomous County, Zhangye City, Gansu Province, China (E 99°28′43.15″, N 38°58′40.62″). In the field, fresh leaves and stems of *G. dahurica* from the single plant were harvested, immediately frozen at −80°C and kept in liquid nitrogen for extracting the genomic DNA (gDNA) or total RNA. Leaves from a single individual were used for gDNA extraction and genome assembly. Leaves and stems from the other individuals were used for RNA-seq.

2.2. DNA extraction and genome sequencing

In order to extract high-quality gDNA, we used the QIAGEN Blood & Cell Culture DNA Kit. We then selected the high molecular weight gDNA (targeting 10–50 kb) using a Blue Pippin system (Sage Science, Beverly, MA, USA) and further processed the Nanopore sequencing library with the Ligation sequencing 1D kit (SQK-LSK108, ONT, UK) according to the manufacturer's instructions. We sequenced the resulting library through the GridION X5 sequencer (ONT, UK) at the Genome Center of Nextomics (Wuhan, China). Base calling was further carried out on fast5 files using the ONT Albacore software v0.8.4, and low-quality reads (mean_qscore <7) and adapter sequences were filtered. Sequencing libraries were also prepared with gDNA using Illumina Genomic DNA Sample Preparation Kit and sequenced on an Illumina HiSeq X Ten system in paired-end mode (2 × 150 bp). Adapter sequences and low-quality reads were removed using NGS QC Toolkit v 2.3.3.[17] The obtained clean data were used for error correction and k-mer analysis. The Hi-C library was prepared from 3 g of freshly ground young leaves, using liquid nitrogen with a mortar and pestle. The chromatin extraction, digestion, DNA ligation, purification, and fragmentation were all performed as previously described.[18]

2.3. Heterozygosity evaluation

We used Illumina sequencing reads to evaluate the level of heterozygosity in *G. dahurica*. The heterozygosity level was estimated using GenomeScope v2.0[19] with 17-mers. The *k-mer* analysis was performed by Jellyfish v2.29.[20]

2.4. Genome assembly and chromosome construction

Genome size was confirmed by *k-mer* analysis using findGSE v0.1[21] with Illumina short reads. All Nanopore long reads were corrected using canu-correct and trimmed by canu-trim for low-quality bases, and the assembly was performed with Canu v1.7[22] and polished the chromosome-level genomes with two iterations using Pilon v1.23.[23] Then, the Hi-C reads were aligned to the assembly using the Juicer v1.6.2.[24,25] The assembly was scaffolded with Hi-C data using the 3D-DNA v180922 with default parameters,[26] and manually curated using the Juicebox Assembly Tools v1.11.08.[27] The Hi-C scaffolding resulted in six chromosome-level super scaffolds, representing a total of 95.36% of the assembled sequence. The completeness of the assembly was evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) v4.1.2 (embryophyte_odb10, 2020-08-05).

2.5. Transcriptome library preparation and sequencing

RNA-seq experiments (three biological replicates) used RNAs extracted from different tissues of *G. dahurica:* old-roots, roots, old-stems, stems, old-leaves, leaves, flowers, and corollas (Fig. 1A). All materials were collected from three mature and flowering individuals with nearly same ages. All samples were transported on dry ice, washed with ultrapure water three times, immediately frozen in liquid nitrogen, and stored at −80°C before RNA extraction. The total RNAs were extracted using Qiagen RNeasy Plant Mini Kits, and sequenced using Illumina HiSeq X Ten system in paired-end mode (2 × 150 bp). Prepared libraries were sequenced on the Illumina HiSeq 2500 platform according to the manufacturer's recommended protocol.

2.6 Repeats annotation

Orthologous repetitive elements in the *G. dahurica* genome were identified using RepeatMasker v4.0[28] and RepeatModeler v4.07[29] with default settings. Intact long terminal repeat (LTR) retrotransposons were identified with LTRharvest v1.5.10[30] and LTR_Finder v1.06[31] with LTR length set to range from 100 to 5,000 bases and the length between two LTRs set to 1,000–20,000 bases. The LTR_retriever v1.9[32] was used to combine results from LTRharvest and LTR_Finder, and estimate the insertion times of LTR retrotransposon. The insertion times were estimated using $T = K/2\mu$,[33] where K is the divergence rate and μ is the neutral mutation rate[34] (115–130 million years).

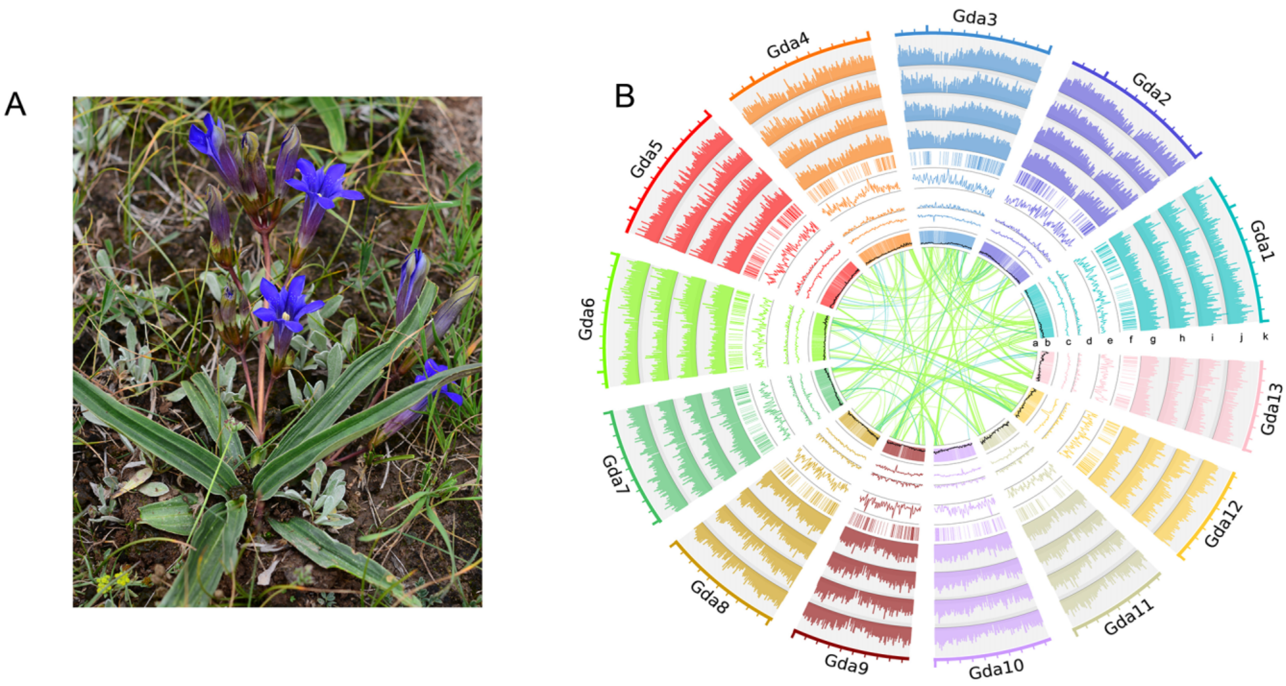

Figure 1. Summary of *G. dahurica* genome assembly. (A) Photo of *G. dahurica*. (B) *G. dahurica* genomic landscape of diversity and expression data. (a) Syntenic blocks in *G. dahurica* genomes. The band width is proportional to syntenic block size. (b–f) The distribution of the gene density, Gypsy elements, Copia elements and GC density, tandem density, respectively. (g–j) Expression of organ-specific genes (from inside to outside tracks: root, stem, leaf, and flower). (k) Circular representation of the pseudomolecules.

2.7. Gene prediction and annotation

A combination of *de novo*-, homology-, and transcript-based methods was used for gene prediction. After quality filtering with Trimmomatic v0.33,[35] a *de novo* and a genome-guided transcripts assembly was performed on Illumina RNA-seq reads using Trinity v2.6.6.[36] Then, transcript-based gene predictions were built with the PASA pipeline v2.1.0.[37] Homologs were predicted by mapping protein sequences from *Arabidopsis thaliana*,[38] *Capsicum baccatum*,[39] *C. roseus*,[40] *Camptotheca acuminata*,[8] *Coffea canephora*,[12] *Daucus carota*,[41] *Calotropis gigantea*,[42] *Olea europaea*,[43] *Dorcoceras hygrometricum*,[44] *G. sempervirens*,[45] *Striga asiatica*,[46] and *Salvia splendens*[47] (Supplementary Table S2) to the *G. dahurica* genome using exonerate v2.4.0.[48] A *de novo* gene prediction was performed with Augustus v3.2.3[49] and GlimmerHMM v3.0.4.[50] Augustus parameters are trained using ORFs predicted by PASA.[49] Gene models from the three main sources (i.e. aligned transcripts, *de novo* predictions and aligned proteins) were merged to produce consensus models by EVidenceModeler v1.1.1.[51] The functional annotation for all genes were generated by alignment to public protein databases including Swiss-Prot and TrEMBL.[52] Protein domains were annotated by searching against InterPro database.[53] The GO terms and metabolic pathways were annotated using Blast2GO v2.5[54] and KEGG Pathway databases.[55]

2.8. Phylogenetic tree construction and divergence time estimation

A phylogenetic tree was built from clusters of gene families for the *G. dahurica* and several other species: *C. baccatum*, *O. europaea*, *C. roseus*, *C. canephora*, *C. gigantea*, *D. hygrometricum*, *G. sempervirens*, and *Gardenia jasminoides*[56] (Supplementary Table S2). Gene families were firstly constructed using OrthoFinder v2.3.12.[57] The longest protein encoding sequence at each gene locus for each gene model was retained to remove redundancy caused by alternative splicing. Orthogroups with only one gene copy per species (Single-copy orthogroups) were collected, and aligned using MUSCLE v.3.8.31.[58] Subsequently, the phylogenetic trees were constructed by RAxML v8.2.11.[59] Divergence time was estimated from the phylogenetic tree using MCMCTree from PAML v4.9.[60] The calibration time for divergence time estimation was obtained from the TimeTree database.[61]

2.9. Gene family expansion and contraction analyses and WGD events detection

The expansion or contraction of orthologous gene families was determined using CAFE v4.2.[62] The program uses a birth and death process to model gene gain and loss over phylogenic distance. Gene families that had undergone expansion and/or contraction were calculated using the phylogeny and divergence times with the parameters: P-value = 0.05, number of threads = 10. We investigated WGD events in *G. dahurica* (Gentianaceae), *C. roseus* (Apocynaceae), and, *C. acuminata* (Nyssaceae) by selecting two species as references: *Vitis vinifera*[63] and *C. canephora* without more species-specific WGD events after whole-genome triplication (WGT) of the eudicots (WGT-γ).[63] We used WGDI[64] to detect syntenic blocks for three species: *C. canephora*, *V. vinifera*, and *G. dahurica*. Based on genes in syntenic blocks, we calculated synonymous substitution rates (Ks) to examine potential WGD events.

2.10. Transcriptome assembly and gene expression analysis

RNA-seq reads from three replicates of the eight tissue types were preprocessed using Trimmomatic[35] by removing adaptor sequences and filtering low-quality reads. Salmon[65] was used to align the samples with the genome for genome-guided transcript assembly. Read counts extracted from Stringtie were filtered using the R package sva[66] to decrease batch effects and hidden variables. Differentially expressed genes were detected using DESeq2,[67] and calculated based on absolute $\log_2$ transformed fold-change values greater than 2 and *P*-value of 0.05 using the Benjamini–Hochberg correction.[68] A gene set enrichment analysis[69] was performed to determine significant gene sets, and the WGCNA package[70] was applied to perform a multivariate analysis of gene co-expression modules.

Genes encoding key enzymes thought to be involved in the gentiopicroside biosynthetic pathway were annotated by BLAST v2.2.28. The proteins were aligned to the Pfam database using HMMER v3.1b1 for domain annotation. Gene expression levels in different tissues were obtained from transcriptome data. In addition, we identified transcription factors in the *G. dahurica* genome by comparison with the PlantTFDB database.

2.11. Medicinal components extraction and statistical analysis

The samples from the same seven tissues previously for RNA-seq were further used for analyses of chemical contents. For each sample, 20 mg of powder was prepared and further extracted with 400 µl of 80% aqueous methanol at 4°C, followed by centrifugation for 10 min at 12,000 rpm. LC-MS analysis was performed using the Waters Acquity UPLC System connected to an AB SCIEX 5500 QQQ-MS. Gradient elution was achieved on a Waters Acquity UPLC BEH C18 column (100 mm*2.1 mm, 1.7 µm) with water containing 0.1% formic acid (solvent A) and acetonitrile (solvent B) at a flow rate of 0.30 ml/min. The column temperature was maintained at 40°C. The gradient elution program was as follows: 1–10%B (0–1 min), 11–60%B (2–5 min), 60–90%B (5–7 min), held at 99%B (7–9 min), and allowed to equilibrate for a further 3 min before the next injection and the last 8 min of the chromatogram solutions were discarded. The injection volume was 4 µl. MS data were recorded with the following parameters: Ion source, ESI; IonSource temperature, 450°C; IonSource Gas1, 55arb; IonSource Gas2,55arb; IonSpray voltage, 4500 V; Curtain Gas, 35arb; Collision GAS, 7arb.Components eluting from the UPLC-QQQ-MS system were processed in MultiQuant for data preprocessing with default settings, except that each sample was normalized to the internal standard.[71] After filtering for outliers, the data were used for the subsequent statistical analysis. One-way analysis of variance was used to compare various contents among different tissues. If the variances were significantly different, Tamhane's T2 test was used to perform *post hoc* analyses; otherwise, a least significant difference test was used. The results were considered significant at *P*-value <0.05.

3. Results

3.1. Genome assembly and gene annotation

We generated a total of 233.1 Gb sequencing data for genome assembly. After filtering, 173.0 Gb long reads with the N50 size of 27.4 Kb were generated (Supplementary Table S3). The genome size was estimated to be 131, 7.1 Mb by *k-mer* analysis (Supplementary Fig. S1). The genome was firstly assembled into contigs with a total length of 1449.94 Mb and contig N50 was 1.30 Mb. We then anchored these contigs into thirteen pseudochromosomes with Hi-C reads using 3D-DNA.[26,72] The final sequences anchored on pseudochromosomes is 1416.54 Mb in length with scaffold N50 = 113.31 Mb (Supplementary Table S1, Fig. 1B, Supplementary Fig. S2). Genome assembly completeness evaluation suggests a total of 96.5% complete BUSCOs were present (Supplementary Table S4). The heterozygosity level was estimated to be ~1% in the assembled genome (Supplementary Fig. S1).

A total of 995.11 Mb (70.25%) repetitive sequences were identified and most of them were classified as LTR retrotransposons (57.57%) (Supplementary Table S5). In total, 37,988 genes were predicted, with an average gene length, coding sequence length and an average exon number of 4,159 bp, 263 bp, and 4.55 exons, respectively (Supplementary Table S1). The gene prediction showed 96.5% coverage of complete BUSCOs (Supplementary Table S7). In our assembly, 97.57% of the genes (37,065 out of 37,988) were annotated on the thirteen pseudochromosomes, only 2.43% (923 out of 37,988) remained on unplaced scaffolds. These statistics revealed that the newly assembled genome had high coverage and accuracy in genic regions. Among the 37,988 predicted genes, we annotated 44.33%, 89.92%, 39.70%, and 22.84% of the genes using the Swiss-Prot, InterPro, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway databases, respectively (Supplementary Table S8).

3.2. Whole-genome duplication

Gentianales contains five families: Rubiaceae, Gentianaceae, Loganiaceae, Gelsemiaceae, and Apocynaceae. As the first reference Gentianaceae genome, we investigated the phylogenetic position of the *G. dahurica* to explore phylogenetic position of Gentianaceae within the Gentianales. Here, we used 1,045 single-copy orthologs from the genomes of the other seven species (four Gentianales, two Lamiales, and one Solanales) to construct a phylogenetic tree, suggesting that Gentianaceae and Gelsemiaceae have a closer relationship with each other than to Apocynaceae. We estimated that the Gentianales and Lamiales diverged around 83.16 million years ago (Mya) (76.14–90.65 Mya), and Gentianaceae diverged from Gelsemiaceae approximately 47.03 Mya (41.12 –53.18 Mya) (Fig. 2A).

Firstly, we identified collinear blocks pairs of *G. dahurica* with *V. vinifera* and *C. canephora* and recovered 1:4 syntenic ratios in *G. dahurica—V. vinifera and G. dahurica—C. canephora* comparisons (Fig. 2C and Supplementary Figs S3 and S4), suggesting that *G. dahurica* experienced more WGD events after WGT-γ. Secondly, we calculated the value of synonymous substitutions per synonymous site (Ks) for the collinear gene pairs. The Ks distribution of *C. canephora* and *V. vinifera* (Fig. 2D) recovered the WGT-γ, consistent with the previous reports.[73] The Ks distribution between *G. dahurica* and *C. canephora* indicated their divergence after WGT-γ. The Ks distribution of *G. dahurica* showed two peaks at 0.766 and 1.064, suggesting that this species had experienced two more species-specific WGD events occurred 42.78–53.14 Mya and 30.39–37.75 Mya (Fig. 2D and Supplementary Fig. S5). We found that the earlier WGD was also shared by *G. sempervirens* (Gelsemiaceae) and an independent WGD event was detected between 67.11 and 73.83 Mya for *C. acuminata* after WGT-γ as found before[8,9] (Supplementary Fig. S6).

3.3. Gene family analyses

We clustered the annotated genes into gene families for *G. dahurica* and other four species of Gentianales, *C. roseus*, *C. canephora*, *G. sempervirens*, and *C. gigantea*. A total of 34,960 *G. dahurica* genes

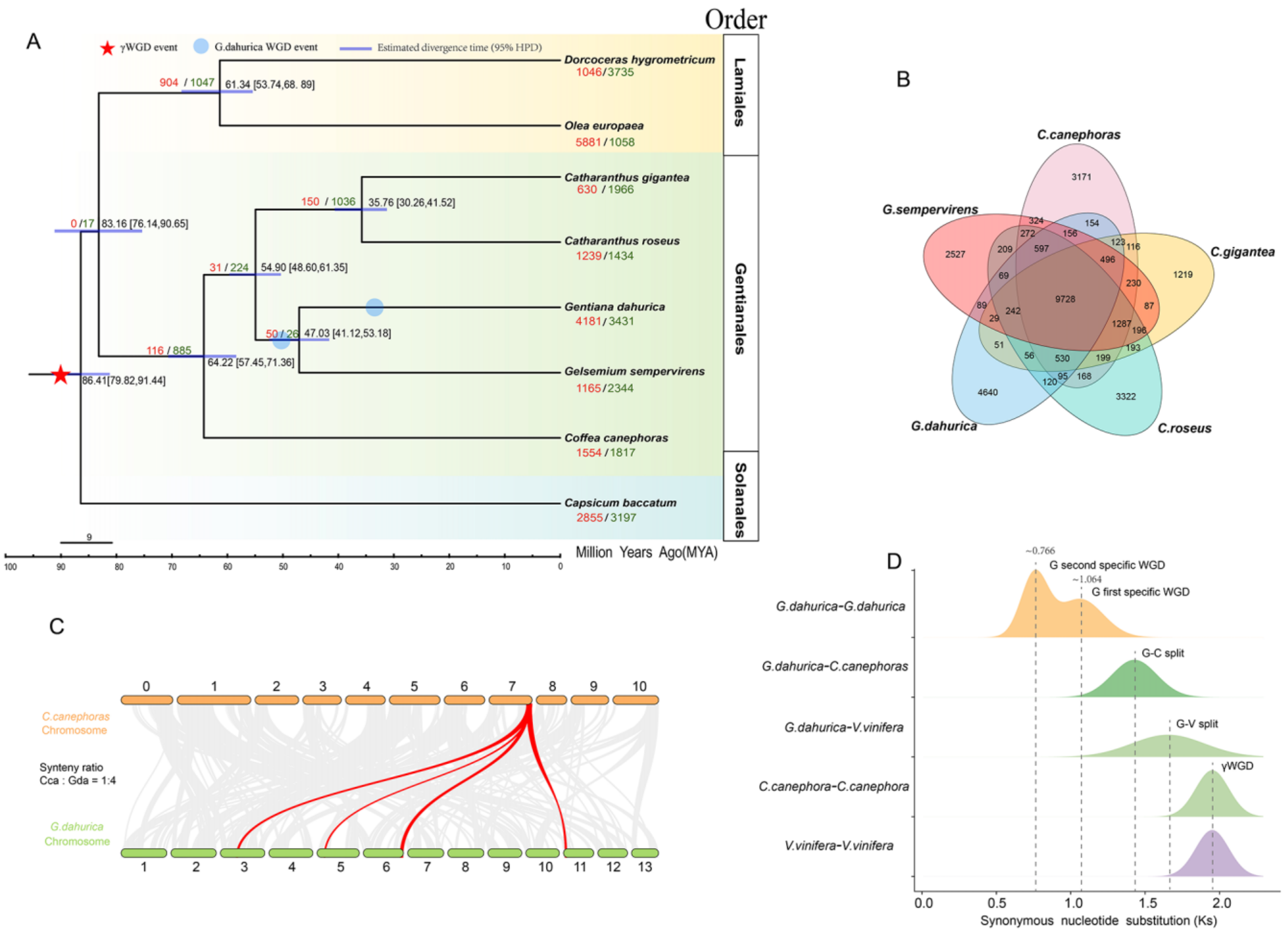

Figure 2. Phylogenetic analysis of the *G. dahurica* genome. (A) Phylogenetic tree for *G. dahurica* and seven other plants: (*Dorcoceras hygrometricum*, *Catharanthus roseus*, *Coffea canephora*, *Gelsemium sempervirens*, *Calotropis gigantea*, *Olea europaea*, *Capsicum baccatum*), divergence time and gene family expansions and contractions displayed on a maximum likelihood tree. Divergence times were estimated using MCMCTree and are indicated by bars at the internodes with 95% highest posterior density (HPD). Circles represent recent whole-genome duplication (WGD) events. (B) Shared gene families by *G. dahurica* and four other species. The numbers indicate gene families identified among all selected species. (C) Collinear relationship between *G. dahurica* and *C. canephora* chromosomes. The collinearity pattern shows that typically an ancestral region in the *C. canephora* genome can be traced to four regions in *G. dahurica*. Grey bands in the background indicate syntenic blocks between the genomes spanning more than 15 genes; example of the 1:4 blocks are highlighted in red. (D) Evolutionary rate correction. Distribution of corrected Ks values in syntenic blocks and age estimates for the WGD events.

(89.81%) were clustered into 17,175 gene families, which included 9,728 (56.64%) gene families shared by all five species and 4,640 (27.02%) *G. dahurica* specific families (Fig. 2B). Gene ontology (GO) term enrichment analysis (P-value < 0.01) revealed that those *G. dahurica* unique genes were involved in the negative regulation of transcription, cellular macromolecule metabolic process, amino acids biosynthetic process, flavonoid metabolic process, and so on (Supplementary Table S9).

By conducting gene family expansion and contraction analysis, we discovered 4,181 expanded and 3,431 contracted gene families in *G. dahurica* relative to *G. sempervirens*. Among them, 104 and 11 gene families were significantly (P-value < 0.01) expanded and contracted respectively (Fig. 2A). The significantly expanded gene families derived mainly from the tandemly duplicated genes (24.92%) (Supplementary Table S10). The GO enrichment analysis of these expanded genes revealed that they were mainly enriched in monoterpenoid and (-)-secologanin biosynthetic process and flavonoid metabolic process (Supplementary Figs S7 and S8, Supplementary Tables S11 and S12). Therefore, expansions of these gene families may play important roles in gentiopicroside biosynthesis in *G. dahurica*.

3.4. Key genes involving in gentiopicroside biosynthesis

We reconstructed the putative gentiopicroside biosynthetic pathway based on the KEGG database and previously published results.[40,74] Homologous alignment and a Pfam database searching were conducted subsequently. A total of 135 genes were classified into 33 enzyme categories related to four metabolic pathways leading to the likely gentiopicroside biosynthesis[75] (Fig. 3A, Supplementary Table S14). Previous studies found that secologanin synthase[76] (SLS/CYP72A219) and 7-deoxyloganic acid 7-hydroxylase[77] (7-DLH/CYP72A224) involved in MIA biosynthesis in *C. roseus*, always show high amino acid residue sequence identity.[78] We therefore constructed phylogenetic trees to explore potential genes encoding 7-DLH and SLS in *G. dahurica*. All candidate genes that may encode these enzymes in

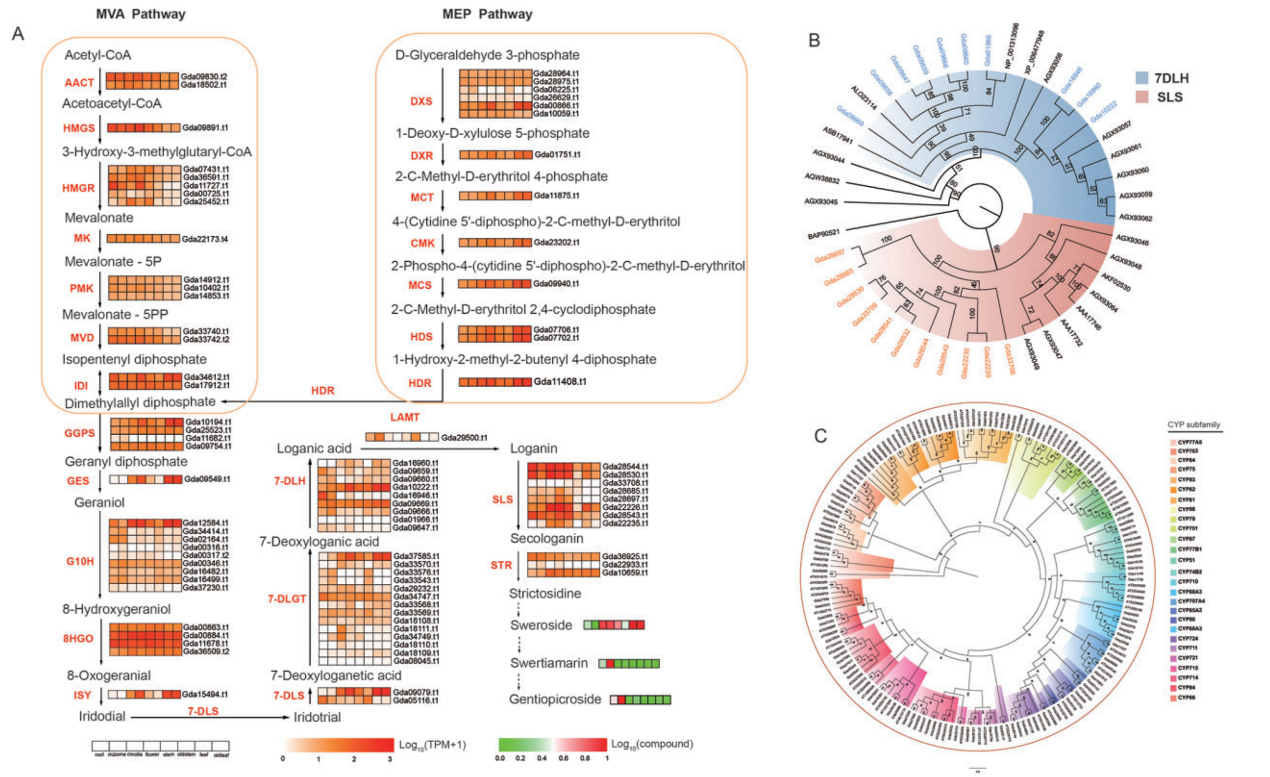

Figure 3. Genes involved in gentiopicroside biosynthesis. (A) A simplified representation of the gentiopicroside biosynthetic pathway. Top hits for pathway genes identified by blast and the expression value for each gene is indicated in color on a $\log_{10}$ (TPM + 1) (transcripts per million) scale for eight tissues: root, rhizome, corolla, flower, stem, old stem, leaf, and old leaf. (B) A phylogenetic tree of all candidate genes in the *CYP72A* subfamily. The sequences shown in figure indicate previously published *7-DLH* and *SLS* genes, and candidate genes identified in the *G. dahurica* genome. (C) Phylogenetic tree of the candidate CYP450 enzyme encoding genes identified in the *G. dahurica* genome. Different colors indicate different CYP450 family members classified based on the protein domain annotation, homolog searching by *A. thaliana* and SwissProt protein database. Families involved in gentiopicroside biosynthesis were marked by arcs.

the subfamily CYP72A were retrieved from two previously published SLS and 7-DLH sequences in *C. roseus* and *C. acuminata*.[8,9] Our phylogenetic trees showed that 11 genes clustered with the previously reported *SLS genes*, while the other 10 clustered with the previously reported *7-DLH genes* with high statistical support values (Fig. 3B).

In order to identify candidate genes for the gentiopicroside biosynthesis in the later stages, we screened candidate *CYP450* genes in *G. dahurica* by homology searching and structural domain alignment, and constructed their phylogenetic relationships (Fig. 3C) according to previous studies.[79,80] We then clustered them into families and subfamilies based on domain annotation, functional descriptions of the homologs in *A. thaliana* and the SwissProt database. These genes may be involved in gentiopicroside biosynthesis, although further verification of the functions of these proteins are needed to be determined. We therefore completed the set of candidate genes for gentiopicroside biosynthesis based on our chromosome-scale genome assembly and gene annotation. In total we identified 135 candidate genes involving in gentiopicroside biosynthesis. We further found that the detected WGD event shared by *G. dahurica* and *G. sempervirens* and the other only for *G. dahurica* contributed to 26 and 28 genes of these candidates identified here for gentiopicroside biosynthesis.

3.5. Gene expressions and metabolome analysis of gentiopicroside biosynthesis

We performed detailed transcriptome analyses of *G. dahurica* based on our high-quality reference genome. We calculated the gene expression level of each gene and carried out analyses of differential gene expressions. We found that 21,827 genes were expressed significantly different between tissues. We then constructed weighted gene co-expression networks with WGCNA based on the differentially expressed genes. The results revealed that 25 co-expression network modules comprised 940 differentially expressed TFs and 80 genes related to the gentiopicroside biosynthesis pathway (Figs 3 and 4 and Supplementary Figs S8 and S9). We found that modules 1 and 2 were mainly related to the seroside biosynthesis, which showed the high expressions in the leaf, while modules 21 to the secologanin and gentiopicroside biosynthesis pathways contained the vast identification genes. In modules 1 and 2, genes from the MEP pathway and those *SLS* and *LAMT* genes were included. In addition, genes related to the MVA pathway and some *SLS* genes were grouped in the module 9, which showed the high concentration in roots. Therefore, we suspected that the upstream pathway of the gentiopicroside biosynthesis might occur in the leaves and the produced intermediates were then transferred to the other tissues and continued the downstream biosynthesis. In addition, we found that all candidate genes derived from two WGD events for *G. dahurica* were also included these modules.

We determined concentrations of the four compounds among the gentiopicroside biosynthesis pathway, loganin, sweroside, swertimarin, and gentiopicroside (Fig. 4). These metabolites had two contrasted accumulation level in contents. Gentiopicroside, swertimarin, and loganin, had the highest content in rhizomes but lowest in leaf, while sweroside, highest in leaf but lowest in rhizomes. This was similar to the clustered

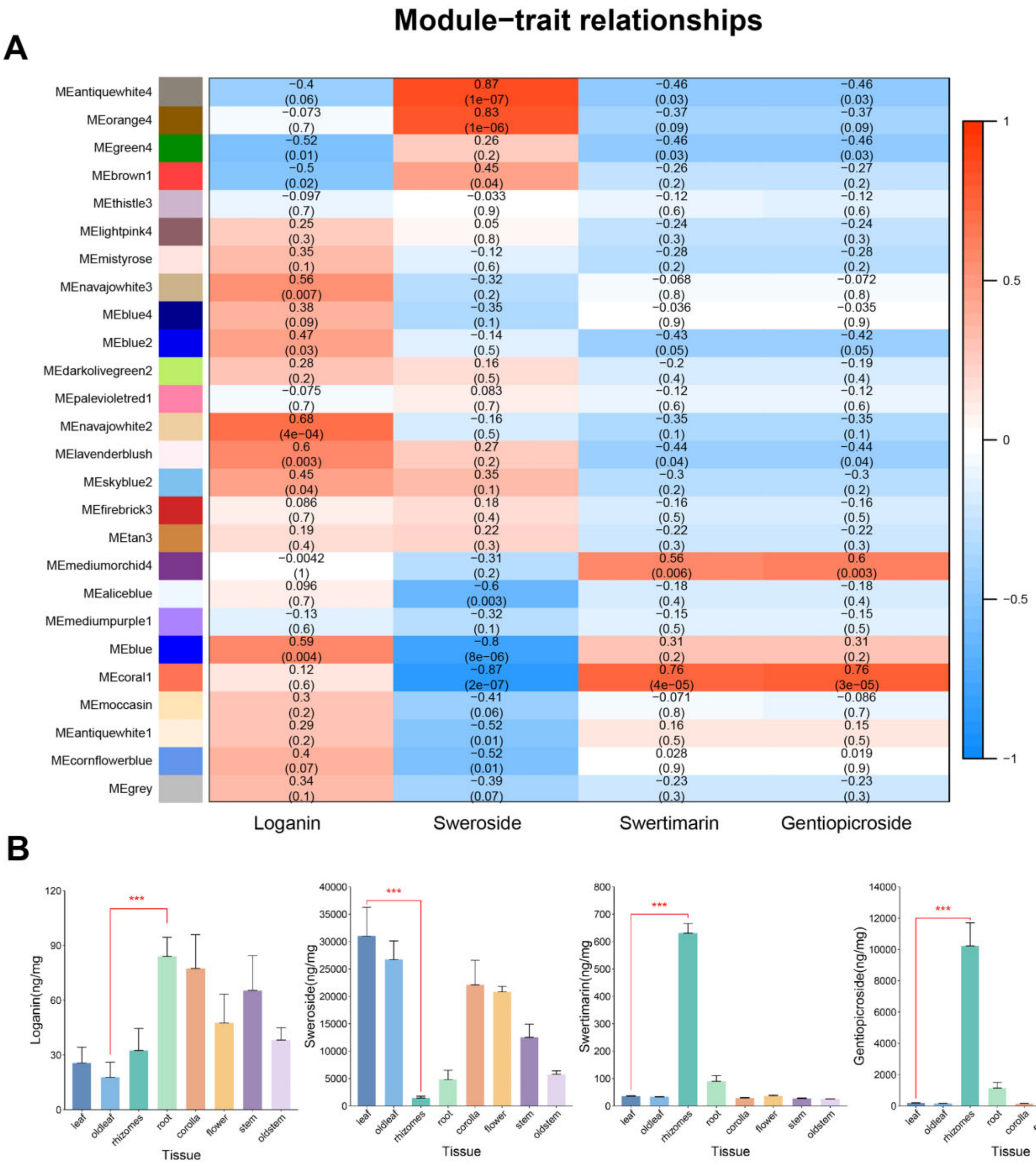

Figure 4. Co-expression gene clusters of gentiopicroside biosynthesis and concentrations of the related chemicals. (A) Heatmap of correlation matrix with 25 gene modules in WGCNA and contents of four major chemical concentrations (Loganin, Sweroside, Swertimarin, and gentiopicroside) in different tissues of *G. dahurica.* (B) Contents of four major chemical concentrations (Loganin, Sweroside, Swertimarin and gentiopicroside) in different tissues of *G. dahurica* determined based on HPLC. Values are means ± SE. The *** within each medicinal component type between different tissues indicates *P*-value <0.01.

gene modules and their tissue expressions, which indicates that gentiopicrosides may be mainly synthesized in the rhizomes although the synthesis of some intermediate chemicals likely occurs in the other tissues.

4. Discussion

In this study, we assembled a chromosome-level genome for one medicinal plant *G. dahurica.* The genome length is 1416.54 Mb, with 97.29% (1378.19 Mb) assigned to 13 pseudochromosomes. The assembled genome has a high quality and completeness, with total complete BUSCOs of 96.5%. Moreover, 37,988 genes were predicted in total and 97.57% of them were annotated on the pseudochromosomes. These results show the high accuracy of genome assembly and gene annotation. This is the first chromosome-level reference genome for the Gentianaceae. Such a genome sequence will accelerate evolutionary, ecological, and pharmaceutical studies of this family in the future.

Base on the selected representative species with genome available, we found that Gentianales and the closely related Lamiales diverged very early. In addition, within Gentianales, Gentianaceae was dated to diverge from Gelsemiaceae approximately 47.03 Mya. These two families shared one WGD event that occurred 42.78–53.14 Mya. However, we further identified one species-specific WGD event 30.39–37.75 Mya in *G. dahurica*. We did not identify the additional WGD event in the family Apocynaceae of Gentianales. However, we recovered an independent but more ancient WGD for the distantly related *Camptotheca* in the family Nyssaceae.[8] All of these WGDs concur basically with phylogenetic relationships and diversification orders of these groups.

TIAs occur in all of these families in Gentianales but the final productions are different, for example, camptothecin in *Camptotheca*[8] while gentiopicroside for *Gentiana*.[4–6] The genes from independent WGD events were repeatedly to be reported to be related to biosynthesis of these specific chemicals.[8] Through the annotated genome sequence and the likely gentiopicroside biosynthetic pathway based on the KEGG database, we identified 135 candidate genes related to gentiopicroside synthesis. Our transcriptome analyses and measurements of these intermediate chemicals also suggest that these candidate genes have been involved in gentiopicroside synthesis. Similarly we found that two WGDs, one specific to *G. dahurica* and the other shared with the closely related *G. sempervirens* contributed greatly (40%) to the identified candidate genes for gentiopicroside synthesis (Supplementary Tables S15–S17). It is highly likely that evolutionary synthesis of gentiopicroside derived from the step-by-step origin of the related genes through two consecutive WGDs that also leaded to phylogenetic diversification of *G. dahurica* and related species. In addition, the high concentration of gentiopicroside in the rhizomes suggests that this chemical is finally synthesized there (Fig. 4B). However, some intermediate chemicals, for example, sweroside, seem to be synthesized in the leaf where the concentration is the highest. The expressions of the related genes also are consistent with the concentrations of these chemicals (Fig. 3). Therefore, the intermediate chemicals may be transferred to rhizomes from other tissues. The underlying molecular mechanism needs further investigation. Overall, our genome sequence data and preliminary analyses advance our understanding of the origin of the gentiopicroside synthesis in the family Gentianaceae.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (32171606, 41771055, and 31700323) and the National Key Research and Development program (2017YFC0505203).

Accession numbers

NCBI under the BioProject: PRJNA799480.

Conflict of interest

None declared.

Data availability

Raw sequencing data and genome assembly have been deposited at the NCBI under the BioProject: PRJNA799480.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Zhang, X-L., Wang, Y.-J., Ge, X.-J., Yuan, Y.-M., Yang, H. and Liu, J. 2009, Molecular phylogeny and biogeography of *Gentiana sect. Cruciata* (Gentianaceae) based on four chloroplast DNA datasets, *Taxon*, **58**, 862–70.
2. Ho, T-N. and Liu, S. 2001, *A Worldwide Monograph of Gentiana*. Science Press: Beijing.
3. Meng, J., Chen, X.F., Song, J.H., et al. 2013, Research progress in classification and identification of *Sect. Cruciata* Gaudin in *Gentiana* (Tourn.) L. *Chin. Tradit. Herb. Drugs*, **44**, 2330–5.
4. Hua, W-P., Zheng, P., He, Y., Cui, L., Kong, W. and Wang, Z-Z. 2014, An insight into the genes involved in secoiridoid biosynthesis in *Gentiana macrophylla* by RNA-seq, *Mol. Biol. Rep.*, **41**, 4817–25.
5. Zhou, D., Gao, S., Wang, H., et al. 2016, De novo sequencing transcriptome of endemic *Gentiana straminea* (Gentianaceae) to identify genes involved in the biosynthesis of active ingredients, *Gene*, **575**, 160–70.
6. Zhang, X., Allan, A.C., Li, C.-H., Wang, Y-Z. and Yao, Q. 2015, De novo assembly and characterization of the transcriptome of the Chinese Medicinal Herb, *Gentiana rigescens*, *Int. J. Mol. Sci.*, **16**, 11550–73.
7. Geu-Flores, F., Sherden, N.H., Courdavault, V., et al. 2012, An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis, *Nature*, **492**, 138–42.
8. Kang, M., Fu, R., Zhang, P., et al. 2021, A chromosome-level *Camptotheca acuminata* genome assembly provides insights into the evolutionary origin of camptothecin biosynthesis, *Nat. Commun.*, **12**, 1–12.
9. Salim, V., Wiens, B., Masada-Atsumi, S., Yu, F. and De Luca, V. 2014, 7-deoxyloganetic acid synthase catalyzes a key 3 step oxidation to form 7-deoxyloganetic acid in *Catharanthus roseus* iridoid biosynthesis, *Phytochemistry*, **101**, 23–31.
10. Zhan, G., Miao, R., Zhang, F., et al. 2020, Monoterpene indole alkaloids with diverse skeletons from the stems of *Rauvolfia vomitoria* and their acetylcholinesterase inhibitory activities, *Phytochemistry*, **177**, 112450.
11. Vranová, E., Coman, D. and Gruissem, W. 2013, Network analysis of the MVA and MEP pathways for isoprenoid synthesis, *Annu. Rev. Plant Biol.*, **64**, 665–700.
12. Denoeud, F., Carretero-Paulet, L., Dereeper, A., et al. 2014, The coffee genome provides insight into the convergent evolution of caffeine biosynthesis, *Science*, **345**, 1181–4.
13. Yang, Z., Liu, G., Zhang, G., et al. 2021, The chromosome scale high-quality genome assembly of *Panax notoginseng* provides insight into dencichine biosynthesis, *Plant Biotechnol. J.*, **19**, 869–71.
14. Seemann, M., Tse Sum Bui, B., Wolff, M., Miginiac-Maslow, M. and Rohmer, M. 2006, Isoprenoid biosynthesis in plant chloroplasts via the MEP pathway: direct thylakoid/ferredoxin-dependent photoreduction of GcpE/IspG, *FEBS Lett.*, **580**, 1547–52.
15. Sun, P., Song, S., Zhou, L., Zhang, B., Qi, J. and Li, X. 2012, Transcriptome analysis reveals putative genes involved in iridoid biosynthesis in *rehmannia glutinosa*, *Int. J. Mol. Sci.*, **13**, 13748–63.
16. Guo, K., Liu, X., Zhou, T., et al. 2020, Gentianelloids A and B, immunosuppressive 10,11-seco-gentianellane sesterterpenoids from the traditional uighur medicine *Gentianella turkestanorum*, *J. Org. Chem.*, **85**, 5511–5.
17. Patel, R.K. and Jain, M. 2012, NGS QC toolkit: a toolkit for quality control of next generation sequencing data, *PLoS One*, **7**, e30619.
18. van Berkum, N.L., Lieberman-Aiden, E., Williams, L., et al. 2010, Hi-C: a method to study the three-dimensional architecture of genomes, *J. Vis. Exp.*, **39**, 1869.
19. Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. 2020, GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes, *Nat. Commun.*, **11**, 1432.

20. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
21. Sun, H., Ding, J., Piednoël, M. and Schneeberger, K. 2018, findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies, *Bioinformatics*, **34**, 550–7.
22. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. 2017, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Res.*, **27**, 722–36.
23. Walker, B.J., Abeel, T., Shea, T.P., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.
24. Durand, N.C., Shamim, M.S., Machol, I., et al. 2016, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments, *Cell Syst.*, **3**, 95–8.
25. Durand, N.C., Robinson, J.T., Shamim, M.S., et al. 2016, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom, *Cell Syst.*, **3**, 99–101.
26. Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds, *Science*, **356**, 92–5.
27. Dudchenko, O., Shamim, M.S., Batra, S.S., et al. 2018, The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000, *bioRxiv*, 254797.
28. Tarailo-Graovac, M. and Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **25**, 1–14.
29. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**, i351–8.
30. Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinformatics*, **9**, 18.
31. Xu, Z. and Wang, H. 2007, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res.*, **35**, W265–8.
32. Ou, S. and Jiang, N. 2018, LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, *Plant Physiol.*, **176**, 1410–1422.
33. Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., et al. 2010, The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*, *Science*, **327**, 92–4.
34. Yang, Y., Sun, P., Lv, L., et al. 2020, Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution, *Nat. Plants.*, **6**, 215–22.
35. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
36. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc.*, **8**, 1494–512.
37. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.*, **31**, 5654–66.
38. Kaul, S., Koo, H.L., Jenkins, J., et al. 2000, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
39. Kim, S., Park, J., Yeom, S.-I., et al. 2017, New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication, *Genome Biol.*, **18**, 1–11.
40. Miettinen, K., Dong, L., Navrot, N., et al. 2014, The seco-iridoid pathway from *Catharanthus roseus*, *Nat. Commun.*, **5**, 3606–16.
41. Iorizzo, M., Ellison, S.L., Senalik, D.A., et al. 2016, A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution, *Nat. Genet.*, **48**, 657–66.
42. Hoopes, G.M., Hamilton, J.P., Kim, J., et al. 2018, Genome assembly and annotation of the medicinal plant *Calotropis gigantea*, a producer of anticancer and antimalarial Cardenolides, *G3 (Bethesda)*, **8**, 385–91.
43. Unver, T., Wu, Z., Sterck, L., et al. 2017, Genome of wild olive and the evolution of oil biosynthesis, *Proc. Natl. Acad. Sci. USA*, **114**, E9413–22.
44. Xiao, L., Yang, G., Zhang, L., et al. 2015, The resurrection genome of *Boea hygrometrica*: a blueprint for survival of dehydration, *Proc. Natl. Acad. Sci. USA*, **112**, 5833–37.
45. Franke, J., Kim, J., Hamilton, J.P., et al. 2019, Gene discovery in gelsemium highlights conserved gene clusters in monoterpene indole alkaloid biosynthesis, *ChemBioChem*, **20**, 83–7.
46. Yoshida, S., Kim, S., Wafula, E.K., et al. 2019, Genome sequence of *Striga asiatica* provides insight into the evolution of plant parasitism, *Curr. Biol.*, **29**, 3041–52.e4.
47. Dong, A., Xin, H-B., Li, Z., et al. 2018, High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant, *GigaScience*, **7**,
48. Slater, G.S.C. and Birney, E. 2005, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics.*, **6**, 31.
49. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004, AUGUSTUS: a web server for gene finding in eukaryotes, *Nucleic Acids Res.*, **32**, W309–312.
50. Majoros, W.H., Pertea, M. and Salzberg, S.L. 2004, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878–79.
51. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biol.*, **9**, R7.
52. Bairoch, A. and Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.*, **28**, 45–8.
53. Zdobnov, E.M. and Apweiler, R. 2001, InterProScan—an integration platform for the signature-recognition methods in InterPro, *Bioinformatics*, **17**, 847–8.
54. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–76.
55. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. 2012, KEGG for integration and interpretation of large-scale molecular data sets, *Nucleic Acids Res.*, **40**, D109–14.
56. Chen, L., Li, M., Yang, Z-Q., et al. 2020, *Gardenia jasminoid*es Ellis: ethnopharmacology, phytochemistry, and pharmacological and industrial applications of an important traditional Chinese medicine, *J. Ethnopharmacol.*, **257**, 112829.
57. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.*, **20**.
58. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772–80.
59. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–13.
60. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
61. Hedges, S.B., Dudley, J.T. and Kumar, S. 2006, TimeTree: a public knowledge-base of divergence times among organisms, *Bioinformatics*, **22**, 2971–72.
62. Bie, T.D., Cristianini, N., Demuth, J.P. and Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–1271.
63. Jaillon, O., Aury, J.M., Noel, B., et al. 2007, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, *Nature*, **449**, 463–7.
64. Sun, P., Jiao, B., Yang, Y., et al. 2021, WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes, *bioRxiv*, https://doi.org/10.1101/2021.04.29.441969.

65. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. and Kingsford, C. 2017, Salmon provides fast and bias-aware quantification of transcript expression, *Nat. Methods*, **14**, 417–19.
66. Leek, J.T., Johnson, W., Parker, H.S., Jaffe, A. and Storey, J.D. 2012, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, *Bioinformatics*, **28**, 882–3.
67. Love, M.I., Huber, W. and Anders, S. 2014, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.*, **15**, 1–21.
68. Haynes, W., 2013, Benjamini–Hochberg method. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H. and Yokota, H. (eds), *Encyclopedia of Systems Biology*, Springer, New York, New York, NY, p. 78.
69. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P. 2007, GSEA-P: a desktop application for Gene Set Enrichment Analysis, *Bioinformatics*, **23** 23, 3251–53.
70. Langfelder, P. and Horvath, S. 2008, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics*, **9**, 559.
71. Yang, J., Song, Y., Liu, Y., et al. 2021, UHPLC-QQQ-MS/MS assay for the quantification of dianthrones as potential toxic markers of Polygonum multiflorum Thunb: applications for the standardization of traditional Chinese medicines (TCMs) with endogenous toxicity. *Chin. Med.*, **16**, 51.
72. Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. 2012, Hi-C: a comprehensive technique to capture the conformation of genomes, *Methods*, **58**, 268–76.
73. van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L. and Vandepoele, K. 2009, The flowering world: a tale of duplications, *Trends Plant Sci.*, **14**, 680–8.
74. Guo, K., Liu, Y.-C., Liu, Y., et al. 2021, Immunosuppressive gentianellane-type sesterterpenoids from the traditional Uighur medicine *Gentianella turkestanorum*, *Phytochemistry*, **187**, 112780.
75. Thamm, A.M., Qu, Y. and De Luca, V. 2016, Discovery and metabolic engineering of iridoid/secoiridoid and monoterpenoid indole alkaloid biosynthesis, *Phytochem. Rev.*, **15**, 339–61.
76. Guo, K., Zhou, T., Ren, X., et al. 2021, Secoiridoids and triterpenoids from the traditional Tibetan medicine *Gentiana veitchiorum* and their immunosuppressive activity, *Phytochemistry*, **192**, 112961.
77. Asada, K., Salim, V., Masada-Atsumi, S., et al. 2013, A 7-deoxyloganetic acid glucosyltransferase contributes a key step in secologanin biosynthesis in *Madagascar Periwinkle*[C][W][OPEN], *Plant Cell*, **25**, 4123–34.
78. Sadre, R., Magallanes-Lundback, M., Pradhan, S., et al. 2016, Metabolite diversity in alkaloid biosynthesis: a multilane (diastereomer) highway for camptothecin synthesis in *Camptotheca acuminata*, *Plant Cell.*, **28**, 110.
79. Sun, W., Ma, Z. and Liu, M. 2020, Cytochrome P450 family: genome-wide identification provides insights into the rutin synthesis pathway in *Tartary buckwheat* and the improvement of agricultural product quality, *Int. J. Biol. Macromol.*, **164**, 4032–45.
80. Zhang, D., Li, W., Xia, E.-H., et al. 2017, The Medicinal Herb *Panax notoginseng* genome provides insights into ginsenoside biosynthesis and genome evolution, *Mol. Plant.*, **10**, 903–7.