DNA Research, 2022, **29**, 1–15
https://doi.org/10.1093/dnares/dsac034
Advance access publication 5 October 2022
Resource Article: Genomes Explored

Resource Article: Genomes Explored

De novo genome assembly of the medicinal plant *Gentiana macrophylla* provides insights into the genomic evolution and biosynthesis of iridoids

Tao Zhou[1,*,†], Guoqing Bai[2,3,†], Yiheng Hu[4,†], Markus Ruhsam[5], Yanci Yang[6], and Yuemei Zhao[7,*]

[1]School of Pharmacy, Xi'an Jiaotong University, Xi'an 710061, China
[2]Shaanxi Engineering Research Centre for Conservation and Utilization of Botanical Resources, Xi'an Botanical Garden of Shaanxi Province (Institute of Botany of Shaanxi Province), Xi'an 710061, China
[3]Shaanxi Eins Ecological Science & Technology Co. Ltd., Xi'an, China
[4]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences, Beijing, China
[5]Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK
[6]School of Biological Science and Technology, Baotou Teachers' College, Baotou, China
[7]School of Biological Sciences, Guizhou Education University, Guiyang, China

*Corresponding author: Tel. +86 29 8265 5424. Email: zhoutao196@mail.xjtu.edu.cn (T.Z.); yezi19820320@163.com (Y.Z.)
†These authors contributed equally to this work.

Abstract

Gentiana macrophylla is a perennial herb in the Gentianaceae family, whose dried roots are used in traditional Chinese medicine. Here, we assembled a chromosome-level genome of *G. macrophylla* using a combination of Nanopore, Illumina, and Hi-C scaffolding approaches. The final genome size was ~1.79 Gb (contig N50 = 720.804 kb), and 98.89% of the genome sequences were anchored on 13 pseudochromosomes (scaffold N50 = 122.73 Mb). The genome contained 55,337 protein-coding genes, and 73.47% of the assemblies were repetitive sequences. Genome evolution analysis indicated that *G. macrophylla* underwent two rounds of whole-genome duplication after the core eudicot γ genome triplication event. We further identified candidate genes related to the biosynthesis of iridoids, and the corresponding gene families mostly expanded in *G. macrophylla*. In addition, we found that root-specific genes are enriched in pathways involved in defense responses, which may greatly improve the biological adaptability of *G. macrophylla*. Phylogenomic analyses showed a sister relationship of asterids and rosids, and all Gentianales species formed a monophyletic group. Our study contributes to the understanding of genome evolution and active component biosynthesis in *G. macrophylla* and provides important genomic resource for the genetic improvement and breeding of *G. macrophylla*.

Key words: medicinal plant, Gentiana macrophylla, genome assembly, whole-genome duplication, iridoid biosynthesis

1. Introduction

The genus *Gentiana* contains about 360 species which are widely distributed throughout the northern hemisphere.[1] *Gentiana* species have attracted interest due to their important medical, chemical, ecological, and horticultural properties.[2,3] Among them, *Gentiana macrophylla* ($2n = 2x = 26$) is a well-known medicinal plant which was first recorded in the Chinese Medicine monograph 'Divine Husbandman's Classic of the Materia Medica' (Shen Nong Ben Cao Jing) of the latter Han dynasty. The dried roots of *G. macrophylla* are used as a traditional Chinese medicine called 'Qinjiao' which is one of the famous 'Qin medicines' in Shaanxi province of northern China. In Traditional Chinese Medicine, Qinjiao is used to dispel rheumatism, to relieve arthralgia and deficiency-heat, to clear dampness-heat and to activate Qi (energy).[4] Therefore, the dried roots of *G. macrophylla* are commonly used to treat diabetes, apoplexy, paralysis, and rheumatism.[5–7]

Phytochemical studies have indicated that *G. macrophylla* contains various compounds such as iridoids, triterpenes, flavonoids, sterols, and alkaloids.[8] Of these compounds, derivates of iridoids and triterpenes are the predominant bioactive ingredients which are responsible for the biological activities of Qinjiao. For instance, gentiopicroside and loganic acid are two secoiridoid glucosides in *G. macrophylla* with powerful anti-inflammatory and osteoprotective effects.[4,9,10] Due to the medicinal properties, there is considerable commercial interest in *G. macrophylla*, however, the distribution and number of wild *G. macrophylla* populations is very limited. Overexploitation driven by economic interests and the lack of effective protective measures has resulted in a dramatic decline of wild *G. macrophylla* plants in recent years and *G. macrophylla* is now listed on the National Key of Protected Wild Herbs in China.[6,11] To date, no fully sequenced genome is available for *G. macrophylla*, which largely hinders conservation and cultivation efforts. Therefore, it is important

Received 28 April 2022; Revised 24 August 2022; Accepted 12 September 2022

© 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

to obtain the whole genome of *G. macrophylla* in order to facilitate molecular breeding and germplasm conservation programmes as well as scientific research of this and other *Gentiana* species.

Results from flow cytometry studies suggest that *Gentiana* species usually have relatively large genomes.[12–15] The genome size variation of plant species is related to the chromosome number changes caused by whole-genome duplication (WGD), hybridization, and chromosome loss.[16] It has been shown that adaptive radiation as well as introgression and polyploidization events have contributed considerably to the species richness in the genus *Gentiana*.[17–19] Therefore, WGD and/or hybridization had a profound impact on the evolution of *Gentiana* species facilitating diversifications of phenotypes and secondary metabolites via gene divergence and structural variants.[20–22] High-quality genomes enable comparative analyses of the genome architecture and the evolution of key traits for plant species.[23] Therefore, the complete genome of *G. macrophylla* can help to unravel the role of WGD events in *Gentiana* and facilitate the discovery of functional genes.

Large genome sizes have presented geneticists with many challenges regarding sequencing and bioinformatics, making it difficult to produce high-quality reference genomes. However, the newly developed long-read sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore combined with the high-throughput chromatin conformation capture (Hi-C) scaffolding strategy have greatly enhanced chromosomal level assemblies for plant species.[24] Recently, several plant genomes have been sequenced, facilitating the study of the genetic regulation mechanisms of various biosynthetic pathways.[25–29] Here, we assembled a chromosome-level genome of *G. macrophylla* which consists of 13 chromosomes (scaffold N50 =122.73 Mb, totalling ~1.79 Gb) using Oxford Nanopore, Illumina, and Hi-C scaffolding strategies. This will facilitate the study of genetic mechanisms underlying the biosynthesis of bioactive ingredients and provide useful information for further improvement of the genome-assisted cultivation and conservation efforts of *Gentiana*.

2. Materials and methods

2.1. Plant materials, DNA library construction, and sequencing

The *G. macrophylla* plant used in this study was collected from the Liupan mountain, Ningxia autonomy region (35.28°N, 106.66°E, 2,234 m asl), China. Approximately 500 mg of young leaf tissue was stored in liquid nitrogen and used for DNA extraction following a cetyltrimethylammonium bromide (CTAB) protocol.[30] The quantity and quality of the extracted DNA were established with a NanoDrop ND-2000 spectrophotometer (NanoDrop products, Wilmington, DE, USA) and a Qubit 2.0 Fluorometer (Invitrogen Ltd, Paisley, UK). 100 ng of genomic DNA was used to prepare the library for Illumina sequencing. A paired-end DNA library with the insert size of 500 bp was constructed using a NEB Ultra DNA library prep kit (NEB, UK) and then sequenced on an Illumina NovaSeq 6000 platform producing 350 bp reads. After sequencing, duplicate reads, reads with ≥20% low-quality bases, or reads with ≥5% unknown ('N') bases were filtered using fastp v. 0.12.6[31] with default parameters.

2 μg of gDNA was recovered to construct the DNA library for Nanopore sequencing using a NEB Next FFPE DNA Repair Mix kit (M6630, USA) and subsequently processed with the ONT template prep kit (SQK-LSK109, UK). The large fragment library was premixed with loading beads and then pipetted into a R9 flow cell. The library was sequenced on the ONT PromethION platform with Corresponding R9 cell and the ONT sequencing reagent kit (EXP-FLP001.PRO.6, UK). Long raw reads were converted into fastq format using the Guppy tool.[32]

2.2. RNA isolation and sequencing

The genome sequencing sample and other two samples collected from the same sites were used for RNA-seq. Roots, stems, leaves, and opening flowers of these three biological replicates were selected for RNA isolation. Total RNA was isolated using the Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, Hammersmith, UK). RNA purity and concentration were checked using a NanoDrop ND-2000 spectrophotometer (NanoDrop products, Wilmington, DE, USA) and a Qubit 2.0 Fluorometer (Invitrogen Ltd, Paisley, UK), respectively. The isolated RNA was treated with RNase-free DNase I and eluted in RNase-free water. The cDNA library with insert size of 250–350 bp was constructed using the NEBNext® Ultra™ RNA Library Prep Kit (NEB, UK) and then sequenced producing 150 bp paired-end reads on NovaSeq 6000 sequencing platform. The raw reads were trimmed by removing adaptor sequences, reads with more than 5% of unknown base calls (N), and low-quality reads, i.e. >20% of the bases with a quality score ≤10 using fastp v. 0.12.6.[31]

2.3. Hi-C library construction and sequencing

Young leaves from the same plant which was used for the Nanopore sequencing were collected for Hi-C library construction applying the method in Rao *et al*.[33] Briefly, leaf samples were fixed using formaldehyde to produce cross-linked DNA which was subsequently digested with the restriction endonuclease Hind III. Biotin-labelled bases were introduced during the sticky end repairing process to facilitate DNA purification and capture. After repairing, the interacting DNA fragments were cyclized to determine their location during subsequent sequencing and analyses. Finally, DNA fragments were purified and fragmented to a size of 300–700 bp, capturing interacting DNA fragments using streptavidin beads for library construction. The libraries were sequenced on an Illumina NovaSeq 6000 sequencing platform producing paired-end 150 bp reads. The raw reads were trimmed by removing adaptor sequences and low-quality reads using fastp v. 0.12.6[31] with default parameters, and the quality of the Hi-C sequencing data was evaluated with HiC-Pro v2.10.0.[34]

2.4. Genome size estimation

To predict the genome characteristics of *G. macrophylla*, we used a K-mer-based method to evaluate the genome size and the level of heterozygosity.[35] Approximately 177.79 Gb (~131.78×) high-quality Illumina sequencing reads were generated and used for a 21 K-mer analysis. Based on the frequency distribution of the 21 K-mers, GenomeScope was used to count the general characteristics of the genome, including the total genome size, repeat content, and heterozygosity.[36]

2.5. Genome assembly and Hi-C scaffolding

In order to recover high-accuracy reads for assembly, raw Nanopore sequencing reads (8,501,481) were self-corrected using Canu[37] with the following parameters: genomeSize

= 2 G, minReadLength = 2,000, minOverlapLength = 500, -nanopore-corrected. The corrected Nanopore reads were then assembled using SMARTdenovo (https://github.com/ruanjue/smartdenovo) with default parameters apart from the parameter -k which was set to 21. After initial assembly, Racon (https://github.com/lbcb-sci/racon) was used to further correct the assembly result with raw Nanopore reads and default parameters via three iterations. Finally, Pilon v1.23[38] was utilized to fine tune the assembled contigs via three iterations using the 177.79 Gb high-quality Illumina paired sequencing reads. The Pilon parameters were set as follows: -fix bases, -changes, -vcf, -diploid.

Before chromosome assembly, the preassembled contigs were split into fragments with an average length of 50 kb for error correction. Any two segments which showed inconsistent connection with information from the raw scaffold were checked manually. Approximately 154.41 Gb of high-quality Hi-C reads were mapped to these fragments using BWA v 0.7.10.[39] The uniquely mapped Hi-C data were retained to perform a chromosome-level assembly by using LACHESIS[40] with the following parameters: CLUSTER_MIN_RE_SITES = 76; CLUSTER_MAX_LINK_DENSITY = 2; ORDER_MIN_N_RES_IN_TRUNK = 59; ORDER_MIN_N_RES_IN_SHREDS = 63. After this step, placement and orientation exhibiting obvious discrete chromatin interaction patterns were adjusted manually. A Hi-C contact map was drawn to show the scaffold interactions which could be used to evaluate the order and direction of the contigs on the pseudochromosomes. Finally, 4,154 scaffolds (representing 98.89% of the total length) were anchored to 13 pseudochromosomes which were named Chr01 to Chr13 according to the scaffold total length.

2.6. Genome assembly quality assessment

To evaluate the accuracy and completeness of the assemblies, the overall alignment rates were calculated through mapping Illumina clean reads to the assembled sequences using BWA with the default parameters.[39] The RNA sequencing reads from four tissues were also mapped against the genome assembly with HISAT v2.0.5,[41] and the overall alignment rates were estimated. Additionally, CEGMA (Core Eukaryotic Genes Mapping Approach)[42] and BUSCO (Benchmarking Universal Single-Copy Orthologs, embryophyta_odb10 database) analyses[43] were used to assess the quality of the assembly.

2.7. Repetitive elements annotations

Repetitive sequences such as tandem repeats and transposable elements (TEs) are ubiquitous in the genome.[44] In the current study, TEs were identified by a combination of homology-based and *de novo* strategies. Firstly, a *de novo* repeats library was modelled for the *G. macrophylla* genome using RepeatModeler (http://www.repeatmasker.org/RepeatModeler), and RECON v1.08[45] and RepeatScout[46] were used for the *de novo* repeat identification. Full-length long terminal repeat retrotransposons (fl-LTR-RTs) were identified using both LTR harvest[47] with the following parameters: -minlenltr 100 -maxlenltr 40000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes and LTR_finder[48] with the parameters: -D 40000 -d 100 -L 9000 -l 50 -p 20 -C -M 0.9. The high-quality intact fl-LTR-RTs and non-redundant LTR library were then produced using LTR_retriever.[49] Finally, TEs in the *G. macrophylla* genome were identified and classified using RepeatMasker v4.10.[50] Tandem repeats were annotated using Tandem Repeats Finder[51] with the parameters: 1 1 2 80 5 200 2000 -d -h and the MIcroSAtellite identification tool (MISA v2.1).[52]

2.8. Gene prediction and genome annotations

Ab initio prediction, homolog protein-based, and transcriptome-aided methods were integrated to predict genes in the *G. macrophylla* genome. Augustus v2.4[53] and SNAP[54] were used for the *ab initio* prediction. GeMoMa v1.7[55] was used for homology prediction with protein sequences from *Arabidopsis thaliana*, *Coffea canephora*, *Gardenia jasminoides*, *Olea europaea*, and *Solanum tuberosum*. For transcriptome-aided predictions, the *de novo* assembly was firstly conducted to obtain the transcripts for gene predictions. Four RNA libraries generated from root, stem, leaf, and flower tissues were sequenced and *de novo* assembled with Trinity v2.1.1.[56] In addition, genome-guided assembly was also performed to recover transcripts for the transcriptome-aided predictions. The RNA sequencing reads were mapped to the *G. macrophylla* genome assembly using HISAT version 2.0.5,[57] and transcripts were reconstructed using StringTie v5.1.[58] The transcript output derived from *de novo* and genome-guided assemblies were combined to construct comprehensive transcripts for gene prediction using PASA v2.0.2.[59] Finally, all the gene prediction results produced from the three abovementioned methods were combined to generate the final consensus gene models with EVM v1.1.1,[60] and the final gene models were further updated by PASA v2.0.2.[59]

Functional annotation of protein-coding genes was conducted by searching the predicted amino acid sequences of *G. macrophylla* against public databases, including NCBI non-redundant protein (Nr), TrEMBL, Swiss-Prot, euKaryotic Ortholog Group (KOG), and Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) using BLASTP with an *E*-value threshold of 1*E*−5.[61] Protein family (Pfam) alignments were performed using HMMER (v3.0, http://hmmer.org/) with an *E*-value of 1*E*−5, and the Gene Ontology (GO) classification was conducted based on the annotation results of Nr using Blast2GO v2.5 with an *E*-value of 1*E*−5.[62]

The tRNA genes were predicted using tRNAscan-SE,[63] and the rRNA genes were identified by searching the genes in Rfam v 12.0[64] with barrnap v 0.9 (https://github.com/tseemann/barrnap). The snoRNA and snRNA genes were predicted by searching the genes against the Rfam database using Infernal 1.1.[65] The miRNA was identified by searching in the miRbase databases.[66]

2.9. Gene family construction and phylogeny

Protein and nucleotide sequences from *G. macrophylla* and 15 other angiosperm species (*Solanum lycopersicum*, *Aquilegia coerulea*, *Glycine max*, *Theobroma cacao*, *O. europaea*, *C. canephora*, *Ophiorrhiza pumila*, *Camptotheca acuminata*, *Helianthus annuus*, *Oryza sativa*, *G. jasminoides*, *Populus trichocarpa*, *A. thaliana*, *Vitis vinifera*, and *Gossypium raimondii*) were recovered to identify the orthologous groups and construct gene families using OrthoFinder v2.4[67] based on diamond all-to-all BLASTP alignment with an *E*-value cutoff of 1*E*−3. The obtained gene families were annotated using PANTHER V15[68] database, and GO and KEGG enrichment were conducted for the specific gene family of *G.*

macrophylla using clusterProfile v3.14.0.[69] The single-copy orthologous genes were extracted from the orthologous clustering results to reconstruct the phylogenetic relationships. Protein sequences of abovementioned 16 species were concatenated and aligned using MAFFT v7.453 with default parameters.[70] The ambiguous and most variable sites in the multiple protein sequence alignments were removed with Gblocks v0.91b.[71] Afterwards, the concatenated protein matrices were used to construct phylogenetic trees with the maximum-likelihood (ML) method. The ML analysis was conducted using IQ-TREE[72] under the best-fit model (GTR + F + I + G4) selected by ModelFinder[73] with 1,000 bootstrap replicates.

The best ML tree was retrieved as a starting tree to estimate the divergence time of each species using the MCMC tree in PAML v. 4.[74] Tree nodes were calibrated by the TimeTree web service (http://www.timetree.org/), including the split between *C. canephora* and *G. jasminoides* (15.8–27.5 million yrs ago, Ma), *T. cacao* and *G. raimondii* (30–60 Ma), and the split between *O. sativa* and other angiosperms (0–112.6 Ma). According to the results of the gene families and phylogenetic tree with divergence time, CAFÉ v4.2 software[75] was utilized to determine the expanded and contracted gene families of 16 species, which were further subjected to KEGG and GO enrichment analysis.

2.10. WGD and genome synteny analyses

In order to explore the origin of predicted protein-coding genes (55,337) in *G. macrophylla*, genome sequences of *G. macrophylla* were used to infer WGD events based on a synonymous substitution rate (Ks) estimation and a 4-fold synonymous degenerative third-codon transversion (4DTv) method. In addition, protein sequences of *G. jasminoides*, *V. vinifera*, and *C. canephora* were also recovered for analysing the WGD events. Protein sequences for each of these abovementioned species were compared using the all-to-all blastp strategy in BLAST with an *E*-value threshold of 1*E*–5 to predict the conserved paralogs in each species. After then, Ks values for gene pairs were calculated with PAML v. 4,[74] and potential WGD events in each genome were highlighted based on their Ks and 4DTv distribution. It is thought that the only well-established whole-genome triplication occurred in *V. vinifera* and *C. canephora*.[76,77] We therefore conducted the synteny searches with MCscanX[78] using the protein-coding genes of *G. macrophylla*, *V. vinifera*, and *C. canephora*. The derived dot plots were examined to predict the paleoploidy level of *G. macrophylla* compared with that of the other angiosperms by determining the syntenic depth in each genomic region. In addition, genomes of two iridoid-producing plants *O. pumila* and *C. acuminata* were also recovered in the synteny analyses as previous studies have indicated that *O. pumila* did not show signs of WGD and only one recent WGD was found for *C. acuminata*.[79,80]

2.11. Terpene synthase and iridoid-related genes identification and expression analyses

Terpene synthase (TPS) is essential for the biosynthesis of terpenoids. Two Pfam domain models (PF03936 and PF01397) were retrieved to search the *G. macrophylla* genome to identify the TPS proteins using HMMER v. 3.0 with an *E*-value cutoff of 1*E*–5.[81] Previously annotated TPS genes of *C. canephora*, *P. trichocarpa*, *O. sativa*, and *A. thaliana* were downloaded to perform a comparative TPS gene family analysis together with *G. macrophylla*. Protein sequences of all TPS genes were aligned using MAFFT v7.453 with default parameters[70] and trimmed using trimAL.[82] An ML phylogenetic tree was constructed using IQ-TREE[72] under the best-fit model (JTT + F + I + G4) selected by ModelFinder[73] with a 1,000 bootstrap replicates, with the subfamily TPS-c as outgroup. Besides, it has been well established that iridoids are synthesized from either the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway in the plastid or the mevalonic acid (MVA) pathway in the cytoplasm.[83–85] All key enzyme genes related to the MEP, MVA, and iridoid pathway were searched according to the integrated annotations.

RNA-seq clean reads were produced from the roots, stems, leaves, and flowers of *G. macrophylla*. These reads were mapped to the reference genome by using HISAT 2.[41] The fragments per kilobase of transcript per million of fragments mapped (FPKM) were calculated using eXpress[86] to estimate the gene expression levels of different tissues. DESeq2 was used to analyse the significantly differentially expressed genes (DEGs) with the thresholds $P < 0.05$ and $|\log2(\text{FoldChange})| > 1$, and threshold of *P* value was inferred from the false discovery rate. Afterwards, a *k*-means cluster analysis was conducted to show the similarities and differences of gene expression patterns in different tissues, and gene expression profiles of TPS and iridoid-related genes were plotted using TBtools[87] based on the RNA-seq data from different tissues of *G. macrophylla*.

3. Results

3.1. Sequencing information, genome size estimation, and heterozygosity

Illumina sequencing generated on average a ×99 coverage of paired-end short reads (177.79 Gb), Oxford Nanopore sequencing a ×110 coverage of single-molecule long reads (a total of 198. 87 Gb, with an average length of 23.39 kb and a read N50 length of 29.21 kb), and the Hi-C sequencing produced ~154.42 Gb of data (Supplementary Table S1). Transcriptome sequencing generated 93.64 Gb reads for roots, stems, leaves, and flowers (Supplementary Table S1). K-mer frequency distribution analyses indicated a K-mer peak with a depth of 25 and an estimated genome size of 1.42 Gb (Supplementary Fig. S1), which is less than the 1.78 Gb determined using flow cytometry.[13] Based on the total number of K-mers, the heterozygosity and the repetitive content of *G. macrophylla* were estimated to be 1.49% and 72.4%, respectively (Supplementary Fig. S1). These results indicate that the genome of *G. macrophylla* is large with high levels of heterozygsity.

3.2. Whole-genome assembly and quality evaluation

The genome of *G. macrophylla* was *de novo* assembled using the Oxford Nanopore long reads output incorporating Hi-C data for scaffolding. After the results derived from different assembly strategies were evaluated, we selected the SMARTdenovo results as the final assembly (Supplementary Table S2). The total length of contigs assembled by SMARTdenovo is about 1.791 Gb across 4,024 contigs, with contig N50 size of 720.80 kb (Table 1), which is larger than the estimated genome size but close to the genome size determined by flow cytometry. Employing the Hi-C scaffolding strategy, 516.3 million clean reads from the Illumina sequencing platform were used for chromosome construction

Table 1. Statistics of final genome assembly

	Nanopore assembly (polished)	Hi-C assembly
Assembly		
Total scaffold length (Gb)		1.792
Number of scaffolds		1,264
Scaffold N50 (Mb)		122.726
Scaffold N90 (Mb)		94.709
Max scaffold length (Mb)		195.951
Total contig length (Gb)	1.791	1.792
Scaffold GC content		37.66%
Scaffold N content		0.017%
Number of contigs	4,024	4,301
Contig N50 (kb)	720.804	647.199
Contig N90 (kb)	210.320	203.182
Max contig length (kb)	4,866.657	4,866.657
Contig GC content	37.66%	37.66%
Complete BUSCOs	95.11%	94.73%
Complete and single-copy BUSCOs	69.89%	87.36%
Complete and duplicate BUSCOs	25.22%	7.37%
Fragmented BUSCOs	1.12%	1.36%
CEGs present in assemblies	97.82%	98.20%
Annotation		
Repetitive density	77.40%	
Number of non-coding RNAs	2,669	
Number of protein-coding genes	55,337	

to further refine the *G. macrophylla* genome assembly. Approximately 98.89% of the assembly was anchored to 13 pseudochromosomes with LACHESIS, and the total length of the assembly was 1.792 Gb with a contig N50 of 647.20 kb and scaffold N50 of 122.73 Mb (Table 1). Ultimately, the assembled contig sequences were connected in the determined order and direction by adding 100 N to obtain the final chromosome-level genome sequence with a chromosome mount rate of 92.77% (Supplementary Table S3 and Fig. S2). A Hi-C contact map indicated that the clustering, ordering, and orientation of the contigs were valid, providing the first chromosome-scale genome assembly for *G. macrophylla* (Supplementary Fig. S2).

Different methods have been adopted to evaluate the assembly quality of the genome. The Illumina reads were aligned to the assembled genome of *G. macrophylla* and showed a mapping rate of 99.60% (Supplementary Table S4). RNA sequencing data from four different tissues for genome annotation were also mapped to the assembly to assess the quality of assembly, resulting in 96.12%, 96.10%, 96.53%, and 95.90% of the total mapped RNA-seq reads for roots, stems, leaves, and opening flowers, respectively (Supplementary Table S5). The completeness of our genome assembly was also evaluated by CEGMA and BUSCO assessment. The results indicated that a total of 232 (93.55%) of 248 core eukaryotic genes (CEGs) and 1,535 (94.73%) complete gene models among 1,614 conserved genes from BUSCO assessment were identified (Table 1). DNA/RNA read mapping using CEGMA and BUSCO could map more than 93% of the reads suggesting that the completeness of the reference genome of *G. macrophylla* is good.

3.3. Repeat annotations

Homology-based annotation and a *de novo* approach were utilized to predict TEs and tandem repeats in the *G. macrophylla* genome. The total length of TEs was 1,386,789,613 bp (73.47% of the genome length) in the genome assemblies of *G. macrophylla*. LTR elements were the dominant repeat type, accounting for 68.87% of the genome length (Supplementary Table S6). Two LTR superfamily elements (*Copia* and *Gypsy* elements) constituted 424,457,684 and 415,291,490 bp, accounting for 34.40% and 33.65% of the total LTR repeat length, respectively. DNA transposons (58,429,067 bp) and long interspersed nuclear elements (21,887,473 bp) accounted for 3.26% and 1.22% of the genome assembly (Supplementary Table S6), respectively. The tandem repeats constituted 70,335,884 bp accounting for 3.93% of the genome length.

3.4. Gene annotation and function prediction

Based on the assembled genome, 55,337 protein-coding genes, 1,406 tRNAs, 755 rRNAs, 72 miRNAs, 293 snRNAs, and 143 snoRNA were predicted (Supplementary Table S7). The gene density, GC content, *Gypsy* density, and *Copia* density were mapped onto the individual chromosomes using the Circos tool (http://www.circos.ca) (Fig. 1). The protein-coding genes in the *G. macrophylla* genome had an average gene length of 3,603.34 bp, an average coding DNA sequence length of 1,096.01 bp, and an average exon number per gene of 4.71. The different gene structure parameters were compared with those of five selected species: *C. canephora*, *A. thaliana*, *S. tuberosum*, *G. jasminoides*, and *O. europaea*. Unexpectedly, *G. macrophylla* had the highest numbers of predicted genes and the largest average intron length (~2,243.04 bp) among abovementioned species (Supplementary Table S7), which appears to be related to the relatively large genome size of *G. macrophylla*. The completeness of gene prediction was evaluated using 1,614 BUSCO genes from the embryophyta_odb10 database. The results suggested that 1,567 (97.09%) BUSCOs were complete in the annotated results, indicating the reliability of the predicted results (Supplementary Table S7).

Overall, 51,148 (92.43%) genes were functionally annotated in at least one of the public databases (Supplementary Table S8). A total of 50,528 (91.31%) genes showed homologous genes in the NR database, while 37,975 (68.62%) genes were similar to proteins in the SwissProt database. The BLASTX top-hit species distribution showed highest homology to *Coffea arabica* (25.5%), *C. eugenioides* (13.9%), and *C. canephora* (7.8%) (Supplementary Fig. S3). A total of 41,688 (75.33%) genes contained Pfam domains, and 41,756 (75.46%) genes were assigned to at least one GO term and classified into 42 GO functional subcategories (Supplementary Fig. S4). In order to predict the metabolic pathways of *G. macrophylla*, 38,293 (69.2%) genes were annotated in the KEGG database and classified into 136 pathways. KEGG pathways involved in 'sesquiterpenoid and triterpenoid biosynthesis' (ko00909, 397 genes), 'terpenoid backbone biosynthesis' (ko00900, 97 genes), and 'monoterpenoid biosynthesis' (ko00902, 97 genes) can be used to explore the biosynthetic pathways of medicinally effective compounds in *G. macrophylla* (Supplementary Table S9).

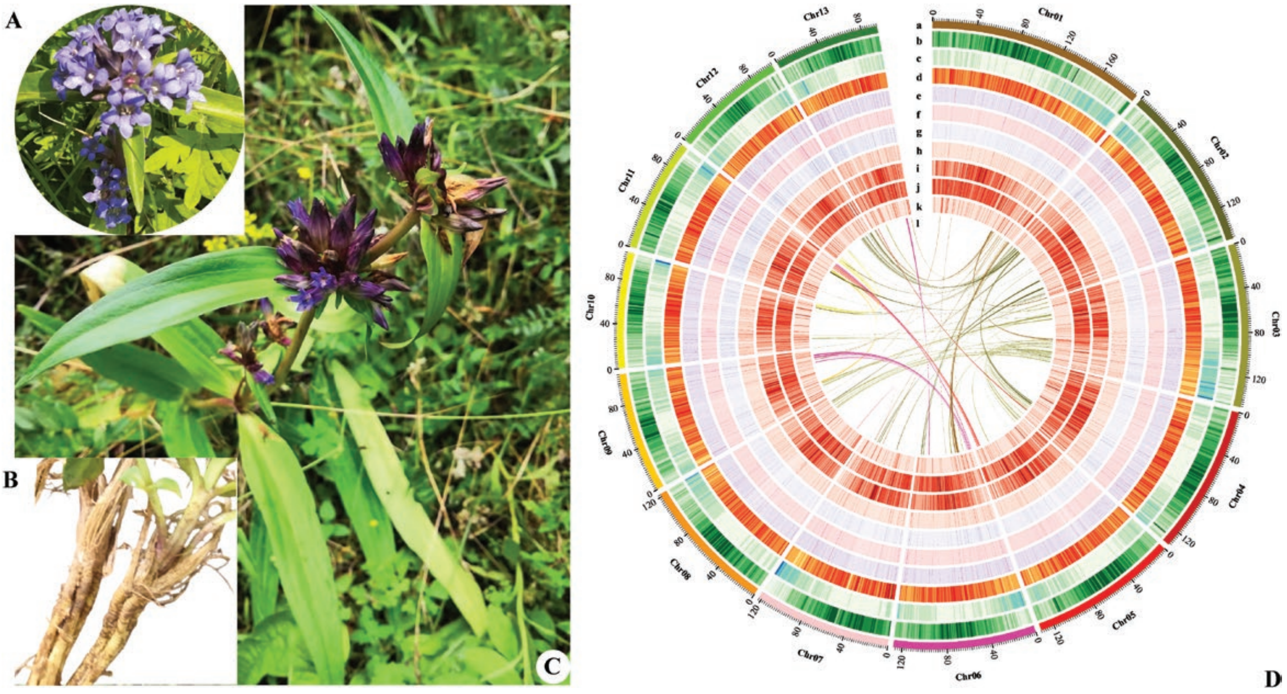

Figure 1. Morphological and genomic characteristics of *Gentiana macrophylla*. (A) *Gentiana macrophylla* in flower. (B) The roots of *G. macrophylla* which are used for medicinal purposes. (C) *Gentiana macrophylla* whole plant. (D) Genome assembly of 13 pseudochromosomes. (a) Assembled pseudochromosomes, (b) TE density, (c) gene density, (d) GC content, (e) gene expression profiles in flower, (f) gene expression profiles in leaf, (g) gene expression profiles in root, (h) gene expression profiles in stem, (i) LTR-Copia density, (j) LTR-Gypsy density, (k) CACTA DNA transposon density, and (l) chromosome synteny (from outside to inside).

3.5. Gene family analysis

In order to explore the genome evolution of *G. macrophylla*, protein-coding genes of other angiosperms were retrieved and clustered into gene families. After cluster analysis of all gene families, a total of 48,374 genes were clustered into 19,134 gene families in the genome of *G. macrophylla*. *Gentiana macrophylla* shared a total of 2,726 gene families with the 15 other species, and 11,778 genes and 2,149 gene families were specific to *G. macrophylla* (Fig. 2A and B). We found that 8,573 gene families were shared among the four Gentianales species, including *G. macrophylla*, *C. canephora*, *G. jasminoides*, *O. pumila*, and iridoid-producing medicinal plant *C. acuminata* (Supplementary Fig. S5). KEGG pathway enrichment analysis indicated that *G. macrophylla*-specific gene families were mainly enriched in sesquiterpenoid and triterpenoid biosynthesis, linoleic acid metabolism, and other secondary metabolites (Supplementary Fig. S6). These results are consistent with the presence of iridoids and terpenoids in *G. macrophylla*.

3.6. Phylogenetic relationships

Base on the gene families of the above mentioned 16 species, 100 single-copy orthologous genes were determined for the subsequent phylogenetic analysis with *O. sativa* as the outgroup. *Aquilegia coerulea* (Ranunculales) and other core eudicots formed an early branching clade with high BS support. All included species which belong to the Asterids formed a monophyletic group, which was sister to the Rosids (Fig. 3). Expectedly, *G. macrophylla* displayed a close relationship with *O. pumila*, *G. jasminoides*, and *C. canephora* all of which belong to the Gentianales (Fig. 3). Gentianales and Lamiales (Lamiids) diverged ~78.37 Ma (71.69–84.01 Ma) and the estimated divergence of the Asterids and Rosids clade occurred ~113.73 Ma (108.22–120.68 Ma).

3.7. Gene family expansion and contraction

The orthologous gene clusters generated from the OrthoFinder and the phylogenetic tree estimated by IQ-TREE were the input for CAFÉ v4.2 to infer whether expansion or contraction occurred in each gene family across species (Fig. 3). Among the 19,134 gene families identified in the *G. macrophylla* genome, 8,481 expansions and 7,440 contractions were detected. Compared with other three species in the Gentianales, *G. macrophylla* had the most expansions and contractions (*G. jasminoides*/*C. canephora*/*O. pumila* had 1,083/1,144/1,525 expansions and 3,049/3,416/6,363 contractions, respectively). GO enrichment analysis showed that the expanded gene families in *G. macrophylla* were involved in plant organ development, regulation of catalytic activity, root development, chloroplast thylakoid membrane, cytoskeleton, nucleosome, hydrolase activity and UDP-glycosyltransferase activity (Supplementary Fig. S7). KEGG pathway analysis suggested that these expanded gene families were enriched in inositol phosphate metabolism, phosphatidylinositol signalling system, pentose phosphate pathway, cutin, suberine and wax biosynthesis, diterpenoid biosynthesis, and carotenoid biosynthesis (Supplementary Fig. S8).

3.8. Analyses of genome synteny and WGD

MCScanX analysis was conducted to screen the genome synteny between *G. macrophylla* and two other species (*C. canephora* and *V. vinifera*), investigating potential WGD events during the evolutionary course of *G. macrophylla*. A total of 18.16% (10,050/55,337) colinear gene pairs on 740 colinear blocks were detected within the *G. macrophylla* genome

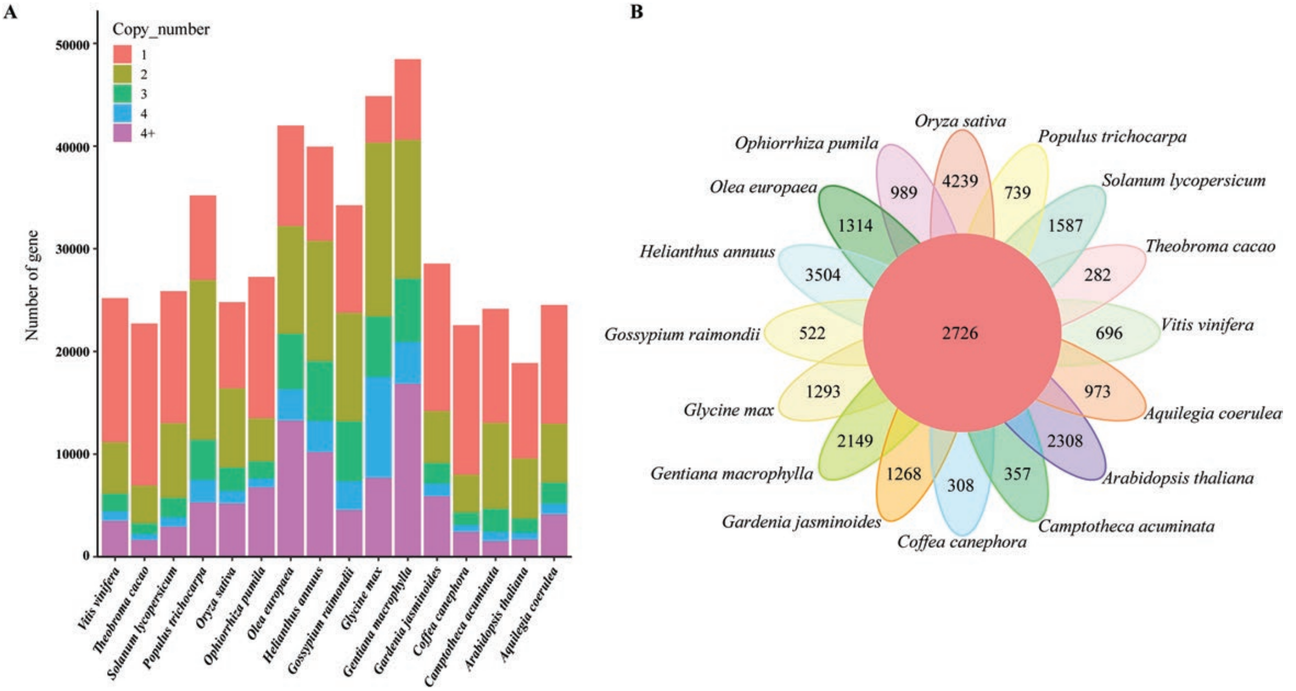

Figure 2. Gene family statistics in the comparative genomics analysis. (A) Number of genes in selected 16 angiosperm species, with *Gentiana macrophylla* having the highest number. (B) Venn diagram showing overlap of gene families between *G. macrophylla* and the other 15 angiosperm species.

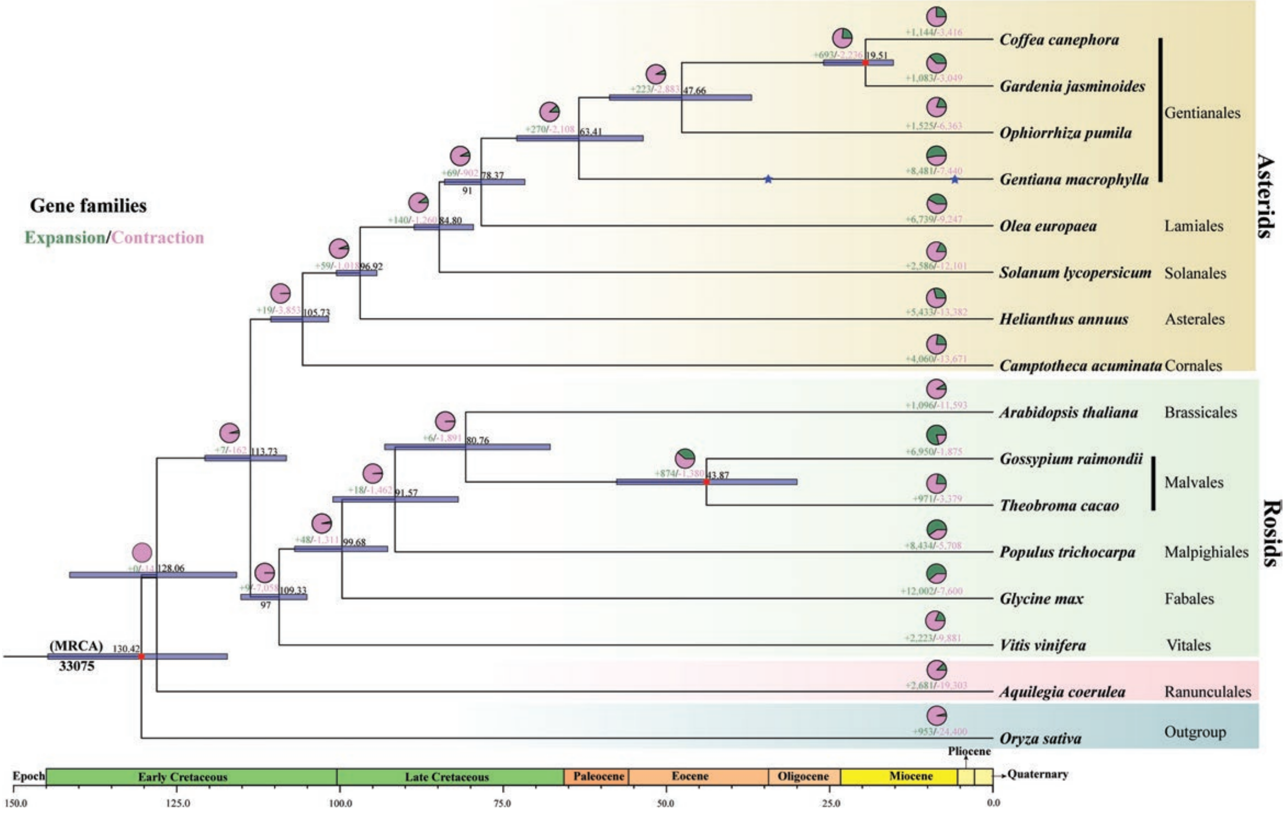

Figure 3. ML phylogeny and number of gene families that have expanded or contracted among 16 plant species. Confidence intervals of estimated divergence times are indicated at each node as teal bars. Calibrated nodes are labelled by red dots. The two WGD specific to *Gentiana macrophylla* are indicated by a blue star. All the branches are supported with a bootstrap value of 100 unless indicated otherwise below the branches.

(Supplementary Fig. S9). 14,681 colinear gene pairs from 892 colinear blocks were detected between *G. macrophylla* and *C. canephora*, and 12,361 colinear gene pairs from 1,087 colinear blocks between *G. macrophylla* and *V. vinifera* (Fig. 4A and D, Supplementary Fig. S10). Synteny analyses between *G. macrophylla* and *C. canephora*/*V. vinifera* provided clear

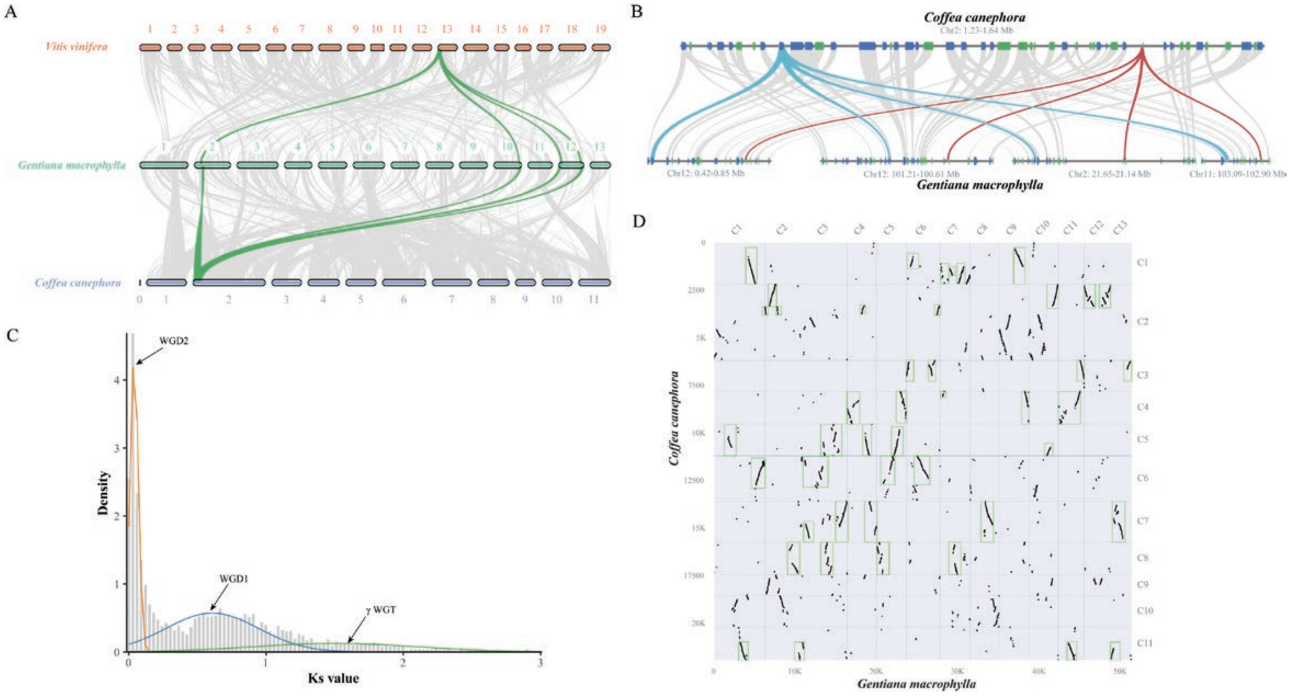

Figure 4. Inter-genomic comparisons revealed that *Gentiana macrophylla* experienced two rounds of WGD events. (A) Syntenic comparisons of *G. macrophylla*, *Coffea canephora*, and *Vitis vinifera* chromosomes. The collinearity pattern shows that typically an ancestral region in the *C. canephora* and *V. vinifera* genome has four corresponding copy regions in *G. macrophylla*. This 1:4 collinear relationship is highlighted in green. Syntenic blocks with more than 15 genes are linked by grey lines. (B) A representative block in the microsynteny comparison between *G. macrophylla* and *C. canephora* shows a 1:4 collinear relationship. Rectangles represent annotated genes with orientation on the same strand (blue) or reverse strand (green) and the grey lines connect syntenic gene pairs, with the 1:4 blocks are highlighted in colour. (C) Distribution of synonymous substitutions (Ks) of paralogous genes in the *G. macrophylla* genome recovered from syntenic blocks. The green, blue, and orange peaks represent WGT-γ, WGD1, and WGD2 during the evolutionary processes of the *G. macrophylla* genome, respectively. (D) Dot plot of inter-genomic comparison of *G. macrophylla* and *C. canephora* (15,882 gene pairs).

structural evidences for two WGDs in *G. macrophylla* with a 1:4 syntenic depth ratio in *G. macrophylla*–*C. canephora*/*V. vinifera* comparison (Fig. 4A and B, Supplementary Fig. S11). For instance, we identified the region of Chr2: 1.23–1.64 Mb in *C. canephora* corresponded to four genomic regions in *G. macrophylla* (Chr12: 0.42–0.85 Mb, Chr12: 101.21–100.61 Mb, Chr2: 21.65–21.14, and Chr11: 103.09–102.90) (Fig. 4B). Dot plots of syntenic blocks in *G. macrophylla*–*C. canephora*/*V. vinifera* also revealed a nearly 1:4 orthology ratio (Fig. 4D, Supplementary Fig. S10), indicating two WGD events might have occurred during *G. macrophylla* genome evolution. Besides, dot plots of syntenic blocks in *G. macrophylla*–*O. pumila* and *G. macrophylla*–*C. acuminata* also showed 1:4 and 2:4 orthology ratios, respectively (Supplementary Figs S12 and S13), which further supported two WGDs for *G. macrophylla*. However, we found that dot plots of syntenic blocks within *G. macrophylla* genome did not show a clear 1:3 syntenic depth ratio (Supplementary Fig. S9). We therefore analysed the duplicate gene origins in *G. macrophylla*. The results indicated that the dispersed duplication is the predominant type of gene duplication (57.12%, 31,607) compared with WGD or segmental duplication (18.16%, 10,050), proximal duplication (11.94%, 6,607), or tandem duplication (8.18%, 4,527).

To further validate and estimate the timing of the two WGD events in the *G. macrophylla* genome, synonymous nucleotide substitutions (Ks) were characterized between collinear homeologs within or between *G. macrophylla* and other five species (*C. canephora*, *G. jasminoides*, *O. pumila*, *C. acuminata*, and *V. vinifera*). The Ks distributions of one-to-one orthologs identified between *G. macrophylla* and the other five species showed different Ks peaks, indicating divergent evolutionary rates among these species (Supplementary Fig. S14). After dividing Ks values of *G. macrophylla* paralogs into three groups with Gaussian mode, Ks distribution showed two clear peaks of duplicated genes at Ks values of approximately 0.1 (WGD-1) and 0.6 (WGD-2) after the peak representing WGT-γ (Fig. 4C). Based on the abundance of 4DTv sites, two significant peaks were found in *G. macrophylla*, which also implied that *G. macrophylla* underwent two independent WGD events (Supplementary Fig. S15). These WGD events in the *G. macrophylla* genome might date back to ~5.80 and 34.60 Ma by mapping the WGD events onto the phylogeny, respectively. The peak value of orthologs between *G. macrophylla* and *O. pumila* (Ks ≈ 1.1) was coincident with the value of Ks between *G. macrophylla* and *G. jasminoides*/*C. canephora* (Ks ≈ 1.1) (Supplementary Fig. S14), implying that the split between *G. macrophylla* and these three species occurred simultaneously, which was also supported by the phylogenetic analyses. We thus inferred that WGD events in *G. macrophylla* occurred after the divergence from the *O. pumila*, *C. canephora*, and *G. jasminoides* lineage.

3.9. Expression pattern analyses of tissue-specific genes

The analysis of gene expression levels in the transcriptome showed that 1,323 genes were expressed in roots, 233 in stems, 600 in leaves, and 1,252 in flowers together with 17,701 genes

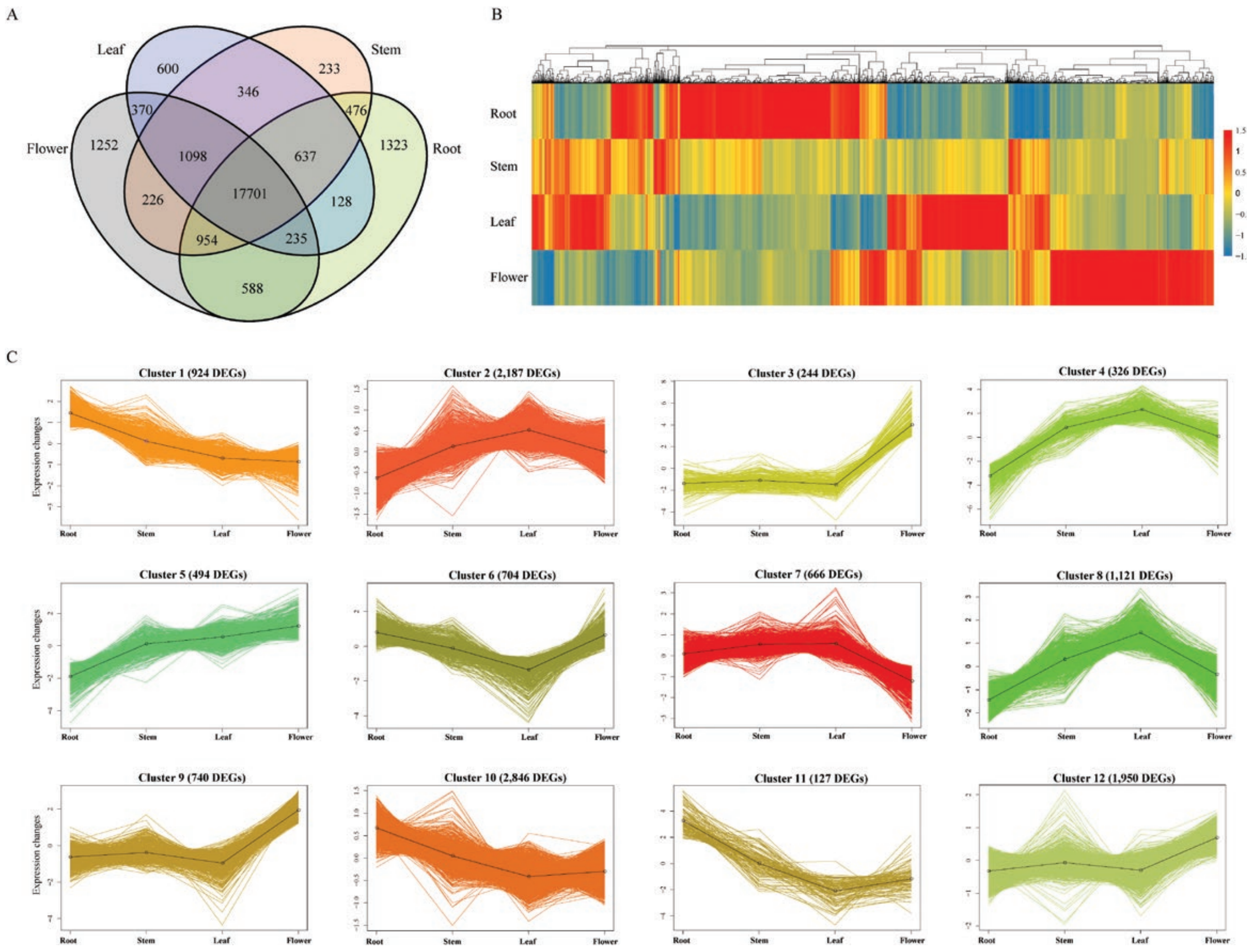

Figure 5. Gene expression patterns in different tissues of *Gentiana macrophylla*. (A) Venn diagram showing the expressed genes in roots, stems, leaves, and flowers. (B) Heatmap of all DEGs from different tissues. Each column in the figure represents one sample, each row represents one gene. The colour indicates the normalized gene expression level (log10(FPKM + 1)) in different tissues. (C) Hierarchical clustering showing the expression patterns of DEGs. The *y* axis represents different tissues, and the *x* axis represents the normalized gene expression level.

which did not show tissue-specific expression levels (Fig. 5A). 12,329 DEGs could be divided into 12 clusters (Fig. 5B and C) which also included the genes with tissue-specific expression. Clusters 1 and 11 contained genes related to defense responses such as 'plant–pathogen interaction', 'plant hormone signal transduction', and 'MAPK signalling pathway-plant' which were highly expressed in roots (Fig. 5C, Supplementary Table S10). In contrast, highly expressed genes in leaves derived from clusters 4 and 8 which were mainly correlated with fundamental pathways such as 'carbon metabolism' and 'photosynthesis'. Genes in clusters 3 and 9 showed high expression levels in flowers and were mainly enriched in 'phenylpropanoid biosynthesis', 'pentose and glucuronate interconversions', and 'starch and sucrose metabolism'. Unexpectedly, no stem-specific gene clusters were identified. Stems also had the smallest number of tissue-specific genes. These results provide the basis for the following analysis to investigate the gene expression regulatory networks and the mechanisms for the biosynthesis of bioactive metabolites in *G. macrophylla*.

3.10. TPS gene predictions and identification of genes related to the biosynthesis of iridoids

TPS is necessary to catalyse geranyl diphosphate, GGPP, and farnesyl diphosphate which function as precursors to produce monoterpenes, diterpenes, sesquiterpenes, and triterpenes. Using the Pfam domain models PF03936 and PF01397, a total of 39 putative TPS genes were identified in the genome of *G. macrophylla*. In order to classify the TPS proteins, we constructed a phylogenetic tree using all the TPS protein sequences from *G. macrophylla*, *A. thaliana*, *C. canephora*, *P. trichocarpa*, and *O. sativa*. The TPS genes found in *G. macrophylla* could be assigned to six subfamilies: TPS-a (26 genes), TPS-b (3 genes), TPS-c (2 gene), TPS-e/f (4 genes), and TPS-g (4 genes) (Fig. 6A, Supplementary Table S11). The TPS-a genes contained the largest number of genes and had a tandem duplication on Chr2 (9 genes) and Chr8 (12 genes) (Fig. 6B). Therefore, it is likely that the tandem duplication could be responsible for the TPS gene family expansion in *G. macrophylla*. In addition, we found that most TPS genes belong to the same group showed similar expression profiles (Fig. 6C). Especially, TPS-a genes were mainly expressed in leaves and TPS-g genes showed relatively high expression levels in stems, which may be the causes for the tissue-specific differences in substance synthesis.

Iridoids isolated from *G. macrophylla* are important terpenoid compounds which belong to the group of monoterpene analogs. It has been shown that the precursor for iridoid biosynthesis (Geranyl diphosphate) is derived from

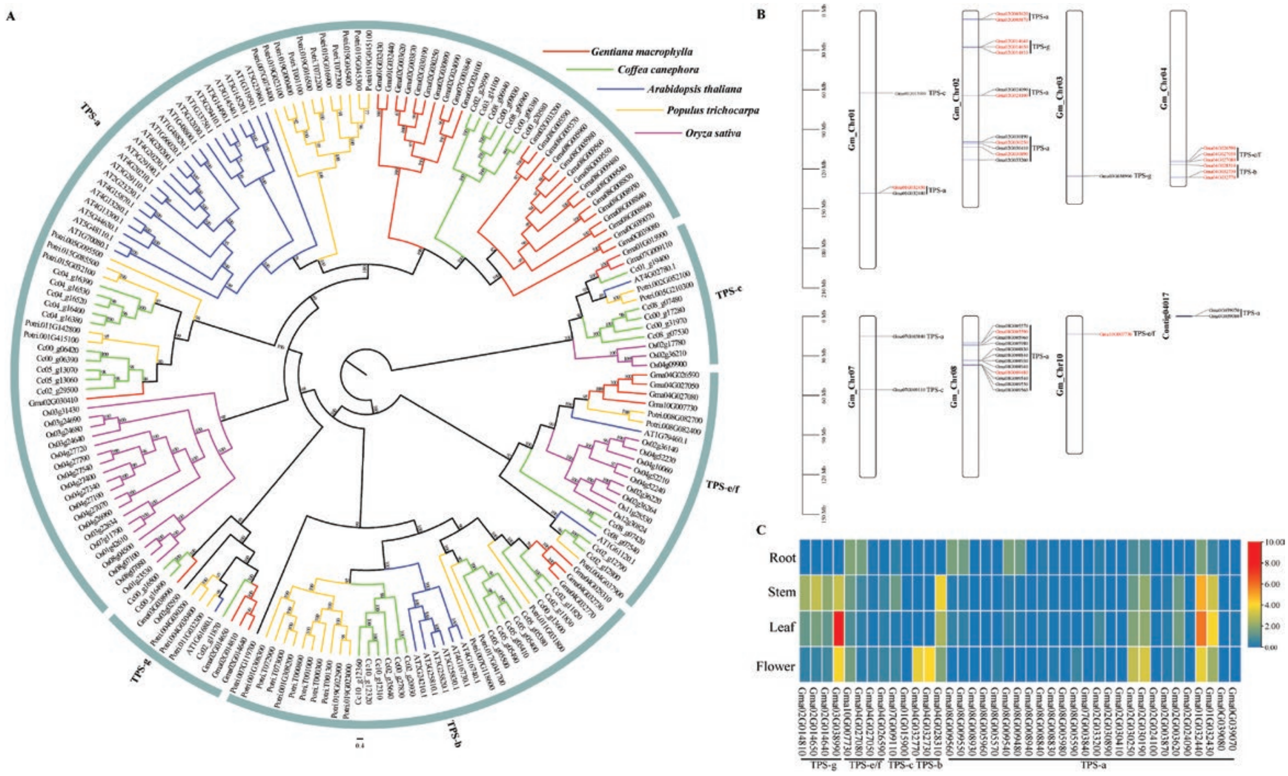

Figure 6. TPS gene family in *Gentiana macrophylla*. (A) The phylogenetic tree of the TPS gene family using an ML method. Branches are coloured based on the species colour scheme on the top right. (B) Chromosomal locations of TPS gene family in the *G. macrophylla* genome. The expanded gene families in *G. macrophylla* genome are labelled with red colour. (C) Heatmap showing the standardized gene expression levels (log10(FPKM + 1)) of TPS genes in different tissues.

dimethylally diphosphate synthesized via the MVA and MEP pathways.[85] Based on our genome annotation, all structural enzyme genes involved in the MVA and MEP pathways which may regulate the biosynthesis of iridoids could be identified (Supplementary Table S12). Three genes (*DXR*, *CMK*, and *MDS*) in the MEP pathway and one gene (*MVD*) in the MVA pathway were single-copy genes. Other genes in the MEP/MVA pathways had two to eight copies, where the corresponding gene families have experienced an expansion (Fig. 7). Gene expression analyses indicated that most MVA enzyme genes showed high expression levels in flowers, while enzyme genes in the MEP pathway were highly expressed in leaves. In the iridoid biosynthesis pathway, key enzyme genes such as *GES* (TPS-g subfamily), *G10H*, *8-HGO*, *IS*, *NEPS*, *IO*, *7-DLGT*, *7-DLNGT*, *7-DLH*, *LAMT*, *SLS*, and *STR* with different copy numbers were identified (Fig. 7). To investigate the copy number variations of iridoid-annotated genes in iridoid-producing species, iridoid-related genes from *O. pumila* and *C. acuminata* genomes were compared with *G. macrophylla*. The results indicated that the copy number of 15 iridoid genes in *G. macrophylla* was higher than in these species (Supplementary Table S13). Unexpectedly, all gene families related to structural genes in the iridoid biosynthesis pathway experienced expansion except for GES and genes such as *G10H*, *8-HGO*, *NEPS*, *7-DLGT*, *7-DLH*, *LAMT*, *SLS*, and *STR* may have undergone tandem duplication in the *G. macrophylla* genome based on the chromosome location (Fig. 8). We compared the expression profiles of enzyme genes related to the iridoid biosynthesis in different tissues and identified a total of 40 genes (*GES*, 7G10Hs, 7 *8-HGO*s, 2 *IS*s, 1 *NEPS*, 2 *IO*s, 6 *7-DLNGT*s, 4 *7-DLGT*s, 3 *DLH*s, 2 *SLS*s, 5 *STR*s) that were differentially expressed. Among these DEGs, 19 iridoid-related genes exhibited higher expression levels in leaves, and 13 were highly expressed in flowers (Supplementary Table S12), indicating that these genes may be the reason for the different expression levels of iridoid derivates in different tissues of *G. macrophylla*.

4. Discussion

Gentiana macrophylla is one of the important Qin herbs in TCM and is threatened due to over-harvesting of wild populations, which may result in the loss of germplasm diversity. In the past decades, most studies were carried out on the pharmacology of *G. macrophylla*, with little focus on the genomics of this species. Therefore, the availability of genomic information on *G. macrophylla* might contribute to its future genetic improvement and the development of molecular breeding programmes. Here, we combined Nanopore and Hi-C sequencing technologies to assemble a chromosome-level genome of *G. macrophylla*. Although the contig N50 of the assembly did not reach a high length (>1 M), approximately 98.89% of the assembly could be anchored to 13 chromosomes. In addition, the BUSCO assessment in this study also showed a relatively high value. Previous genome-wide assemblies of plants with large genomes or high heterozygosity such as *G. jasminoides*,[25] *Magnolia biondii*,[88] and *Rehmannia glutinosa*,[89] also had comparable contig N50 lengths.

About 47% of annotated genes could be matched to the proteins of the genus *Coffea*, the closest available genomic resource which belongs to the same order as the genus *Gentiana* (Gentianales). Phylogenomic analyses of 100 single-copy

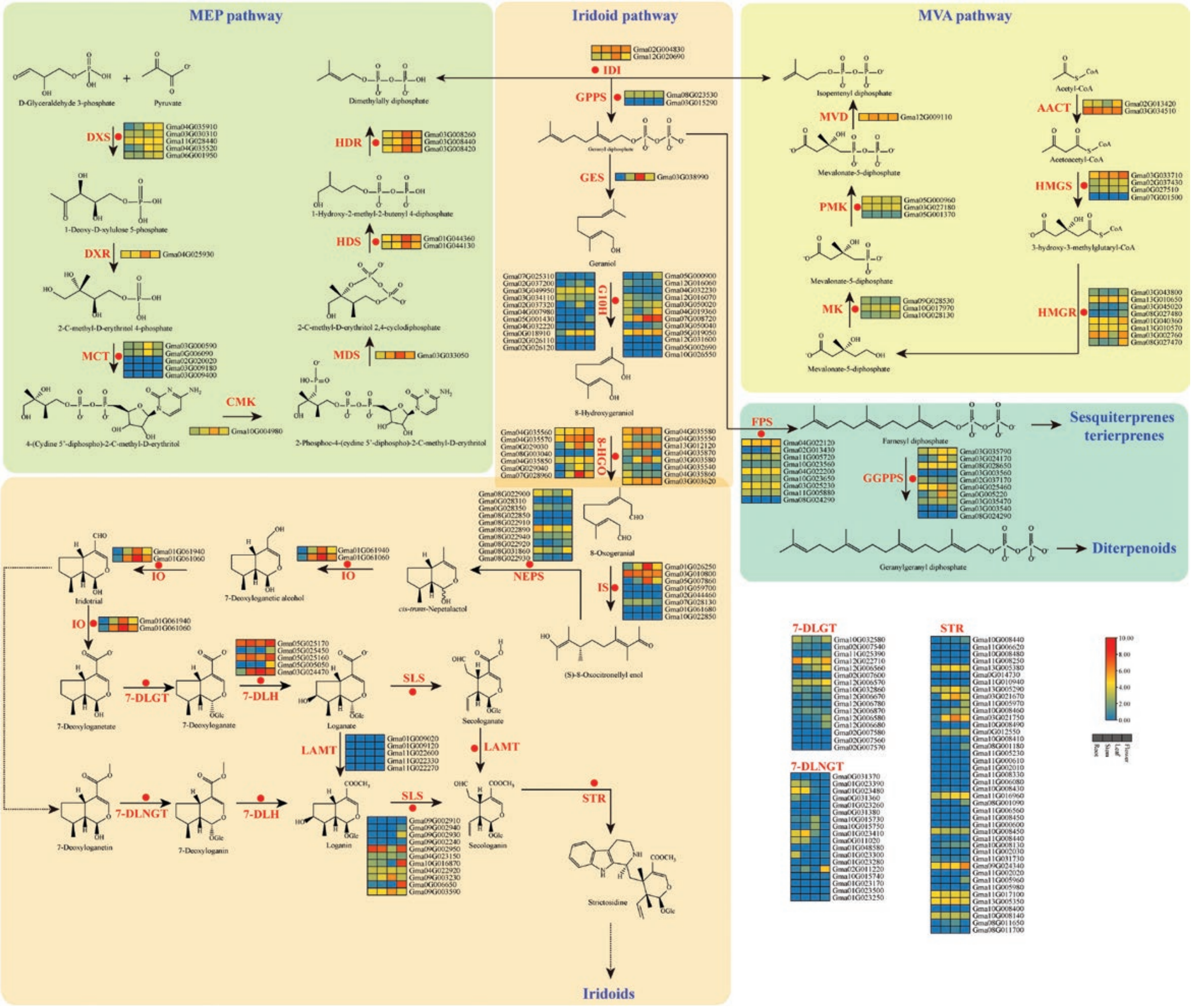

Figure 7. Expression analysis of genes involved in iridoids and other terpenoid biosynthesis. Different colour blocks represent the normalized gene expression levels (log10(FPKM + 1)) of all genes in different tissues. Red dots represent expanded gene families in *Gentiana macrophylla*.

orthologs from 16 seed plant genomes strongly supported the sister relationship of asterids and rosids, which is consistent with previous research on the phylogeny of angiosperms.[90] Our phylogenomic results also supported the monophyly of Gentianales with a sister relationship to Lamiales. This contrasts with a recent asterid phylotranscriptomics study which indicated that Gentianales and Boraginales formed a clade.[91] It is likely that this discrepancy could be due to the unbalanced lineage sampling in this study and the lack of genomic resources for Boraginales.

Polyploidization and WGD events, resulting in the expansion of repeated genes and increased genetic variation, are the prominent drivers for diversity in metabolites, species diversification and evolutionary novelty in plants.[22,92,93] A recent phylotranscriptomic study indicated extensive gene duplication events in the subtribe Gentianinae (Gentianaceae), particularly within the species-rich genus *Gentiana*.[19] We detected a large proportion of repeated genes in the *G. macrophylla* genome, and gene family analyses also indicated that *G. macrophylla* displayed the greatest number of expanded gene families apart for *G. max*. Unexpectedly, 19.75% of duplicated genes are syntenic and originated from WGD, however we did not find a clear 1:3 syntenic depth ratio in the dot plot. This is consistent with the chromosomal-level genome assembly of siberian ginseng (*Eleutherococcus senticosus*) which has a comparable genome size and N50 contig value (1.3 Gb and 309.43 kb, respectively), where a clear 1:3 syntenic depth ratio could not be detected.[94] Dispersed duplication (TEs) was the dominant type in the *G. macrophylla* genome. It has been showed that TEs shaped and (re)organized the chromosome structure of plant species such as *A. thaliana* and wheat.[95,96] It is therefore likely that TEs are responsible for the large genome size of *G. macrophylla*, and may also have caused large-scale chromosome rearrangements which might have compromised the intragenomic collinearity. The published genomes of *Taxus wallichiana* and *Magnolia biondii* which are also large (10.9 and 2.22 Gb, respectively) and have high TE content, show equally low proportions of intragenomic collinearity.[88,97] The quality of genome assembly may have a considerable effect on the retrieval of intragenomic synteny. Therefore, it is possible that current levels of synteny in *G. macrophylla* were underestimated as large scale syntenic

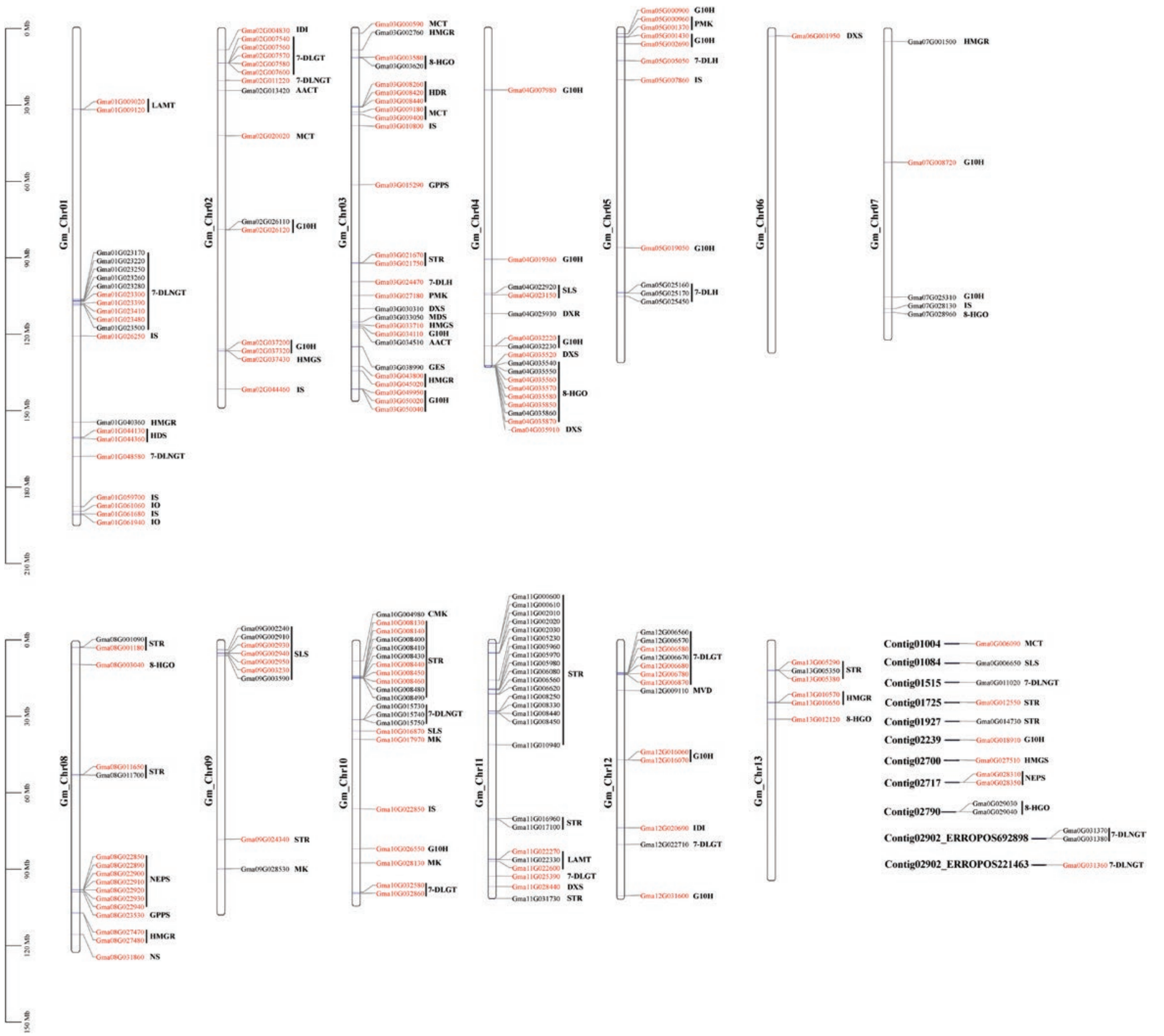

Figure 8. Chromosomal locations of genes related to biosynthesis of iridoids in the *Gentiana macrophylla* genome. The red gene name indicates that the corresponding gene family expanded in the *G. macrophylla* genome.

intragenomic regions may not have been detected because the current assembly of *G. macrophylla* did not achieve the more desirable N50 length >1 Mb (current contig N50 = 720.804 kb). However, both our inter-genomic synteny and Ks distribution analyses suggested two WGDs that occurred independently after WGT-γ in the *G. macrophylla* genome. It is likely that the WGDs may have facilitated gene family expansions and genome evolution in *G. macrophylla*.

Our analyses indicated that the two WGDs in *G. macrophylla* occurred ~5.8 and ~34.6 Ma which is after the divergence time (38.1 Ma) of *Gentiana* species.[98] A previous study also reported two rounds of independent WGD events for a different Gentianaceae species (*Sinoswertia tetraptera*), but the estimated times (41–46 and 67–75 Ma) were earlier than the ones for *G. macrophylla*.[99] This might indicate that the two rounds of WGD detected in *G. macrophylla* could be specific to the genus *Gentiana*. However, it is not clear if all the *Gentiana* species have undergone WGD events due to the limited genomic resources for this genus.

Root rot caused by pathogens is a major disease affecting the yield of *G. macrophylla*.[100] Based on expression patterns, genes highly expressed in roots related to plant–pathogen interactions, which may be an adaptative response to various pathogens. In addition, genes associated with a defense response such as ‘Plant hormone signal transduction’ and ‘MAPK signalling pathway-plant’ which play a vital role in plant disease resistance by regulating biotic defenses[101,102] also showed high expression levels in roots compared with other tissues. These results provide new insight into the molecular mechanism underlying the interaction between *G. macrophylla* and pathogens.

The major medicinally effective compounds of *G. macrophylla* are iridoids that belong to the terpenoids. The biosynthesis of terpenoids is initiated by the precursors dimethylally diphosphate and isopentenyl diphosphate, which are then converted by TPS to compounds such as monoterpenes, diterpenes, sesquiterpenes, and triterpenes.[102] All the TPS genes screened in the *G. macrophylla* genome were classified into six subfamilies based on the gene tree topologies of five angiosperm TPS proteins which is consistent with the classification of TPS genes in *Magnolia biondii* and *Aristolochia fimbriata*.[88,103] The gene tree topologies inferred from TPS proteins and our gene family analyses indicated the expansion of TPS gene family (TPS-a, TPS-b, and TPS-e/f) in *G. macrophylla* which are responsible for the high accumulation of terpenoids in *G. macrophylla*. Iridoids are the downstream derivates in the monoterpenoid biosynthesis pathway. We identified the candidate genes related to the biosynthesis of iridoids in *G. macrophylla* and analysed their copy number variance among iridoid-producing plants suggesting that 15 out of 27 iridoid-related genes had higher copy numbers compared with *O. pumila* and *C. acuminata*. No WGD signals were detected for *O. pumila*, and *C. acuminata* which experienced WGD.[79,80] We showed that *G. macrophylla* experienced two rounds of WGDs which likely increased the copy numbers of iridoid-related genes. In addition, we found that most gene families of key enzyme genes involved in the synthesis of iridoids also expanded in *G. macrophylla*. The expansion of key enzyme genes in specific metabolic pathways facilitates the biosynthesis and accumulation of active ingredients and is a ubiquitous phenomenon in medicinal plants during their evolutionary history.[89,102,104,105] However, we found that several expanded genes with the same function showed different expression patterns, indicating a potential functional divergence of these gene families.

5. Conclusions

The chromosome-level reference genome of *G. macrophylla* provides insight into the genome evolution and active component biosynthesis of this species. However, it will also be an important resource for researchers working on the genetic improvement and breeding of *G. macrophylla* and other *Gentiana* species. Furthermore, it will be useful to address evolutionary questions and investigate patterns of genome within the Gentianaceae.

Funding

This study was co-supported by the National Natural Science Foundation of China (31900273, 31800554), the Outstanding Youth Scientific Research Project of Shaanxi Academy of Sciences (2017K-12), the Scientific Research Project of Shaanxi Province Bureau of Traditional Chinese Medicine (610902), and the National Traditional Chinese Medicine Resources Survey Project in 2017 (2017-66). The Royal Botanic Garden Edinburgh is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division.

Authors' contributions

T.Z. and Y.Z. designed the research. T.Z. collected samples. T.Z., Y.H., G.B., and Y.Y. performed the experiments and analysed the data. T.Z., Y.H., M.R., and Y.Z. wrote the paper. All authors read and approved the final manuscript.

Conflict of interest

All authors did not have any conflict of interest.

Data availability

Raw sequencing data and genome assembly have been deposited at the NCBI under the BioProject: PRJNA831327.

Code availability

The custom scripts can be found at Github (https://github.com/yihenghu/Gentiana_macrophylla_genome_analysis).

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Zhang, X.L., Wang, Y.J., Ge, X.J., et al. 2009, Molecular phylogeny and biogeography of *Gentiana* sect. *Cruciata* (Gentianaceae) based on four chloroplast DNA datasets, *Taxon*, **58**, 862–70.
2. Rybczyński, J.J., Davey, M.R. and Mikuła, A. 2015, *The Gentianaceae—volume 2: biotechnology and applications*. Springer: New York.
3. Rybczyński, J.J., Davey, M.R. and Mikuła, A. 2014, *The Gentianaceae—volume 1: characterization and ecology*. Springer: New York.
4. Committee, C.P. 2020, *Pharmacopoeia of the People's Republic of China, Part 1*. China Medical Science Press: Beijing.
5. Chang-Liao, W.-L., Chien, C.-F., Lin, L.-C. and Tsai, T.-H. 2012, Isolation of gentiopicroside from *Gentianae* Radix and its pharmacokinetics on liver ischemia/reperfusion rats, *J. Ethnopharmacol.*, **141**, 668–73.
6. Yin, H., Zhao, Q., Sun, F.-M. and An, T. 2009, Gentiopicrin-producing endophytic fungus isolated from *Gentiana macrophylla*, *Phytomedicine*, **16**, 793–7.
7. Yu, F., Yu, F., Li, R. and Wang, R. 2004, Inhibitory effects of the *Gentiana macrophylla* (Gentianaceae) extract on rheumatoid arthritis of rats, *J. Ethnopharmacol.*, **95**, 77–81.
8. Zhang, X., Zhan, G., Jin, M., et al. 2018, Botany, traditional use, phytochemistry, pharmacology, quality control, and authentication of Radix Gentianae Macrophyllae-A traditional medicine: a review, *Phytomedicine*, **46**, 142–63.
9. Niu, Y.-T., Zhao, Y.-P., Jiao, Y.-F., et al. 2016, Protective effect of gentiopicroside against dextran sodium sulfate induced colitis in mice, *Int. Immunopharmacol.*, **39**, 16–22.
10. Park, E., Lee, C.G., Lim, E., et al. 2021, Osteoprotective effects of loganic acid on osteoblastic and osteoclastic cells and osteoporosis-induced mice, *Int. J. Mol. Sci.*, **22**, 233.
11. Hua, W., Zheng, P., He, Y., et al. 2014, An insight into the genes involved in secoiridoid biosynthesis in *Gentiana macrophylla* by RNA-seq, *Mol. Biol. Rep.*, **41**, 4817–25.
12. Tomiczak, K., Sliwinska, E. and Rybczyński, J.J. 2016, Comparison of the morphogenic potential of five *Gentiana* species in leaf mesophyll protoplast culture and ploidy stability of regenerated calli and plants, *Plant Cell Tissue Organ Cult.*, **126**, 319–31.
13. Zhang, L., Cao, B. and Bai, C. 2013, New reports of nuclear DNA content for 66 traditional Chinese medicinal plant taxa in China, *Caryologia*, **66**, 375–83.
14. Tomiczak, K., Mikuła, A., Sliwinska, E. and Rybczyński, J.J. 2015, Autotetraploid plant regeneration by indirect somatic embryogenesis from leaf mesophyll protoplasts of diploid *Gentiana decumbens* L.f., *In Vitro Cell. Dev. Biol. Plant*, **51**, 350–9.
15. Pustahija, F., Brown, S.C., Bogunić, F., et al. 2013, Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa, *Plant Soil*, **373**, 427–53.
16. Moeglein, M.K., Chatelet, D.S., Donoghue, M.J. and Edwards, E.J. 2020, Evolutionary dynamics of genome size in a radiation of woody plants, *Am. J. Bot.*, **107**, 1527–41.

17. Hu, Q., Peng, H., Bi, H., et al. 2016, Genetic homogenization of the nuclear ITS loci across two morphologically distinct gentians in their overlapping distributions in the Qinghai-Tibet Plateau, *Sci. Rep.*, **6**, 34244.
18. Zhang, X., Ge, X., Liu, J. and Yuan, Y. 2006, Morphological, karyological and molecular delimitation of two gentians: *Gentiana crassicaulis* versus *G. tibetica* (Gentianaceae), *Acta Phytotax. Sin.*, **44**, 627–40.
19. Chen, C.-L., Zhang, L., Li, J.-L., et al. 2021, Phylotranscriptomics reveals extensive gene duplication in the subtribe Gentianinae (Gentianaceae), *J. Syst. Evol.*, **59**, 1198–208.
20. Carretero-Paulet, L. and Van de Peer, Y. 2020, The evolutionary conundrum of whole-genome duplication, *Am. J. Bot.*, **107**, 1101–5.
21. Soltis, P.S. and Soltis, D.E. 2016, Ancient WGD events as drivers of key innovations in angiosperms, *Curr. Opin. Plant Biol.*, **30**, 159–65.
22. Soltis, P.S., Marchant, D.B., Van de Peer, Y. and Soltis, D.E. 2015, Polyploidy and genome evolution in plants, *Curr. Opin. Genet. Dev.*, **35**, 119–25.
23. Chen, D., Pan, Y., Wang, Y., et al. 2021, The chromosome-level reference genome of *Coptis chinensis* provides insights into genomic evolution and berberine biosynthesis, *Hortic. Res.*, **8**, 121.
24. Ma, D., Dong, S., Zhang, S., et al. 2021, Chromosome-level reference genome assembly provides insights into aroma biosynthesis in passion fruit (*Passiflora edulis*), *Mol. Ecol. Resour.*, **21**, 955–68.
25. Xu, Z., Pu, X., Gao, R., et al. 2020, Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants, *BMC Biol.*, **18**, 63.
26. Li, Y., Wei, H., Yang, J., et al. 2020, High-quality de novo assembly of the *Eucommia ulmoides* haploid genome provides new insights into evolution and rubber biosynthesis, *Hortic. Res.*, **7**, 183.
27. Zhao, Q., Yang, J., Cui, M.-Y., et al. 2019, The reference genome sequence of *Scutellaria baicalensis* provides insights into the evolution of wogonin biosynthesis, *Mol. Plant*, **12**, 935–50.
28. Yan, L., Wang, X., Liu, H., et al. 2015, The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb, *Mol. Plant*, **8**, 922–34.
29. Zhang, L., Li, X., Ma, B., et al. 2017, The tartary buckwheat genome provides insights into rutin biosynthesis and abiotic stress tolerance, *Mol. Plant*, **10**, 1224–37.
30. Doyle, J.J. 1987, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, *Phytochem. Bull.*, **19**, 11–5.
31. Chen, S., Zhou, Y., Chen, Y. and Gu, J. 2018, fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, **34**, i884–90.
32. Wick, R.R., Judd, L.M. and Holt, K.E. 2019, Performance of neural network basecalling tools for Oxford Nanopore sequencing, *Genome Biol.*, **20**, 129.
33. Rao, S.S.P., Huntley, M.H., Durand, N.C., et al. 2014, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell*, **159**, 1665–80.
34. Servant, N., Varoquaux, N., Lajoie, B.R., et al. 2015, HiC-Pro: an optimized and flexible pipeline for Hi-C data processing, *Genome Biol.*, **16**, 259.
35. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics*, **27**, 764–70.
36. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., et al. 2017, GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, **33**, 2202–4.
37. Koren, S., Walenz, B.P., Berlin, K., et al. 2017, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, *Genome Res.*, **27**, 722–36.
38. Walker, B.J., Abeel, T., Shea, T., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PLoS One*, **9**, e112963.
39. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
40. Burton, J.N., Adey, A., Patwardhan, R.P., et al. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol.*, **31**, 1119–25.
41. Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. 2019, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, *Nat. Biotechnol.*, **37**, 907–15.
42. Parra, G., Bradnam, K. and Korf, I. 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics*, **23**, 1061–7.
43. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
44. Britten, R.J. and Kohne, D.E. 1968, Repeated sequences in DNA: hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms, *Science*, **161**, 529–40.
45. Bao, Z. and Eddy, S.R. 2002, Automated de novo identification of repeat sequence families in sequenced genomes, *Genome Res.*, **12**, 1269–76.
46. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, *De novo* identification of repeat families in large genomes, *Bioinformatics*, **21**, i351–8.
47. Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinf.*, **9**, 18.
48. 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.
49. Ou, S. and Jiang, N. 2017, LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, *Plant Physiol.*, **176**, 1410–22.
50. Chen, N. 2004, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **5**, 4.10.1–14.10.14.
51. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–80.
52. Beier, S., Thiel, T., Münch, T., Scholz, U. and Mascher, M. 2017, MISA-web: a web server for microsatellite prediction, *Bioinformatics*, **33**, 2583–5.
53. Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. 2008, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, **24**, 637–44.
54. Korf, I. 2004, Gene finding in novel genomes, *BMC Bioinf.*, **5**, 59.
55. Keilwagen, J., Wenk, M., Erickson, J.L., et al. 2016, Using intron position conservation for homology-based gene prediction, *Nucleic Acids Res.*, **44**, e89.
56. 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.
57. Kim, D., Langmead B. and Salzberg, S.L. 2015, HISAT: a fast spliced aligner with low memory requirements, *Nat. Methods*, **12**, 357–60.
58. Pertea, M., Pertea, G.M., Antonescu, C.M., et al. 2015, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.*, **33**, 290–5.
59. 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.
60. 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.
61. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
62. Conesa, A., Götz, S., García-Gómez, J.M., et al. 2005, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674–6.
63. Lowe, T.M. and Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res.*, **25**, 955–64.

64. Griffiths-Jones, S., Moxon, S., Marshall, M., et al. 2005, Rfam: annotating non-coding RNAs in complete genomes, *Nucleic Acids Res.*, **33**, D121–4.
65. Nawrocki, E.P. and Eddy, S.R. 2013, Infernal 1.1: 100-fold faster RNA homology searches, *Bioinformatics*, **29**, 2933–5.
66. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. 2006, miRBase: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Res.*, **34**, D140–4.
67. Emms, D.M. and Kelly, S. 2019, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.*, **20**, 238.
68. Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P.D. 2018, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, *Nucleic Acids Res.*, **47**, D419–26.
69. Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. 2012, clusterProfiler: an R package for comparing biological themes among gene clusters, *OMICS*, **16**, 284–7.
70. 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.
71. Talavera, G. and Castresana, J. 2007, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *Syst. Biol.*, **56**, 564–77.
72. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. 2015, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, *Mol. Biol. Evol.*, **32**, 268–74.
73. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. 2017, ModelFinder: fast model selection for accurate phylogenetic estimates, *Nat. Methods*, **14**, 587.
74. Yang, Z. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
75. Han, M.V., Thomas, G.W.C., Lugo-Martinez, J. and Hahn, M.W. 2013, Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3, *Mol. Biol. Evol.*, **30**, 1987–97.
76. 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.
77. 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.
78. Wang, Y., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, **40**, e49.
79. Rai, A., Hirakawa, H., Nakabayashi, R., et al. 2021, Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis, *Nat. Commun.*, **12**, 405.
80. 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**, 3531.
81. Finn, R.D., Clements, J. and Eddy, S.R. 2011, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res.*, **39**, W29–37.
82. Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. 2009, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics*, **25**, 1972–3.
83. Banerjee, A. and Sharkey, T. 2014, Methylerythritol 4-phosphate (MEP) pathway metabolic regulation, *Nat. Prod. Rep.*, **31**, 1043–55.
84. Dellas, N., Thomas, S.T., Manning, G. and Noel, J.P. 2013, Discovery of a metabolic alternative to the classical mevalonate pathway, *eLife*, **2**, e00672.
85. Miettinen, K., Dong, L., Navrot, N., et al. 2014, The seco-iridoid pathway from *Catharanthus roseus*, *Nat. Commun.*, **5**, 3606.
86. Roberts, A. and Pachter, L. 2013, Streaming fragment assignment for real-time analysis of sequencing experiments, *Nat. Methods*, **10**, 71–3.
87. Chen, C., Chen, H., Zhang, Y., et al. 2020, TBtools: an integrative toolkit developed for interactive analyses of big biological data, *Mol. Plant*, **13**, 1194–202.
88. Dong, S., Liu, M., Liu, Y., et al. 2021, The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids, *Hortic. Res.*, **8**, 38.
89. Ma, L., Dong, C., Song, C., et al. 2021, De novo genome assembly of the potent medicinal plant *Rehmannia glutinosa* using nanopore technology, *Comput. Struct. Biotechnol. J.*, **19**, 3954–63.
90. Magallon, S. and Sanderson, M.J. 2001, Absolute diversification rates in angiosperm clades, *Evolution*, **55**, 1762–80.
91. Zhang, C., Zhang, T., Luebert, F., et al. 2020, Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications, *Mol. Biol. Evol.*, **37**, 3188–210.
92. Lichman, B.R., Godden, G.T. and Buell, C.R. 2020, Gene and genome duplications in the evolution of chemodiversity: perspectives from studies of Lamiaceae, *Curr. Opin. Plant Biol.*, **55**, 74–83.
93. Jia, K.-H., Liu, H., Zhang, R.-G., et al. 2021, Chromosome-scale assembly and evolution of the tetraploid *Salvia splendens* (Lamiaceae) genome, *Hortic. Res.*, **8**, 177.
94. Yang, Z., Chen, S., Wang, S., et al. 2021, Chromosomal-scale genome assembly of *Eleutherococcus senticosus* provides insights into chromosome evolution in Araliaceae, *Mol. Ecol. Resour.*, **21**, 2204–20.
95. Ito, H. and Kakutani, T. 2014, Control of transposable elements in *Arabidopsis thaliana*, *Chromosome Res.*, **22**, 217–23.
96. Daron, J., Glover, N., Pingault, L., et al. 2014, Organization and evolution of transposable elements along the bread wheat chromosome 3B, *Genome Biol.*, **15**, 546.
97. Cheng, J., Wang, X., Liu, X., et al. 2021, Chromosome-level genome of Himalayan yew provides insights into the origin and evolution of the paclitaxel biosynthetic pathway, *Mol. Plant*, **14**, 1199–209.
98. Fu, P.-C., Sun, S.-S., Twyford, A.D., et al. 2021, Lineage-specific plastid degradation in subtribe Gentianinae (Gentianaceae), *Ecol. Evol.*, **11**, 3286–99.
99. Zhu, M., Wang, Z., Yang, Y., et al. 2022, Multi-omics reveal differentiation and maintenance of dimorphic flowers in an alpine plant on the Qinghai-Tibet Plateau, *Mol. Ecol.*, doi: 10.1111/mec.16449.
100. Cheng, Y., Hou, Z. and Xu, Z. 2014, Prevention and control of *Gentiana macrophylla* Pall root rot disease, *Agric. Sci. Technol.*, **15**, 2039–40.
101. Verma, V., Ravindran, P. and Kumar, P.P. 2016, Plant hormone-mediated regulation of stress responses, *BMC Plant Biol.*, **16**, 86.
102. Wang, J., Xu, S., Mei, Y., et al. 2021, A high-quality genome assembly of *Morinda officinalis*, a famous native southern herb in the Lingnan region of southern China, *Hortic. Res.*, **8**, 135.
103. Qin, L., Hu, Y., Wang, J., et al. 2021, Insights into angiosperm evolution, floral development and chemical biosynthesis from the *Aristolochia fimbriata* genome, *Nat. Plants*, **7**, 1239–53.
104. Xu, Z., Gao, R., Pu, X., et al. 2020, Comparative genome analysis of *Scutellaria baicalensis* and *Scutellaria barbata* reveals the evolution of active flavonoid biosynthesis, *Genomics Proteomics Bioinformatics*, **18**, 230–40.
105. Niu, Z., Zhu, F., Fan, Y., et al. 2021, The chromosome-level reference genome assembly for *Dendrobium officinale* and its utility of functional genomics research and molecular breeding study, *Acta Pharm. Sin. B*, **11**, 2080–92.