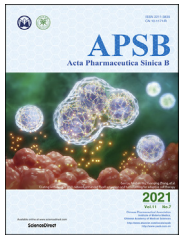
Acta Pharmaceutica Sinica B 2021;11(7):2080–2092

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

The chromosome-level reference genome assembly for *Dendrobium officinale* and its utility of functional genomics research and molecular breeding study

Zhitao Niu, Fei Zhu, Yajuan Fan, Chao Li, Benhou Zhang, Shuying Zhu, Zhenyu Hou, Mengting Wang, Jiapeng Yang, Qingyun Xue, Wei Liu, Xiaoyu Ding*

College of Life Sciences, Nanjing Normal University, China & Jiangsu Provincial Engineering Research Center for Technical Industrialization for Dendrobiums, Nanjing 210023, China

Received 20 October 2020; received in revised form 24 December 2020; accepted 21 January 2021

KEY WORDS

Dendrobium officinale;
Genome;
Active ingredients;
GWAS;
Plant production

Abstract *Dendrobium officinale*, an important medicinal plant of the genus *Dendrobium* in Orchidaceae family, has been used as traditional Chinese medicine (TCM) for nearly thousands of years. Here, we report the first chromosome-level reference genome of *D. officinale*, based on PacBio long-reads, Illumina short-reads and Hi-C data. The high-quality assembled genome is 1.23 Gb long, with contig N50 of 1.44 Mb. A total of 93.53% genome sequences were assembled into 19 pseudochromosomes with a super scaffold N50 of 63.07 Mb. Through comparative genomic analysis, we explored the expanded gene families of *D. officinale*, and also their impact on environmental adaptation and biosynthesis of secondary metabolites. We further performed detailed transcriptional analysis of *D. officinale*, and identified the candidate genes involved in the biosynthesis of three main active ingredients, including polysaccharides, alkaloids and flavonoids. In addition, the *MODIFYING WALL LIGNIN-1* (*MWL1*) gene, which inferred from Genome-Wide Association Studies (GWAS) based on the resequencing date from *D. officinale* and five related species and their morphologic features, may contribute to the plant production (yield of stems) of *D. officinale*. Therefore, the high-quality reference genome reported in this study could benefits functional genomics research and molecular breeding of *D. officinale*.

© 2021 Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Corresponding author. Tel./fax: +86 25 85891605.
E-mail address: dingxynj@263.net (Xiaoyu Ding).

Peer review under responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2021.01.019

2211-3835 © 2021 Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Orchids are popular economic plants worldwide not only for their aesthetic appeal, primarily reflected in their beautiful flowers, but also for their medicinal value. *Dendrobium*, one of the largest genera in Orchidaceae, includes approximately 1800 species mainly distributed in tropical Asia, Australasia, and Australia[1]. There are about 120 species in China and many of them grow in very strict conditions, such as epiphytic on cliffs or tree trunks, and distributed at high altitude above 1200 m[1–3]. To enhance their resistance to harsh environment, *Dendrobium* orchids accumulate high content of secondary metabolites, some of which are important active ingredients in medicinal plants, *e.g.*, polysaccharides, alkaloids, flavonoids, terpenes, benzyl compounds[4]. The high content of active ingredients in *Dendrobium* orchids have resulted in their excellent medicinal merits, such as benefiting the stomach, resisting cancer and enhancing the body's immunity[5]. Consequently, *Dendrobium* species are commonly used as tonic medicine and health food in many Asian countries for nearly thousands of years. However, *Dendrobium* species have many similar morphological characters, causing confusion and difficulty in their classification and taxonoic identification[3,6,7]. Although many *Dendrobium* species have been used as traditional Chinese medicine (TCM) for a long time, the molecular mechanisms underlying the synthesis of active ingredients in *Dendrobium* species are still unclear.

In China, in comparison with other *Dendrobium* species, *Dendrobium officinale* Kimura et Migo is the most popular and in demand, but the supply of *D. officinale* is limited due to its low germination rate, slow growing, and being overexploited[8]. Because of its ornamental and commercial value, *D. officinale* have attracted the intense attention of researchers, leading to the publication of numerous molecular studies. For example, Hou et al.[9] evaluated the processes of geographical distribution of *D. officinale* by using the sequence combination of four molecular markers, including nrDNA ITS sequence and three cpDNA sequences (*rps15-ycf1*, *accD-psaI* and *trnC-petN*). Zhu et al.[10] successfully distinguished *D. officinale* from other *Dendrobium* species based on the complete plastome sequences. Zhang et al.[11] provided useful information of glycosyltransferase genes and its expression patterns according to transcriptome data, which were important clues to explore pathways associated with polysaccharide biosynthesis. Additionally, two versions of *D. officinale* genomes have been sequenced, with contig N50 = 4.7 kb, scaffold N50 = 14.7 kb in genome v.1.0[12], and contig N50 = 33.1 kb, scaffold N50 = 391.4 kb in genome v.2.0[13]. Sequencing and investigating genome evolution are essential for the studies of medicinal plants, however; until now, high-quality genome sequences of *D. officinale* is still limited.

Recently, the rapid development in long-read sequencing technologies (*e.g.*, PacBio sequencing) and Hi-C sequencing approach have facilitated the generation of chromosome-scale genome assemblies in various plant species, especially for the species with large and complex genome[14,15]. Therefore, in this study, we assembled the first chromosome-level genome of *D. officinale* by combining PacBio long-reads, Illumina short-reads and Hi-C sequencing. By using the high-quality genome sequence, we performed genome-wide comparative studies that unraveled the genomic evolution of *D. officinale*, and demonstrated the application in functional genomics research and molecular breeding of *D. officinale*.

2. Materials and methods

2.1. Genomic DNA extraction and sequencing

D. officinale (voucher specimen: Niu2020) used for genome sequencing was sampled from Huoshan, Anhui (116.32 °N, 31.38 °E) and grew in the greenhouse of college of life sciences, Nanjing Normal University (Nanjing, China). Only fresh young leaves were harvested for the high-quality genomic DNA extraction, by using a modified CTAB method[16]. Three different methods were employed for genome sequencing. Firstly, a 400-bp short-insert library was prepared and sequenced using an Illumina Hiseq4000 sequencer. Approximately 125 Gb (100×) raw data were generated. Secondly, two 20-kb SMRT libraries were constructed. Then, the PacBio Sequel II platform were employed for whole-genome sequencing. Thirdly, 5 g of fresh young leaves was harvested for Hi-C experiments and library constructed[17]. In total, we generated approximately 132 Gb of Hi-C reads sequenced using an Illumina HiSeq2500 sequencer.

The individuals of *D. officinale* and five related species, *Dendrobium flexicaule*, *Dendrobium tosaense*, *Dendrobium scoriarum*, *Dendrobium shixingense* and *Dendrobium aduncum* were also sampled. We harvested 3−5 g of young leaves for each individual and extracted their total DNA using the DNeasy Plant Mini Kit (Qiagen). The total genomic DNA were sequenced on Illumina Hiseq4000 platform. Approximately 9.5 Gb raw reads were obtained for each species.

2.2. RNA extraction and sequencing

The stems, leaves, flowers, fruits and different growth stages of *D. officinale* were collected, and their total RNA were isolated using the MiniBEST Plant RNA Extraction Kit (Takara). The RNA sequencing was performed on an Illumina HiSeq2500 platform.

The tissue-cultured plants of *D. officinale*, which grew in the condition of 1/2 MS + 80 g/L banana slurry +60 g/L potato slurry +30 g/L sugar +7.8 g/L agar were used as control groups, while the salicylic acid (SA) stress treatment groups were grew in the condition of 1/2 MS + 80 g/L banana slurry + 60 g/L potato slurry + 30 g/L sugar + 7.8 g/L agar + SA 100 μmol/L. Their RNAs were also extracted and high-quality samples were used for RNA sequencing.

2.3. Genome assembly and pseudochromosome construction

For genome size estimation, we employed *k-mer* method using the clean Illumina short reads. The genome size and heterozygosity of *D. officinale* were calculate at the highest peak of *k-mer* frequency. The PacBio subreads were used for the draft genome assembly using Mecat2[18]. To correct errors in the primary assembly, we first polished the assembled contigs by Racon (v.1.4.3)[19], using PacBio subreads. Then, the modified contigs were polished three times by Pilon (v.1.22)[20], using Illumina short reads. The homologous contigs were further optimized and corrected by Purge_haplotigs (v.1.0.4)[21]. Finally, our contig-scale genome sequence was achieved with the contig N50 of 1.44 Mb.

In total, 132 Gb clean reads were obtained from the Hi-C sequencing and were mapped to the assembled contigs using BWA software[22]. Paired reads with mate mapped to a different contig were used to do the Hi-C associated scaffolding, while self-ligation, non-ligation and other invalid reads were filtered. The Lachesis software[23] was used to cluster and reorder 2430 contigs

into 19 pseudochromosomes. Then, we used Juicer software[24] to examine the direction and order of contigs based on the interaction results in the Hi-C heatmap. Finally, the chromosomal-level genome sequence for *D. officinale* were obtained, with chromosomal lengths from 37.98 to 94.97 Mb containing 93.53% of the total sequence.

The quality and completeness of our newly assembled *D. officinale* genome were evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis based on the gene set from Embryophyta_odb10 database[25].

2.4. Repeat sequence annotation

To identify the repetitive elements in our genome, two methods, homology-based and *de novo* prediction, were employed in repeat annotation. First, the RepeatMasker (v.1.0.10)[26] software was used for the homology-based prediction of the repetitive elements within the *D. officinale* genome based on the Repbase library (http://www.girinst.org/repbase). Then, we used RepeatModeler[27] to constructed a *de novo* repeats library, which employed Racon and RepeatScout to predict repetitive elements. The long terminal repeat (LTR) retrotransposons were *de novo* searched within the *D. officinale* genome using LTR_FINDER (v.1.0.7)[28]. The tandem and non-interspersed repeats were identified using the Tandem Repeat Finder (TRF) package[29] and RepeatMasker, respectively. Finally, we use RepeatMaker to combine the overlapped repeats and calculate the repeat contents.

2.5. Gene prediction and functional annotation

The high-quality protein-coding genes in our newly assembled *D. officinale* genome were predicted using homology-based, transcriptome-based, and *de novo* methods. For homology-based prediction, the annotated genes from the genome sequences of *Apostasia shenzhenica*, *D. officinale* v.1.0, *D. officinale* v.2.0, *Phalaenopsis equestris* and *Zea mays* were retrieved from NCBI. Those genes were aligned to the *D. officinale* genome v.3.0 using TBLASTN with the search parameter of *E*-value of 1×10^{-5}. After filtering low-quality results, we predicted gene structures using GeneWise (v.2.4.1)[30]. For the transcriptome-based gene prediction, two approaches were employed: (1) the transcriptome clean reads were *de novo* assembled and aligned to the genome sequence using Trinity (v.2.3.2)[31] and GMAP[32]; (2) the clean reads were directly mapped to our newly assembled genome v.3.0 using Hisat2, and reconstructed using Stringtie (v.1.3.1c)[33]. Then, we merged and filtered genes from two different methods using MIKADO (v.1.1)[34]. The software Augustus (v.3.3.1)[35] and GlimmerHMM[36] were used to perform *de novo* gene prediction. We generate the final set of gene annotation using the software MAKER2[37], and assessed the gene annotation results through the BUSCO test (BUSCO hits: 1,492, 93.5%).

The predicted genes were functionally annotated based on the publicly available databases including NR, TrEMBL, InterPro, Swiss-Prot, KEGG, Pfam, GO databases. Approximately 25,894 (about 93.71%) of the predicted protein-coding genes in *D. officinale* genome were annotated.

Furthermore, Non-coding RNAs were also annotated in *D. officinale* genome. tRNAscan-SE (v.1.23)[38] was used to find tRNA sequences. rRNA prediction was performed using RNAmmer (v.1.2)[39]. The miRNA and snRNA were predicted using INFERNAL implemented in Rfam_scan.pl (v.1.0.4)[40].

2.6. Gene family and phylogenetic analyses

The amino acid and nucleotide sequences of *D. officinale* and other 12 representative plant species (*A. shenzhenica*, *Arabidopsis thaliana*, *Artemisia annua*, *Cinnamomum kanehirae*, *Ginkgo biloba*, *Glycyrrhiza uralensis*, *Lonicera japonica*, *Oryza sativa*, *Panax ginseng*, *P. equestris*, *Rhodiola crenulate* and *Z. mays*) were retrieved and aligned to each other using BLASTP with the parameter *E*-value $< 1 \times 10^{-5}$. Genes with identity less than 30%, coverage less than 50% were filtered out. The OrthoMCL (v.2.0.9)[41] was used to cluster genes from these different species into gene families with the parameter of percent match cut off = 30, *E*-value exponent cut off = 1×10^{-5}.

The single copy genes from 13 species were extracted and aligned using MAFFT (v.7.313)[42]. The maximum likelihood (ML) phylogeny was reconstructed using RAxML (v.8.2.11)[43] based on the concatenated amino acid alignments with 1000 bootstrap replicates. The Bayesian relaxed molecular clock approach in MCMCtree, which implemented in PAML (v.4.9)[44] was used to estimate divergence time with the parameters of nsample = 1,000,000; burnin = 200,000; seqtype = 0; model = 4. The divergence times of *A. thaliana*−*G. uralensis* (98−117 Mya), *L. japonica*−*P. ginseng* (82−102 Mya) and *O. sativa*−*Z. mays* (42−52 millions of years ago, Mya) were inferred from the TimeTree[45] database.

2.7. Gene family expansion and contraction

The gene families that have experienced expansion or contraction were searched using the CAFE package (v.4.2)[46]. The expanded or contracted genes which specific to *D. officinale* were functional annotated and enriched using Blast2GO.

2.8. Synteny and whole-genome duplication (WGD) analysis

To examine WGD events in *D. officinale*, we isolated the homologous proteins between *D. officinale*−*A. thaliana* and *D. officinale*−*A. shenzhenica* using all-to-all search in BLASTP with the parameters of -*E*-value 1×10^{-5} -outfmt 6. Then, we employed MCScanX package[47] to search collinear blocks (regions with at least five collinear gene pairs) and calculated the synonymous substitution rates (K_s) between each gene pairs using yn00 implemented in PAML. We also constructed a dot plot by ggplot2 package (v.2.2.1) in R (v.2.15). The synonymous substitution rates were further employed to estimated divergence times based on Eq. (1):

$$T = K_s/2r \ (r = 3.39 \times 10^{-9}) \quad (1)$$

The synteny analysis were also performed between *D. officinale*−*O. sativa* and *D. officinale*−*Z. mays* to confirm that *D. officinale* had undergone WGD events. The insertion time of LTR retrotransposons were estimated based on Eq. (2):

$$T = K_s/2r \ (r = 1 \times 10^{-8}) \quad (2)$$

2.9. Transcriptome analysis and CYP450 gene superfamily

The clean reads obtained from RNA-Seq were mapped to the genome of *D. officinale* and assembled using Hisat2 and

Stringtie, respectively. The differential expression genes (DEGs) were identified using the DEseq2 package in R with the standard of adjusted *P*-value of 0.05 and the foldchange more than 1.5×.

The annotated *CYP450* genes of four species: *D. officinale*, *A. thaliana*, *A. shenzhenica* and *P. equestris* were retrieved. Moreover, the Pfam domains: PF00067 was searched against the genome of *D. officinale* using HMMER (v.3.0)[48]. Candidate genes were further confirmed using the SMART tool[49]. Pseudogenes or genes with sequence lengths <200 amino acids were removed. The gene sequences of *CYP450* were aligned using MAFFT and manually adjusted using MEGA 5.2[50]. The *CYP450* gene tree was constructed using RAxML with 1000 bootstrap replicates.

2.10. Single nucleotide polymorphism (SNP) calling and GWAS

The clean reads from 38 individuals of *D. officinale* and five related species were mapped to the newly assembled *D. officinale* genome using Bowtie 2 (v.2.4.1)[51]. All the samples were aligned to the reference genome v.3.0 with the mapping rate >90%. The software of GATK 4.0[52] were used for the SNP calling. The raw SNPs were filtered using Plink (v.1.9)[53] with the parameters of minor allele frequency (MAF) < 0.05 and missing rate < 0.5. Finally, we obtained 1.98 million high-quality SNPs. These SNP data were used to perform GWAS for six traits in TASSEL (v.5.0)[54]. The genome-wide significance thresholds were approximately 6.31×10^{-8}, which calculated using Eq. (3):

$$P = 0.05/n \tag{3}$$

where n represents effective number of SNPs. The Manhattan plots were constructed by CMplot package in R.

2.11. Data availability

All of the raw sequence reads used in this study have been deposited in NCBI under the BioProject accession number PRJNA662181. The final chromosome-scale genome assembly was submitted to the NCBI with accession number JACXSL000000000.

3. Results

3.1. Genome sequencing and assembly of D. officinale

The genome size, heterozygosity and repetitive ratio of *D. officinale* were evaluated using *k-mer* distribution analysis. The highest peak of 17 *k-mer* frequency was occurred at the depth of 90. The estimated genome size of *D. officinale* was 1.21 Gb, with an extremely high heterozygosity of 1.27% and repetitive ratio of 64.39% (Supporting Information Fig. S1 and Table S1). For genome assembly, a total of 131.1 Gb of PacBio subreads (108× coverage) with an average N50 of 22.6 kb were obtained from two SMRT cells in PacBio Sequel platform (Supporting Information Table S2). The assembled contigs were anchored to 19 pseudochromosomes using Hi-C sequencing approach (anchored rate 93.53%, Table 1)[12,13]. The final assembled chromosome-level genome of *D. officinale* was 1.23 Gb in size with 2430 contigs (contig N50 = 1.44 Mb, scaffold N50 = 63.07 Mb, Supporting Information Fig. S2 and Table S3). These results collectively showed that our new genome is well assembled than both previous genome versions (Table 1)[12,13].

The quality and completeness of our newly assembled genome was evaluated based on the BUSCO test. A total of 1614 plant-specific orthologs were searched, 1515 (93.9%) genes were identified in the assembly, of which 1471 (91.2%) genes were complete (Supporting Information Table S4). The accuracy of genome sequence was assessed by continuous long reads mapping analysis. As shown in Supporting Information Tables S5, and 87.86% reads have covered 99.81% genome sequence, and 96.27% of genome sequence have been covered ≥20×. Moreover, Illumina short read mapping results showed a high mapping rate of 99.12% and a low base error percentage of 0.0047%. These evaluation results indicate the high completeness, high continuity, and high accuracy of the present genome assembly.

3.2. Repeat and gene annotations

Two approaches, homology-based and *de novo* prediction, were employed to identified the repetitive sequences in the *D. officinale* genome. In total, 76.77% of assembled sequences were identified as repetitive sequences (Supporting Information Table S6). Transposable elements (TE) account for 76.77% of the *D. officinale* genome, and 59.72% of them are LTR elements (Supporting Information Table S7). A total of 27,631 high-quality protein-coding genes were annotated based on the homology-based, transcriptome-based, and *de novo* gene prediction (Supporting Information Tables S8−S10). The average gene length and coding DNA sequence (CDS) length of annotated genes were 17,023 and 1086 bp, respectively. Among these genes, 93.61% and 89.01% were functionally annotated in the NR and TrEMBL database (Supporting Information Table S9). Functional annotations also showed that 46% of the genes could be classified by GO terms, and 34.4% of the genes could be annotated to KEGG pathways. In addition, we identified 1523 noncoding RNA (ncRNA) genes, including 57 miRNAs, 498 tRNAs, 684 rRNAs and 284 snRNAs (Fig. 1A). Evaluation test showed 93.5% of BUSCO genes were existed in the annotation (Table S4), suggesting that the annotated genome is largely complete. Moreover, the high accurately mapping rate (90.8%) of the RNA-Seq reads from stems, leaves, flowers, fruits and different growth stages further support the completeness of our *D. officinale* genome sequence.

Table 1 A comparison of the three published *D. officinale* genomes.

Item	*D. officinale* v.1.0[12]	*D. officinale* v.2.0[13]	*D. officinale* v.3.0 (this study)
Genome size	1.36 Gb	1.01 Gb	1.23 Gb
Contig N50	4.70 kb	33.09 kb	1.44 Mb
Max contig length	129.06 kb	288.53 kb	8.75 Mb
Contig number	814,881	105,732	2430
Scaffold N50	14.75 kb	391.46 kb	–
Hi-C Scaffold N50	–	–	63.07 Mb
Hi-C Anchored rate	–	–	93.53%

–Not applicable.

3.3. Genome characterization of *D. officinale*

Genome characters of *D. officinale*, *e.g.*, gene density, gene expression, transposon density, GC content etc., were calculated. As shown in Fig. 1A, GC-rich regions exhibit higher gene density, transposon density and repeat density. We identified 618,258 and 508,465 copies of *Gypsy* and *Copia* element, occupied 25.65% and 23.93% of the genome sequences, respectively (Supporting Information Table S11). The *Gypsy* element mainly distributed in GC-rich regions, while the *Copia* element were distributed in opposite regions. Three versions of *D. officinale* genomes including our newly assembled and two previously published were compared in present study. In total, we identified 1,164,757 SNPs and 253,171 Insertion-Deletion sites (InDels) between our newly assembled genome v.3.0 and v.1.0 and 983,912 SNPs and 252,373 InDels between genome v.3.0 and v.2.0 (Fig. 1C[12,13] and Supporting Information Table S13). The distribution of presence/absence variations (PAVs), which only present in genome v.3.0 but entirely missing in v.1.0 and v.2.0 was identified. We identified 24,922 segments with a total length of 877.32 Mb were present in genome v.3.0 but absent in v.1.0. Similarly, there were 24,718 segments with a total length of 909.03 Mb were present in genome v.3.0 but missing in v.2.0 (Fig. 1C[12,13] and Table S12). These results indicated that our genome assembly was more continuity and accuracy than previous studies. Additionally, we found that the gene-rich regions exhibited higher number of SNPs and InDels relative to gene-poor regions, whereas, the gene-poor regions contain higher number of PAV segments than the gene-rich regions.

3.4. Comparative genomic analysis of *D. officinale*

The annotated protein-coding genes were clustered into different gene families among *D. officinale* and other orchid or medicinal plants according to the sequence similarity. A total of 27,631 *D. officinale* genes clustered into 13,903 gene families, of which 12,398 gene families were common to orchid species and 1196 were specific to *D. officinale* (Fig. 2A and Table S13). The GO enrichment analysis showed that the specific genes were largely enriched in terms of biological process, cellular component and molecular function part (Supporting Information Fig. S3A). KEGG enrichment results showed that these genes were enriched mainly in pathways for environmental adaptation and biosynthesis of secondary metabolites (Fig. S3B).

A total of 1456 single copy genes among 13 species were concatenated for the phylogenetic analysis. The ML tree revealed a monophyletic group among orchid species and *D. officinale* showed a sister relationship with *P. equestris*. The divergence time between *D. officinale* and *P. equestris* was estimated to be 66.4 (41.5−141.0) Mya (Fig. 2B). The phylogenetic relationships among 13 species are consistent with previous studies.

The K_s values between collinear genes were calculated to identify potential WGD events. The density plot of K_s values showed three peaks at 0.05, 0.84 and 1.50, which suggested that *D. officinale* has experienced three rounds of WGD (Fig. 2C). The identification of I-WGD and II-WGD events were consistent with previously study[13]. The collinear relationship among *D. officinale*, *O. sativa* and *Z. mays* also confirmed the two WGD events. For each genomic region of *O. sativa* and *Z. mays*, we typically found two matching regions in *D. officinale* (Supporting Information Fig. S4). The K_s distribution showed an obvious peak at $K_s = 0.05$. The time distribution of LTR retrotransposon insertions showed a large number of insertions occurred recently, which result in the independent peak (Fig. 2D). These results indicated that *D. officinale* did not experienced an independent WGD event after the first two rounds of WGD.

The gene family expansion and contraction analysis revealed a total of 820 gene families expanded and 975 gene families contracted in *D. officinale* (Fig. 2B). Functional analysis showed that gene families related to secondary metabolism, especially for the biosynthesis of active ingredients, *e.g.*, polysaccharides, alkaloids and flavonoids, have experienced large expansions. For example, KEGG enrichment analysis suggested that expanded genes were enriched in pentose and glucuronate interconversions (ko00040), starch and sucrose metabolism (ko00500), fructose and mannose metabolism (ko00051), phenylpropanoid biosynthesis (ko00940), flavonoid biosynthesis (ko00941) and isoquinoline alkaloid biosynthesis (ko00950, Supporting Information Table S14). Additionally, genes related to photosynthesis were also expanded, *e.g.*, photosynthesis (ko00195), photosynthesis-antenna proteins

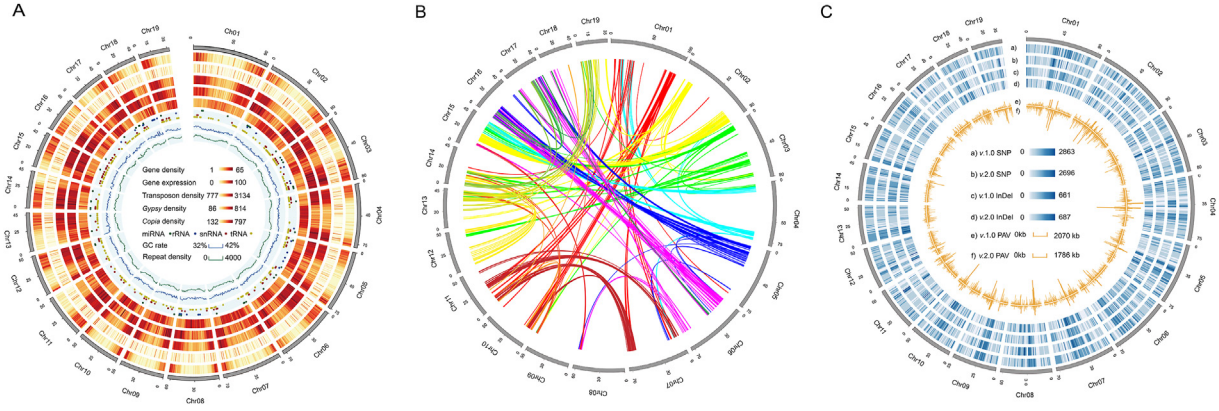


Figure 1 Genomic landscape of the *D. officinale* genome (A) Genome features of *D. officinale* genome v.3.0. (B) Synteny information of *D. officinale* genome v.3.0. The lines with different color indicate synteny blocks of paralogous sequences related to the whole genome duplication event. (C) Comparisons of three published *D. officinale* genomes. a), SNP between genome v.3.0 and v.1.0; b), SNP between genome v.3.0 and v.2.0; c), InDel variation among genome v.3.0 and v.1.0; d), InDel variation among genome v.3.0 and v.2.0; e) and f), PAV distribution among three genomes.

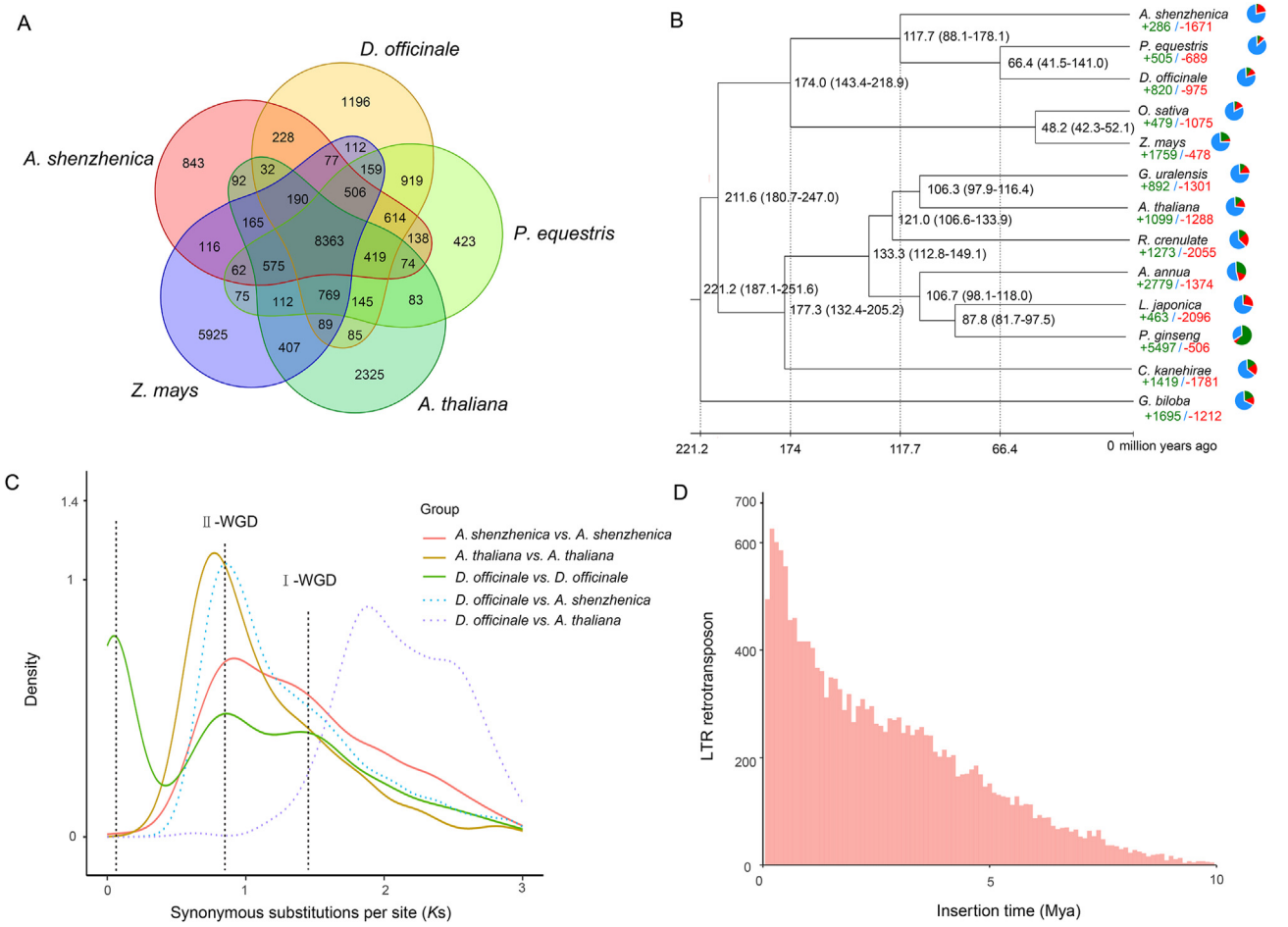


Figure 2 Comparative genomic analysis of *D. officinale*. (A) Gene clusters for orchid species and two model plants, *A. thaliana*. and *Z. mays*. (B) Phylogenetic relationships and divergence times between *D. officinale* and other plant species. Expansions and contractions of gene families were also showed in the tree. The colors in green and red indicate the expanded and contracted gene families, respectively. (C) Distribution of the synonymous substitution rate (K_s) between *A. shenzhenica*, *A. thaliana*. and *D. officinale*. (D) The estimated time distribution of LTR retrotransposon insertions.

(ko00196) and carbon fixation in photosynthetic organisms (ko00710). Fast evolved gene analyses and positive selection results indicated that many genes have experienced fast evolved or positive selection, some of which are key enzyme genes related to the synthesis of active ingredients and photosynthesis in *D. officinale* (Supporting Information Table S15).

3.5. Identification of genes involved in the biosynthesis of active ingredients

Using a combination of homolog searching and functional annotation methods, the candidate genes for the synthesis of three active ingredients, polysaccharides, alkaloids and flavonoids, in *D. officinale* genome were identified. Polysaccharides are major active compounds in *Dendrobium* species, which have been reported to play important roles in demonstrated to show prominent bioactivities, including anti-tumor, immune stimulation and antioxidant. In total, we identified 268 genes (encoded 56 enzymes) were related to polysaccharides biosynthesis (Fig. 3A and Supporting Information Table S16). Besides, we annotated 98 and 52 genes related to the synthesis of alkaloids and flavonoids, respectively (Table S16). Among them, 56 genes (encoded 25 enzymes) were found to be related to the synthesis of sesquiterpene alkaloids, terpenoid indole alkaloids and their upstream pathways, including shikimate pathway, mevalonate (MVA) and methylerythritol 4-phosphate (MEP) pathways (Fig. 3B).

As shown in Supporting Information Fig. S5 the contents of active ingredients, *e.g.*, polysaccharides, alkaloids and flavonoids, were significantly changed in *D. officinale* after SA stress for 35 days. Therefore, we selected non-SA stress and 35 days SA stress treatment groups for transcriptome analysis. Analysis of differential expression showed a total of 1677 DEGs, with 772 up-regulated and 905 down-regulated DEGs (Supporting Information Fig. S6). Among them, 107 DEGs were related to biosynthesis of active compounds, of which 51 DEGs were found to be related to the biosynthesis of polysaccharides, alkaloids and flavonoids. In the biosynthesis of flavonoids, two key enzymes (F3′H, encoded by three genes and DFR, encoded by two genes) related to anthocyanins synthesis were significantly up-regulated, which may contribute to the color variation among different *D. officinale* individuals (as shown in Fig. S3C). Additionally, 156 DEGs of transcript factors were identified (Fig. S6). These genes and pathways will provide valuable information for studies on underlying the molecular mechanisms of active ingredients synthesis.

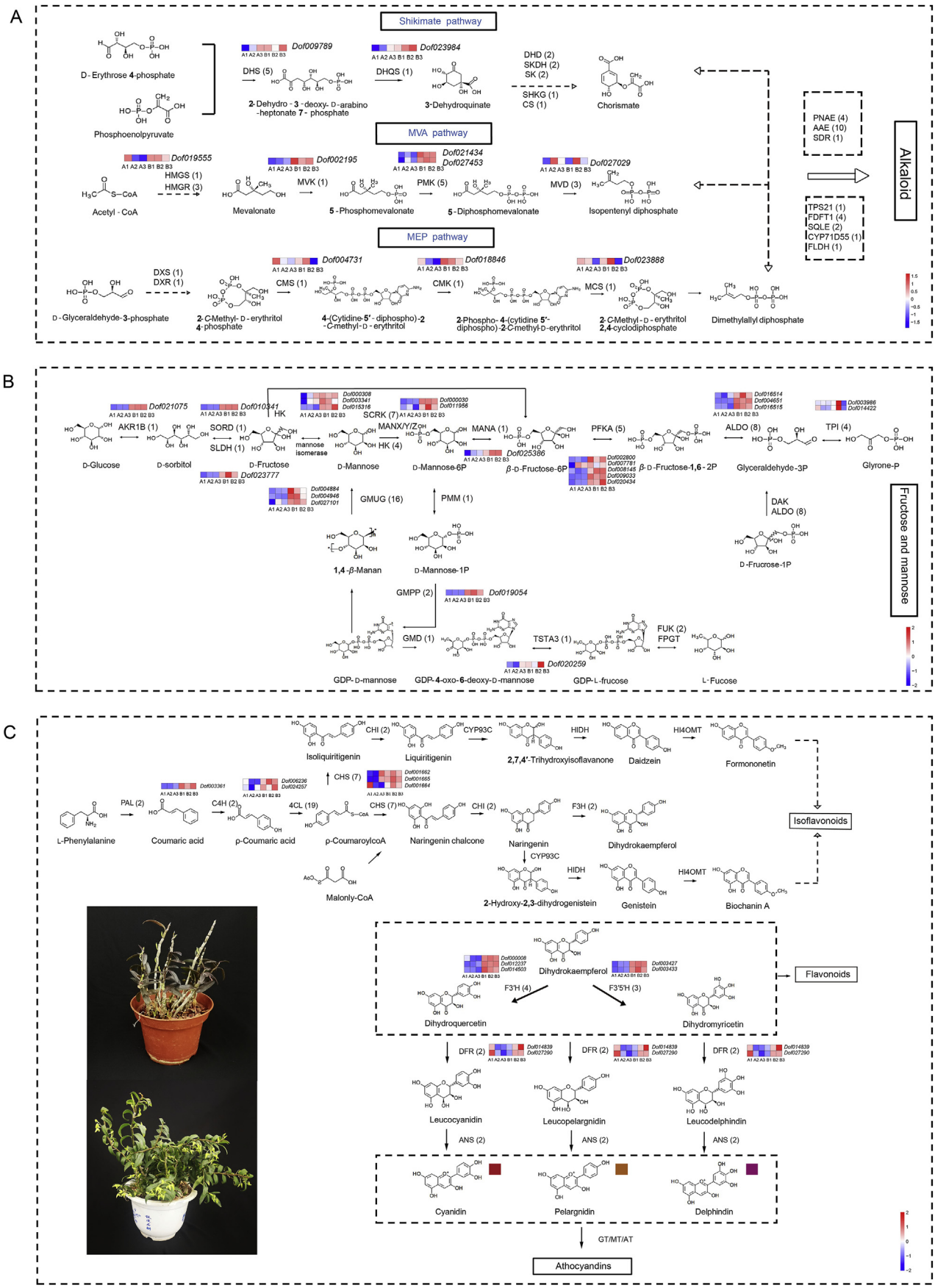


Figure 3 Biosynthesis pathways of three main active ingredients in *D. officinale*. (A) Alkaloid. (B) Fructose and mannose. (C) Flavonoids. Values within brackets indicate the numbers of gene copies corresponding to the genes in the pathways. The differential expression of genes involved in the pathways that under SA stress are also showed. A1−A3: the control groups, B1−B3: the SA treatment groups.

3.6. *The cytochrome P450 gene superfamily*

The cytochrome P450 gene superfamily is one of the largest enzyme family of plant metabolism, which participate in biosynthesis of diverse secondary metabolites, whereas no study has demonstrated interest in comprehensive comparison of the genes in *D. officinale*. In the present study, a total of 218 *CYP450* genes were predicted, annotated and confirmed by comparisons of homologous gene sequences from *A. thaliana*. The identified *CYP450* genes are categorized in 9 groups, namely CYP51 (1 family, 1 gene), CYP71 (13 family, 128 gene), CYP72 (6 family, 26 gene), CYP74 (1 family, 7 gene), CYP85 (7 family, 26 gene), CYP86 (4 family, 23 gene), CYP97 (1 family, 3 gene), CYP710 (1 family, 1 gene) and CYP711 (1 family, 3 gene) (Supporting Information Table S17). By including a genomic data set of *D. officinale*, *A. thaliana* and two other orchid species (*A. shenzhenica*, and *P. equestris*), phylogenetic analyses of *CYP450* genes were performed as shown in Fig. 4A. With 128 genes in 13 families, CYP71 groups are most diversified, which contains three orchid-specific and four *A. thaliana*-specific CYP450 families. Moreover, a total of 29 differentially expressed *CYP450* genes, which encoded key enzymes in active ingredients synthesis were identified according to the comparative transcriptome analysis between control and SA treatment groups. Among them, 26 genes were belonging to CYP71 groups, of which 17 genes were up-regulated (Fig. 4B). For example, one *CYP71D* genes related to the biosynthesis of sesquiterpenoid and triterpenoid, four *CYP75A* genes and six *CYP75B* genes related to the biosynthesis of flavonoids were significantly up-regulated. These results indicate that genes in CYP71 groups play important roles in regulating the biosynthesis of active compounds in *D. officinale*, especially for flavonoid compounds.

3.7. *GWAS reveals significant SNPs associated with gene loci that related to the production of Dendrobium species*

A total of 38 individuals from *D. officinale* and five related species were sampled from 13 regions, which represent most of the phenotypic and genetic diversity known for *D. officinale* species in China (Supporting Information Table S18). We carefully documented the morphologic features which related to plant production, *e.g.*, plant height, stem diameter, leaf length, internode length, leaf width and leaf shape index for each individual, all of which showed a broad phenotypic distribution in the population (Table S18). Approximately 9.5 Gb clean reads were sequenced using an Illumina Hiseq4000 sequencer, with average coverage depth $>8.0\times$ for each individual. Using the newly assembled chromosome-level genome (v.3.0) as reference, we obtained 6.93 million raw SNPs and 1.98 million high-quality SNPs using a strict filtering standard (MAF < 0.05, missing rate < 0.5, Fig. 5B and Supporting Information Table S19). Through conducting GWAS using the PCA values of six morphologic features of the individuals across six groups including *D. officinale* and five related species, we identified 13 GWAS loci based on the suggested threshold *P*-value $< 6.31 \times 10^{-8}$ (Fig. 5A and C). We noticed that the identified GWAS loci included many well characterized genes: four genes for plant height traits, two genes for leaf length traits, three genes for stemdia meter traits and one for internode length traits (Supporting Information Table S20). Among them, *MWL1* gene, responsible majorly for cell wall growing which enhance the plant height growth, may contribute to plant production (yield of stems) variation among *Dendrobium* species. However, there were still a few GWAS loci in which the causal genes remain to be characterized. For example, two genes (*Dof007132* and *Dof007133*) around GW7 (Supporting Information Table S20), which strongly associated with the stem diameter traits have not been reported. Further genetic and functional studies are needed to validate the involvement of these candidate genes in regulating the production of *Dendrobium* species.

4. Discussion

4.1. *The first chromosome-level genome assembly for orchid species*

A contiguous and well-annotated genome sequence is the foundation for gene-function discovery, phylogenetic and genome evolutionary studies of medical plants. The *D. officinale* genome has been reported to have high heterozygosity and high repetitive ratio, which presents great challenge for genome assembly[12]. Neither of the two previously published genomes[12,13], which were assembled from Illumina short-reads, have achieved high-quality sequences due to the frequent fragmentation and low contiguity. Compared with Illumina short-reads sequencing, PacBio sequencing technologies have many advantages, *e.g.*, generated longer subreads (up to 30−40 kb) and facilitated the construction of repetitive regions, and have been successfully used in complex genome assembly, especially for highly heterozygous medicinal plant genomes, such as *A. annua*[55], *Scutellaria baicalensis*[56] and *Tripterygium wilfordii*[57]. Indeed, the long-read sequencing technologies used in this study have provide substantial improvements of the *D. officinale* genome than the v.1.0 and v.2.0 genomes (Table 1)[12,13]. The newly assembled genome v.3.0 was improved with a 306- and 44-fold increase in the value of contig N50. Only 2430 gaps were found in v.3.0, which is 335- and 44-fold better than those of the previous versions. The high-quality genome sequences also facilitated the construction of repetitive regions and gene annotation. For example, (i) the newly assembled genome constructed substantially more TEs, especially for LTR elements (59.72%, Table S7) than in the previous two genomes (38.54% in v.1.0[12] and 45.72% in v.2.0[13]). (ii) The average gene length and average exon number of annotated genes was significantly larger than in genome v.2.0[13], which undoubtedly more accurate and complete. The high-quality and completeness of our newly assembled *D. officinale* genome v.3.0 were further confirmed by BUSCO evaluation (93.9% BUSCOs were identified in the assembly, Table S4). Moreover, this is the first chromosome-level genome assembly for orchid species. We employed Hi-C sequencing data to assemble the contigs into 19 pseudochromosomes with the final super scaffold N50 = 63.07 Mb and the maximum pseudochromosome length = 94.97 Mb. The newly assembled genome sequence could be a valuable genetic resource for underlying the molecular mechanisms of the biosynthesis of active ingredients and provides insight into the evolution of the orchid family.

4.2. *Transposable elements may contribute to the diversification, speciation, and adaptation of D. officinale after the rapid uplift of the Qinghai-Tibetan Plateau (QTP)*

Ever since Darwin published *Fertilization of Orchids* in 1862, the evolutionary relationship among orchid species have attracted great interest from evolutionary biologists and botanists[58−61]. We constructed the phylogenetic tree to infer the relationship among

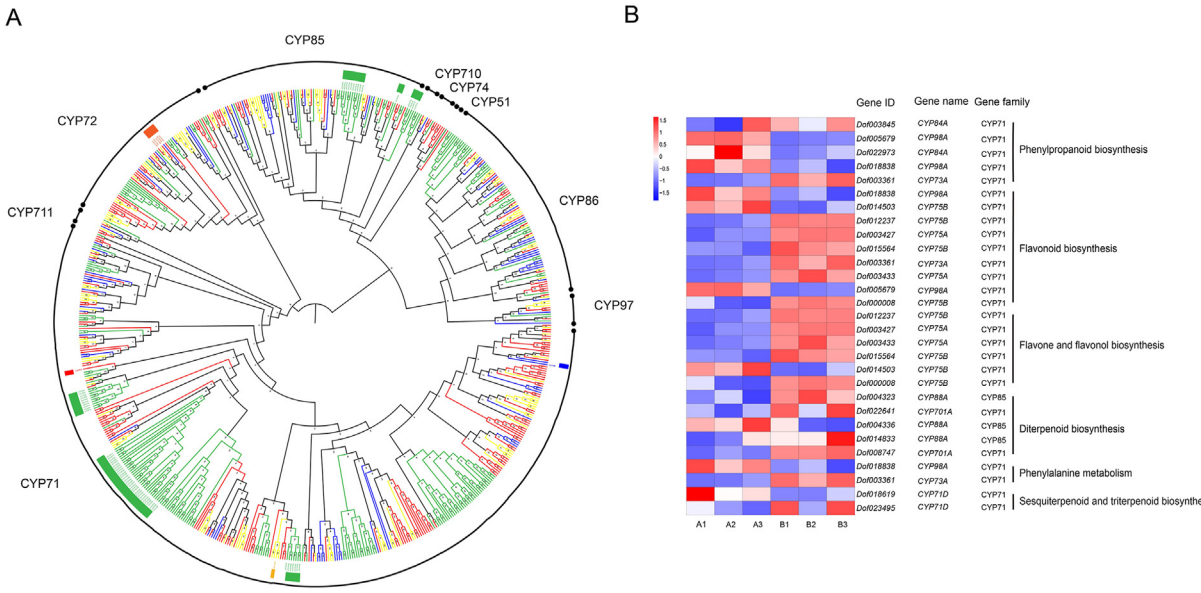


Figure 4 Phylogenetic analysis and expression patterns of the cytochrome P450 gene superfamily. (A) Phylogenetic placements of the 218 *CYP450* genes of *D. officinale*. (B) The differential expression of *CYP450* genes involved in polysaccharides, alkaloids and flavonoids biosynthetic pathways that under SA stress.

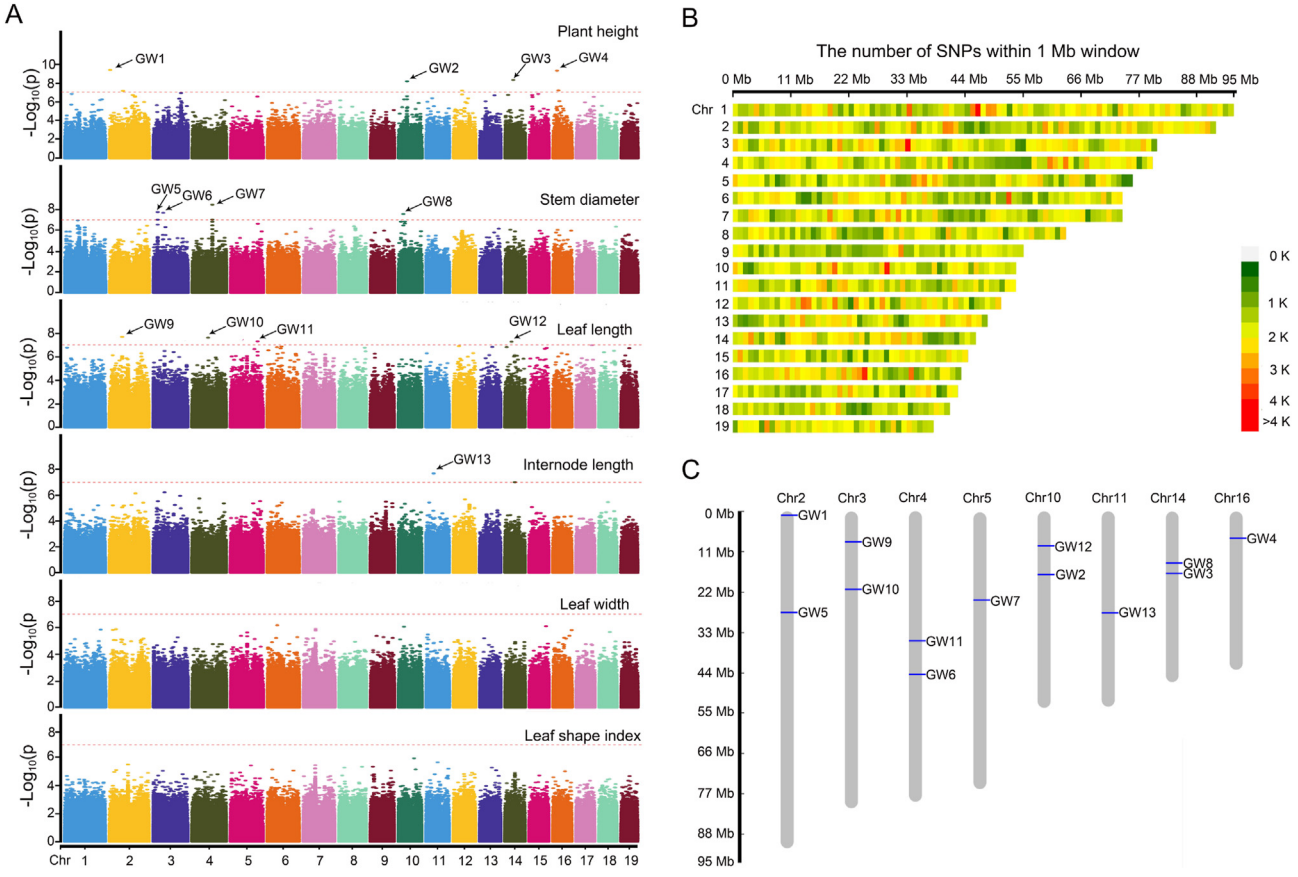


Figure 5 GWAS for *D. officinale* and its 5 closely related species. (A) Manhattan plots of assessed morphologic features which related to plant production, *e.g.*, plant height, stem diameter, internode length, leaf length, leaf width and leaf shape index. (B) SNP distribution among 19 pseudochromosomes of *D. officinale*. (C) The distributions of 13 GWAS loci in the pseudochromosomes of *D. officinale*.

D. officinale and other two orchid species and found that (i) three orchid species formed a monophyletic group, (ii) *D. officinale* is sister to *P. equestris*, which were consistent with the results of previously published phylogenetic analyses[58,59,61].

Compared with the genome size of *A. shenzhenica*, species from the basal-most subfamily with an assembled genome size of 349 Mb[62], the *D. officinale* genome is considerably larger. It has been well documented that (i) WGD events and (ii) TE amplification are major causes of genome expansion[63,64]. In this study, the K_s distribution indicated two rounds of WGD events occurred in *D. officinale* genome, one of which was consistent with *A. shenzhenica* (II-WGD)[62]. Although there is an additional *D. officinale*-specific WGD (I-WGD) occurred around 115.4 Mya, the gene number of *D. officinale* (27,631 protein-coding genes) were nearly the same with *A. shenzhenica* (21,841 protein-coding genes). In contrast, the transposable elements occupied 76.77% of the *D. officinale* genome, which was significantly higher than that reported for *A. shenzhenica* genome (42.05%[62]). In particular, the high proportion of TEs in *D. officinale* were LTR elements, among them 83.03% elements were LTR retrotransposons, whereas other TEs only constituted 17% collectively. The LTR retrotransposons were diversified among orchid species, *e.g.*, 22.06% in *A. shenzhenica*[62], 53.14% in *Gastrodia elata*[65], 46.46% in *P. equestris*[66]. Moreover, the abundant insertion of LTR retrotransposons have resulted in an independent peak at $K_s = 0.05$, which indicated a significant genome expansion occurred after the two rounds of WGDs (Fig. 2C). Therefore, the proliferation of LTR retrotransposons might be responsible for the genome expansion of *D. officinale*.

The activity of TEs, including LTR retrotransposons led to diverse genomic changes, *e.g.*, duplicated or created genes, regulated gene expression and rearrange the chromosomal structures, which may drive lineage-specific diversification and adaptation[67,68]. For example, TEs have caused the rapid phenotypic variation among *Capsella rubella* by regulating the expression levels of their adjacent genes, which resulted in the increasing of species adaptation to changing environments[69]. Our previous researches suggested that *D. officinale* was most likely originated from South Yungui Plateau, which is adjacent to the QTP[8]. The intraspecific divergence of *D. officinale* was occurred at 2.06 Mya, shortly after the dramatic elevational and climatic changes of the QTP (the most recent rapid uplift of the QTP, 8–0 Mya) during the period of late Miocene to late Pliocene[70]. In present study, the density plot of K^s values showed an independent peak occurred at 3.8 Mya (Fig. 2C). Moreover, insertion time estimation using LTR retrotransposons showed that most of LTR retrotransposons were expanded within 5 Mya (Fig. 2D). These results indicated that the proliferation of LTR retrotransposons occurred after rapid uplift of the QTP. Therefore, we hypothesize that transposable elements may contribute to the diversification, speciation, and adaptation of *D. officinale* after the rapid uplift of the QTP.

4.3. The newly assembled genome sequence provides new insight into the evolution of D. officinale

The high-quality genome sequence of *D. officinale* reported here could be an important genetic resource that provides new insight into the evolution of *D. officinale*. We counted four kinds of genes, including *D. officinale*-specific genes, expanded genes, fast evolved genes and positive selected genes from the *D. officinale* genome, many of them were key enzyme genes that involved in the pathways for environmental adaptation. For example, 10 genes involved in the pathway of cutin, suberine and wax biosynthesis (ko00073) were specific to *D. officinale*, which can enhance the adaptative ability of leaf and stem for environmental changes. In addition, the variation of photosynthesis pathway are also important adaptive characteristics of *Dendrobium* orchids, *e.g.*, the photosynthesis pathway of *D. officinale* could shift from C3 to CAM pathway according to the environmental changes[71]. A total of 32 genes that related to photosynthesis pathway were expanded in *D. officinale* genome, of which 18 genes encode 7 key enzymes were involved in the pathway of carbon fixation in photosynthetic organisms (ko00710, Table S14). Among them, we found that (i) seven genes, which encoded the enzymes of RUBISCO, and (ii) four genes, which encoded the enzymes of malate dehydrogenase, were significantly expanded. These two genes are important key enzymes involved in Calvin−Benson cycle and crassulacean acid metabolism, respectively[72]. Moreover, 14 genes, which encoded the 6 enzymes related to photosynthetic electron−transfer reaction in the pathway of photosynthesis (ko00195) and photosynthesis-antenna proteins (ko00196) were also expanded. The expansion of these genes could enhance the ability of photosynthetic capacity and CO_2 fixation, which led to the increase of the adaptation to environmental changes for *D. officinale*.

4.4. The high-quality genome sequence benefits research that underlying biosynthesis of active ingredients in D. officinale

Bioactive ingredients and their pharmacological effects play important roles in the merits of medicinal plants[57,73–75]. Therefore, biosynthetic pathway analysis and key enzyme gene discovery have become the main aims of study in medicinal plants. Previous researches of *D. officinale* have focused on the transcriptome sequencing and identified a lot of genes which involved in the synthesis and metabolic pathways[11,76,77]. However, there is still a limited of comprehensive analysis that underlying the synthesis of active ingredients in *Dendrobium* species. Salicylic acid, functions as a signaling molecule, could be applied to enhance the accumulation of active ingredients in medical plants[78]. Therefore, to investigate the biosynthesis of three main active ingredients in *D. officinale*, including polysaccharides, alkaloids and flavonoids, the expression of candidate genes was analyzed by coupling the high-quality genome sequence generated here with RNA-Seq studies under SA stress.

The polysaccharide of *Dendrobium* species mainly consists of glucose, mannose and galactose, of which mannose are the main monosaccharides in *D. officinale*[4,11]. We totally identified 268 genes, which encoded 56 enzymes involved in polysaccharides biosynthesis. Among them, 67 genes were related to the pathway of fructose and mannose metabolism (ko00051). DEGs analysis showed 24 genes, which encoded 12 key enzymes, *e.g.*, mannose-6-phosphate isomerase (MANA) and hexokinase (HK), were up-regulated. Such data for *D. officinale* could be used as an important resource to investigate key enzyme genes and the metabolic pathway of polysaccharides in *D. officinale*.

Early work focused on the pharmacological effects of *Dendrobium* species have found that dendrobine-type alkaloids can improve the body's immunity, benefiting the stomach and lowing blood pressure[79]. *D. officinale* alkaloids belong to the sesquiterpene alkaloids and terpenoid indole alkaloid classes, which are the downstream products of MVA and MEP pathway[76]. A total of 98 genes that related to the biosynthesis of alkaloids were identified, of which 32 genes encoded 17 key enzymes were found to be related to the pathway of terpenoid backbone biosynthesis

(ko00900). Several MVA and MEP pathway genes were significantly up-regulated by the SA treatment, indicating an improvement of active precursor supply for the alkaloid biosynthesis under SA treatment.

The flavonoids are also important active compounds in medical plants. Moreover, the differential of flavonoids biosynthesis may contribute to the color variation among different plant individuals[80]. For example, in *D. officinale*, the up-regulated of *CYP75A* and *CYP75B* genes, encoded the key enzyme of dihydroflavonol 4-reductase (DFR), have led to the accumulation of cyanidin, which may result in the color changes from green to red among different *D. officinale* individuals. In addition, we found that in some biosynthesis steps, the enzyme-coding genes contents more than two copies, *e.g*., the enzyme of mannan *endo*-1,4-beta-mannosidase (GMUG) involved in ko00051 has 16 gene copies, the enzyme of hydroxymethylglutaryl-CoA reductase (NADPH) involved in ko00900 has 3 gene copies, the enzyme of chalcone synthase (CHS) involved in ko00941 has 7 gene copies. The increase in copy number may drive the production of major active ingredients in *D. officinale* and account for its excellent medicinal effects because gene expansions might be responsible for enhancing the ability for active compounds synthesis.

4.5. GWAS reveals several significant SNPs associated with known genes that contributed to plant production (yield of stems) of D. officinale

The high-quality genome sequence could not only benefit functional genomics research, but also useful for molecular breeding analysis. Based on the high-quality of the genome sequences, GWAS could be performed to quickly and comprehensively identify candidate genes that are related to the production and quality of economic plants, *e.g*., cotton[81], maize[15] and tea plant[82]. Our previous studies showed that the phylogenetic relationship among *D. officinale* and five species are closely related[9,10], while the morphologies among these species are differ significantly. For example, the plant height, which related to production (yield of stems) of *Dendrobium* species, are significantly higher in *D. shixingense* and *D. aduncum* than that of *D. officinale*. Therefore, 38 individuals of *D. officinale* and five relatives with resequencing data were used for association analysis. A total of 11 candidate genes, especially for one key gene of *MWL1* were identified. *MWL1* gene is well documented as a key gene involved in secondary cell wall biology, specifically lignin biosynthesis. After the knockout of *MWL1* gene and its closely related gene *MWL2*, the plant height was significantly decreased[83]. These results indicated that *MWL1* gene, may contribute to plant production (yield of stems) of *D. officinale*.

Overall, our newly assembled high-quality genome sequence of *D. officinale* could provide a platform for elucidating genomic evolution of the orchid species and understanding the genes responsible for biosynthesis of active ingredients in *Dendrobium* species as well as laying a foundation for the molecular breeding of *D. officinale*.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (Grant No. 31900268, 31670330 and 32070353), Natural Science Foundation of Jiangsu Province, China (BK20190699), Natural science fund for colleges and universities in Jiangsu Province, China (19KJB180005).

Author contributions

Xiaoyu Ding and Zhitao Niu designed the study. Zhitao Niu, Fei Zhu, Yajuan Fan, Chao Li, Shuying Zhu and Wei Liu performed the experiments. Zhitao Niu, Benhou Zhang, Zhenyu Hou, Mengting Wang, Jiapeng Yang and Qingyun Xue analyzed the data. Zhitao Niu wrote the manuscript. All authors approved the final version of the manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2021.01.019.

References

1. Wood HP. *The dendrobiums*. Ruggell: ARG Gantner Verlag; 2006.
2. Zhu GH, Ji ZH, Wood JJ, Wood HP. *Flora of China*. Beijing: Scientific Press; 2009.
3. Xiang XG, Schuiteman A, Li DZ, Huang WC, Chung SW, Li JW, et al. Molecular systematics of *Dendrobium* (Orchidaceae, Dendrobieae) from mainland Asia based on plastid and nuclear sequences. *Mol Phylogenet Evol* 2013;**69**:950−60.
4. Bao XS, Shun QS, Chen LZ. *The medicinal plants of Dendrobium (Shi-hu) in China*. Shanghai: Shanghai Medicinal University Press and Fudan University Press; 2001.
5. The State Pharmacopoeia Commission of P. R. China. *Pharmacopoeia of the people's Republic of China 2010, set of 3*. Beijing: China Medical Science and Technology Press; 2010.
6. Xu S, Li D, Li J, Xiang X, Jin W, Huang W, et al. Evaluation of the DNA barcodes in *Dendrobium* (Orchidaceae) from mainland Asia. *PLoS One* 2015;**10**:e0115168.
7. Feng S, Jiang Y, Wang S, Jiang M, Chen Z, Ying Q, et al. Molecular identification of *Dendrobium* species (Orchidaceae) based on the DNA barcode ITS2 region and its application for phylogenetic study. *Int J Mol Sci* 2015;**16**:21975−88.
8. Hou B, Tian M, Luo J, Ji Y, Xue Q, Ding X. Genetic diversity assessment and *ex situ* conservation strategy of the endangered *Dendrobium officinale* (Orchidaceae) using new trinucleotide microsatellite markers. *Plant Syst Evol* 2012;**298**:1483−91.
9. Hou BW, Luo J, Zhang YS, Niu ZT, Xue QY, Ding XY. Iteration expansion and regional evolution: phylogeography of *Dendrobium officinale* and four related taxa in southern China. *Sci Rep* 2017;**7**:43525.
10. Zhu S, Niu Z, Xue Q, Wang H, Xie X, Ding X. Accurate authentication of *Dendrobium officinale* and its closely related species by comparative analysis of complete plastomes. *Acta Pharm Sin B* 2018;**8**:969−80.
11. Zhang J, He C, Wu K, Teixeira da Silva JA, Zeng S, Zhang X, et al. Transcriptome analysis of *Dendrobium officinale* and its application to the identification of genes associated with polysaccharide synthesis. *Front Plant Sci* 2016;**7**:5.
12. Yan L, Wang X, Liu H, Tian Y, Lian J, Yang R, et al. The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb. *Mol Plant* 2015;**8**:922−34.
13. Zhang GQ, Xu Q, Bian C, Tsai WC, Yeh CM, Liu KW, et al. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep* 2016;**6**:19029.

14. Chaw SM, Liu YC, Wu YW, Wang HY, Lin CI, Wu CS, et al. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat Plants* 2019;**5**:63−73.
15. Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat Genet* 2019;**51**:1052−9.
16. Xu H, Hou B, Zhang J, Min T, Yuan Y, Niu Z, et al. Detecting adulteration of *Dendrobium officinale* by real-time PCR coupled with ARMS. *Int J Food Sci Tech* 2012;**47**:1695−700.
17. Louwers M, Splinter E, van Driel R, de Laat W, Stam M. Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C). *Nat Protoc* 2009;**4**:1216−29.
18. Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, et al. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat Methods* 2017;**14**:1072−4.
19. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737−46.
20. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014;**9**:e112963.
21. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf* 2018;**19**:460.
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754−60.
23. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* 2013;**31**:1119−25.
24. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;**3**:95−8.
25. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinf* 2012;**13**:238.
26. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2004;**4**:10.
27. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005;**110**:462−7.
28. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;**35**:W265−8.
29. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**:573−80.
30. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res* 2004;**14**:988−95.
31. Haas BJ, Papanicolaou A, Yassour M, Grabherr M. Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;**8**:1494−512.
32. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005;**21**:1859−75.
33. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290−5.
34. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* 2018;**7**:giy093.
35. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* 2006;**7**:S11.
36. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 2004;**20**:2878−9.
37. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf* 2011;**12**:491.
38. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 2005;**33**:W686−9.
39. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007;**35**:3100−8.
40. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;**29**:2933−5.
41. Li L, Stoeckert Jr CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178−89.
42. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772−80.
43. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**:1312−3.
44. Yang Z. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007;**24**:1586−91.
45. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 2017;**34**:1812−9.
46. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* 2013;**30**:1987−97.
47. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**:e49.
48. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29−37.
49. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* 2018;**46**:D493−6.
50. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;**28**:2731−9.
51. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357−9.
52. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491−8.
53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559−75.
54. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;**23**:2633−5.
55. Shen Q, Zhang L, Liao Z, Wang S, Yan T, Shi P, et al. The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family and artemisinin biosynthesis. *Mol Plant* 2018;**11**:776−88.
56. Zhao Q, Yang J, Cui MY, Liu J, Fang Y, Yan M, et al. The reference genome sequence of *Scutellaria baicalensis* provides insights into the evolution of wogonin biosynthesis. *Mol Plant* 2019;**12**:935−50.
57. Tu L, Su P, Zhang Z, Gao L, Wang J, Hu T, et al. Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat Commun* 2020;**11**:971.
58. Cameron KM, Chase MW, Whitten WM, Kores PJ, Jarrell DC, Albert VA, et al. A phylogenetic analysis of the Orchidaceae: evidence from *rbcL* nucleotide sequences. *Am J Bot* 1999;**86**:208−24.
59. Chase MW, Cameron KM, Barrett RL, Freudenstein JV. DNA data and Orchidaceae systematics: a new phylogenetic classification. In: Dixon KW, Kell SP, Barrett RL, Cribb PJ, editors. *Orchid conservation*. Kota Kinabalu: Natural History; 2003. p. 69−89.
60. Cozzolino S, Widmer A. Orchid diversity: an evolutionary consequence of deception?. *Trends Ecol Evol* 2005;**20**:487−94.

61. Givnish TJ, Spalink D, Ames M, Lyon SP, Hunter SJ, Zuluaga A, et al. Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc Biol Sci* 2015;**282**:20151553.
62. Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, et al. The *Apostasia* genome and the evolution of orchids. *Nature* 2017;**549**:379−83.
63. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 2009;**10**:725−32.
64. Bennetzen JL. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 2002;**115**:29−36.
65. Yuan Y, Jin X, Liu J, Zhao X, Zhou J, Wang X, et al. The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat Commun* 2018;**9**:1615.
66. Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* 2015;**47**:65−72.
67. Oliver KR, McComb JA, Greene WK. Transposable elements: powerful contributors to angiosperm evolution and diversity. *Genome Biol Evol* 2013;**5**:1886−901.
68. Warren IA, Naville M, Chalopin D, Levin P, Berger CS, Galiana D, et al. Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. *Chromosome Res* 2015;**23**:505−31.
69. Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc Natl Acad Sci U S A* 2019;**116**:6908−13.
70. Quade J, Cerling TE, Bowman JR. Development of Asian monsoon revealed by marked ecological shift during the latest Miocene in northern Pakistan. *Nature* 1989;**342**:163−6.
71. Zhang L, Chen F, Zhang GQ, Zhang YQ, Niu S, Xiong JS, et al. Origin and mechanism of crassulacean acid metabolism in orchids as implied by comparative transcriptomics and genomics of the carbon fixation pathway. *Plant J* 2016;**86**:175−85.
72. Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, et al. The *Kalanchoë* genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nat Commun* 2017;**8**:1899.
73. Kang M, Wu H, Yang Q, Huang L, Hu Q, Ma T, et al. A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine: an *Isatis* genome. *Hortic Res* 2020;**7**:18.
74. Lv Q, Qiu J, Liu J, Li Z, Zhang W, Wang Q, et al. The *Chimonanthus salicifolius* genome provides insight into magnoliid evolution and flavonoid biosynthesis. *Plant J* 2020;**103**:1910−23.
75. Jiang ZQ, Tu LC, Yang WF, Zhang YF, Hu TY, Ma BW, et al. The chromosome-level reference genome assembly for *Panax notoginseng* and insights into ginsenoside biosynthesis. *Plant Commu* 2021;**2**:100113.
76. Guo X, Li Y, Li C, Luo H, Wang L, Qian J, et al. Analysis of the *Dendrobium officinale* transcriptome reveals putative alkaloid biosynthetic genes and genetic markers. *Gene* 2013;**527**:131−8.
77. He C, Zhang J, Liu X, Zeng S, Wu K, Yu Z, et al. Identification of genes involved in biosynthesis of mannan polysaccharides in *Dendrobium officinale* by RNA-seq analysis. *Plant Mol Biol* 2015;**88**:219−31.
78. Chen K, Liu J, Ji R, Chen T, Zhou X, Yang J, et al. Biogenic synthesis and spatial distribution of endogenous phytohormones and ginsenosides provide insights on their intrinsic relevance in *Panax ginseng*. *Front Plant Sci* 2019;**9**:1951.
79. Ng TB, Liu J, Wong JH, Ye X, Wing Sze SC, Tong Y, et al. Review of research on *Dendrobium*, a prized folk medicine. *Appl Microbiol Biotechnol* 2012;**93**:1795−803.
80. Pu X, Li Z, Tian Y, Gao R, Hao L, Hu Y, et al. The honeysuckle genome provides insight into the molecular mechanism of carotenoid metabolism underlying dynamic flower coloration. *New Phytol* 2020;**227**:930−43.
81. Du X, Huang G, He S, Yang Z, Sun G, Ma X, et al. Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. *Nat Genet* 2018;**50**:796−802.
82. Wang X, Feng H, Chang Y, Ma C, Wang L, Hao X, et al. Population sequencing enhances understanding of tea plant evolution. *Nat Commun* 2020;**11**:4447.
83. Mewalal R, Mizrachi E, Coetzee B, Mansfield SD, Myburg AA. The *Arabidopsis* domain of unknown function 1218 (DUF1218) containing proteins, MODIFYING WALL LIGNIN-1 and 2 (At1g31720/MWL-1 and At4g19370/MWL-2) function redundantly to alter secondary cell wall lignin content. *PLoS One* 2016;**11**:e0150254.