Plant Communications, Volume *2*

Supplemental Information

The chromosome-level reference genome assembly for *Panax notoginseng* and insights into ginsenoside biosynthesis

Zhouqian Jiang, Lichan Tu, Weifei Yang, Yifeng Zhang, Tianyuan Hu, Baowei Ma, Yun Lu, Xiuming Cui, Jie Gao, Xiaoyi Wu, Yuru Tong, Jiawei Zhou, Yadi Song, Yuan Liu, Nan Liu, Luqi Huang, and Wei Gao

Supporting Information Appendix

The Chromosome-level Reference genome Assembly for *Panax notoginseng* and Insights into Ginsenoside Biosynthesis

Zhouqian Jiang[1], Lichan Tu[1,2], Weifei Yang[3], Yifeng Zhang[1,2], Tianyuan Hu[1], Baowei Ma[1], Yun Lu[1], Xiuming Cui[4], Yuru Tong[2], Jie Gao[1], Xiaoyi Wu[1], Jiawei Zhou[1], Yadi Song[1], Yuan Liu[1], Nan Liu[1], Luqi Huang[5], Wei Gao[1,2,6*]

[1]School of Traditional Chinese Medicine, Capital Medical University, Beijing, China

[2]School of Pharmaceutical Sciences, Capital Medical University, Beijing, China

[3]Annoroad Gene Technology, Beijing, China

[4]Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming, China

[5]State Key Laboratory Breeding Base of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

[6]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China

*Corresponding Authors:

Wei Gao: weigao@ccmu.edu.cn

Luqi Huang: huangluqi01@126.com

Tel/Fax: 010-83916572

Table of contents

Supplementary Section S1 - Genome sequencing, assembly and evaluation

1.1 Plant materials

The *P. notoginseng* plant used for genome sequencing were collected from Wenshan County, Yunnan Province, China in August 2019. Fresh and healthy leaves were harvested and immediately frozen on liquid nitrogen after collection, followed by the preservation at -80℃ in the laboratory prior to DNA extraction.

1.2 Estimation of genome size using *K-mer* analysis

To estimation the genome size of *P. notoginseng* by using the *K-mer* analysis, we selected 231.06 Gb pair-end reads and generated the *21-mer* frequency distribution. The distribution of the *21-mer* depends on the characteristics of the genome and follows a Poisson's distribution. We estimated that the genome size was 1.67 Gb. Based on the *21-mer* analysis, we also estimated the heterozygosity ratio and the proportion of repeated sequences in the genome, which were 0.21% and 70.09% respectively (**Supplementary Table 1**). Considering that the higher repetitive sequence of the genome may lead to inaccurate *K-mer* analysis, we chose *K*=35 for a second prediction and finally the genome size was estimated to be 2.35Gb (**Supplementary Figure 1**).

1.3 Library construction and genome sequencing

High-quality genomic DNA was extracted from the leaves using a phenol chloroform isoamyl alcohol extraction method. The quality and quantity of the isolated DNA were separated checked by Nanophotometer® (IMPLEN, CA, USA) and Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). Then the genomic DNA was broken into random fragments. DNA sequencing libraries were constructed according to the standard Illumina library preparation protocols. Paired-end library with insert size of 350 bp was constructed according to the manufacturer's instructions (Illumina, San Diego, CA). The constructed library was sequenced using Illumina HiSeq X Ten Platform by following the standard Illumina protocols. After filtering out the adapter sequences and the low-quality and duplicated reads, we obtained a total of 231.06

Gb(~86.86x) of clean data.

For PacBio libraries (English et al., 2012), we needed at least 10 μg of sheared DNA. The whole genome was sequenced on the PacBio Sequel System (**Supplementary URLs**) based on the single-molecular real-time (SMRT) sequencing technology. The template library was constructed using SMRTbell Template Prep Kit 1.0 (product code 100-259-100) and SMRTbell Damage Repair Kit (product code 100-465-900). Following the procedure described in the PacBio brochure, the high-quality DNA was fragmented and concentrated. The fragments were bead-purified, damage-repaired, and used as the ~20 kb SMRTbell templates. A total of 284.07 Gb (~106.79x) of data were obtained.

DNA from young leaves of the same *P. notoginseng* plant was used to constructed the Hi-C library. Grind the sample with 2% formaldehyde to fix the chromatin. After the cross-linking of the sample was completed, leaf cell lysis was performed and chromatin digestion was performed using *Mbol* endonucleases. After biotin labeling, blunt end linking and DNA purification, Hi-C sample was prepared and entered into the standard library construction process **(Supplementary Figure 2)**. After the constructed libraries were qualified by quality controlling, Illumina HiSeq X Ten was used for sequencing and the sequencing strategy used was PE150. Finally, a total of 340.83 Gb (~128.13x) data was retained **(Supplementary Table 2**).

1.4 Genome assembly

The reads exported by Sequel II™ Systems were quality evaluated with the in-built High-Quality Region Finder (HQRF) which identified the longest high quality regain each read generated by a singly-loaded DNA polymerase according to the ratio of signal to noise (Chakraborty et al., 2016; Hackl et al., 2014). The quality reads obtained were assembled into contigs using Canu (v1.5; **Supplementary URLs**) (Koren et al., 2017). The consensus genome was subjected to a final round of base-error correction (polish) by referring to the Ilumina reads with BWA (v0.7.9a) and Pilon (v1.22; **Supplementary URLs**) (BJ et al., 2014). The total length of this assembly version was 2.66 Gb with a contig N50 size of 1.21 Mb. Then, the Hi-C sequencing data were aligned to the

assembled scaffold by BWA-mem and the contigs were clustered onto chromosomes with LACHESIS (**Supplementary URLs**; **Supplementary Table 3**), the final genome was 2.66 Gb and the contig and scaffold N50 were 1.12 Mb and 216.47 Mb respectively (**Table 1 and Supplementary Table 4**).

1.5 Evaluation of assembly quality

We applied three methods to evaluate the quality of our assembled genome. First, we mapped clean reads from Illumina PE libraries to the genome using BWA mem. The distribution of the sequencing depth at each position was calculated using SAMtools. Nearly 99.82% of the clean reads could be mapped to the assembly genome, which covered 97.97% of the assembled sequence (**Supplementary Table 5**). Second, the completeness of the genome was evaluated with BUSCO (Benchmarking Universal Single-Copy Orthologs, v3.0.1, default parameters; **Supplementary URLs**) (Simao et al., 2015) based on the homologous database. We found 96.6% complete BUSCOs in the *P. notoginseng* genome (**Supplementary Table 6**). Third, the RNA sequencing (RNA-seq) reads of *P. notoginseng* generated in this study were assembled using Trinity, and these samples came from different tissue of different parts of *P. notoginseng* plants. According to the mapping rate (mostly ranging from 94%~97%) of each sample, the assembly had good coverage of the gene regions. Collectively, BUSCO, short-insert size read mapping and transcriptome analysis proved the high quality of the genome assembly, which was adequate enough for subsequent genome analyses in this study.

Supplementary Section S2 - Genome Annotation

2.1 Annotation of repeat sequences

Repetitive sequences are an important part of the genome including two categories, tandem repeat and interspersed repeats. In this study, two strategies were used to predict the repetitive sequences, which were *de novo* approach and homology approach respectively. For the *de novo* approach, RepeatModeler (**Supplementary URLs**) was used in this strategy. *De novo* repeat sequence library was established firstly, and then repeat sequences were predicted by repeatmasker software RepeatScout (**Supplementary URLs**). In addition, the *ab* initio prediction method was also used to find tandem repeat sequences in the genome by the software Tandem Repeats Finder (TRF). For the homology-based approach, it was based on repeated sequence database Repdase. RepeatMasker (version 3.3.0; **Supplementary URLs**) and RepeatProteinMask used to predict sequences similar to known repeat sequences (**Supplementary Table 7**). According to the integrated statistics of the prediction results obtained above, the proportion of repetitive sequences in the *P. notoginseng* genome was 85.85%. The most abundant repetitive element repeat type was LTR, which accounted for 58.88% of the genome (**Supplementary Figure 3 and Supplementary Table 8**).

2.2 Annotation of protein-coding genes

We used homology-based prediction, *de novo* prediction and transcriptome-based prediction to predict the protein-coding genes in the *P. notoginseng* genome. Proteins from the four known species (*Arabidopsis thaliana*, *Daucus carota*, *Panax ginseng*, *P. notoginseng*-pub) were used as homology evidence to search against *P. notoginseng* genome using tblastn (evalue 1e-5), and the gene structure were predicted by GeneWise with default parameter (**Supplementary URLs**; **Supplementary Figure 4**). For the *de novo* prediction, software based on the statistical characteristics of genomic sequence data (such as codon frequency, exon-intron distribution) was used to predict gene structure. The software used in this study contained Augustus, SNAP and GeneMark

(**Supplementary URLs**). To carry out the RNA-Seq aided gene prediction, clean RNA-Seq reads were assembled into transcripts using Trinity, then aligned to our genome assembly and predicted gene structure using PASA (**Supplementary URLs**). Synthesizing the above forecast results, the gene sets predicted by various strategies were integrated into a non-redundant and more complete gene set by EVidenceModeler (EVM; **Supplementary URLs**). Finally, a total of 37,606 genes were predicted from the *P. notoginseng* genome (**Supplementary Table 9**). By predicting the structure of the genes, we also obtained information of gene features such as the distributions of mRNA length, exon length, exon number, intron length and CDS length and so on (**Supplementary Table 10**).

2.3 Functional annotation protein-coding genes

The function annotation of genes is mainly to compare the predicted gene sets with various functional databases, so that can able to understand the function of genes and their role in life activities. The protein database used in this study included Swissprot, NT, NR, PFAM, eggNOG, and GO (**Supplementary URLs**). A total of 36,154 genes were predicted to be functional, accounting for 96.14% of all genes in the *P. notoginseng* genome (**Supplementary Table 11**).

2.4 Annotation of noncoding RNA genes

Noncoding RNA, refers to RNA that can't translate into proteins, such as rRNA, tRNA, snRNA, miRNA and so on, all have important biological functions. miRNA can degrade its target gene or inhibit translation into protein, and play an important role for gene silencing. tRNA and rRNA directly participate in protein synthesis. As well as, snRNA mainly involves in the processing of RNA precursors, which is the important component of RNA shear body. By comparing with known noncoding RNA libraries, Rfam, we can obtain the prediction of rRNA, snRNA, miRNA and so on. The tRNA sequences in genome were predicted by the software tRNAscan-SE (**Supplementary URLs**). Finally, we obtained 14,430 miRNA genes, 1513 tRNA genes, 3018 rRNA genes and 8174 snRNA genes in *P. notoginseng* genome (**Supplementary Table 12**).

Supplementary Section S3 - Evolution and cluster of gene family

3.1 Identification of gene families

Using the OrthoMCL package (Li et al., 2003) (version 1.4), we identified the gene families (clusters) between *P. notoginseng* and seven other plant species, including *P. ginseng*, *D. carota*, *V. vinifera*, *C. annuum*, *G. uralensis*, *A. thaliana*, *O. sativa*. First, the gene set of each species was filtered (**Supplementary Figure 5 and Supplementary Table 13**). If there were multiple alternative splicing transcripts for a gene, only the transcript with the longest coding region was retained for further analysis. Second, in order to ensure the reliability of the encoded protein, genes encoding length less than 50 amino acids were excluded. Then, an all-vs-all BLASTP (version 2.2.28) was performed with an E-value threshold of 1e-5. Finally, clustering was conducted using the Markov cluster algorithm (MCL) integrated in the OrthoMCL package. In total, 27,501 gene families comprising 232,394 genes were identified among these eight plant species and used for subsequent comparative analysis. According to the classification results of gene families, specific gene families within species and gene families shared between species could be found. A total of 1072 gene families containing 2879 genes unique to the *P. notoginseng* genome were found (**Supplementary Figure 6**). To functionally annotate these unique genes, we performed Gene Ontology (GO) and KEGG pathway enrichment analysis by using Fisher's exact test with false discovery rate (FDR) corrections (**Supplementary Figure 7 and Supplementary Table 14**).

3.2 Phylogenetic tree and divergence estimation

After gene family clustering, we aligned all 458 single-copy gene protein sequences by MUSCLE (**Supplementary URLs**) (Edgar, 2004). Then the four-fold degenerate synonymous site (4DTv) were employed to construct phylogenetic trees. PhyML software (Guindon et al., 2010) used the maximum likelihood method (Guindon and Gascuel, 2003) to construct the species phylogenetic tree (ML TREE). 4DTv of genes in each single-copy gene family are often used to estimate the substitution rate and the

divergence time between species. This analysis needed to be added to the phylogenetic tree of the species with the calibration time first (Benton and Donoghue, 2007; Blanc and Wolfe, 2004). According to the supergene sequence integrated in the phylogenetic analysis, the MCMCtree software (**Supplementary URLs**) in the PAML software package (Yang, 2007) was used to estimate the divergence time using the BRMC method (International Brachypodium, 2010; Sanderson, 2003). The MCMCtree running parameters were as follows: burn-in=20,000; sample-frequency=2. *O. sativa* was designated as an outgroup of the phylogenetic tree. The calibration times of the divergence between *O. sativa* and *A. thaliana* (130.7-160.6 MYA), *A. thaliana* and *G. uralensis* (124.3-132.1 MYA), *G. uralensis* and *V. vinifera* (117.8-127.5 MYA), *V. vinifera* and *C. annuum* (109.6-116.3 MYA), *D. carota* and *P. notoginseng* (48.3-70.1 MYA) were obtained from the TimeTree website (**Supplementary URLs**). The divergence time between *P. notoginseng* and *P. ginseng*, *C. annuum* and *D. carota* were estimated to be approximately 4.2 MYA and 91.6 MYA respectively (**Supplementary Figure 8A**).

3.3 Expansion and contraction of gene family

Based on the cluster analysis results of gene families and after filtering gene families with abnormal gene numbers in individual species, we used the CAFÉ program(De Bie et al., 2006) to identify the expansion and contraction of gene families of each species. A random birth and death model were used to study changes in gene families along each lineage of the phylogenetic tree. We used the probabilistic graphical model (PGM) to simulate the gain and loss of genes under the phylogenetic tree and conducted hypothesis testing to analyze the expansion and contraction of gene families. Using conditional likelihoods as the test statistics, we calculated the corresponding p-values in each lineage, and a p-value of 0.05 was used to identify families that were significantly expanded and contracted. Finally, we determined that 989 gene families were expanded and 1823 gene families were contracted (**Supplementary Figure 8B**). By conducting enrichment analysis of GO and KEGG on gene families, results showed that expanded gene families mainly enriched in GO terms such as transposition, fatty

281 acid biosynthetic process, respiratory chain, catalytic activity and so on
282 (**Supplementary Table 15**). Contracted gene families mainly enriched in GO terms
283 (**Supplementary Table 16**) such as protein phosphorylation, protein modification
284 process, beta-glucan biosynthetic process, 1,3-beta-D-glucan synthase complex, purine
285 nucleotide binding and so on (**Supplementary Figure 9**).
286

Supplementary Section S4 - Analysis of whole-genome duplication

4.1 Identification of WGD events of *P. notoginseng*

To further explore the evolution of the *P. notoginseng* genome, we searched for whole genome duplication (WGD) in our assembled *P. notoginseng* genome. WGD events are widespread in the plant genome and are considered to be an important driving force for the evolution of plant genomes. The protein sequences from *P. notoginseng*, *V. vinifera* and *D. carota* were searched against themselves using blastp ($E<1e^{-5}$) to identify syntenic blocks. Then the alignment results were subjected to McscanX (Huang et al., 2009; Schmutz et al., 2010) to determine syntenic blocks. In addition, the protein sequences from *P. notoginseng* were compared with *V. vinifera*, *D. carota* and *P. ginseng*. We calculated the 4DTv (fourfold degenerate synonymous sites of the third codons) for syntenic segments from the concatenated alignments constructed by fourfold degenerate sites of all gene pairs found in each segment and plotted the distribution of the 4DTv values (**Figure 2B**). There were two peaks at approximately 0.16 and 0.50 found in the *P. notoginseng* genome, and the first peak at approximately 0.50 revealed the core eudicot gamma triplication event. The second peak at approximately 0.16 indicated that *P. notoginseng* underwent another WGD event after diverging from *V. vinifera* and *D. carota*.

To verify the above conjecture, we conducted a collinear comparison analysis of the *P. notoginseng* and *V. vinifera* genome. Jcvi was used for identify syntenic blocks and plotted their relationship (**Figure 2C and Supplementary Figure 10**). From the results, we could find that there was a 1:2 collinear relationship between *P. notoginseng* and *V. vinifera* genome (**Supplementary Figure 10**).

4.2 Estimate the timing of the WGD event in *P. notoginseng*

To estimate the timing of the WGD event in *P. notoginseng*, we calculated the *Ks* (synonymous substitution rate) value of the gene pair within and between species using the default parameters of the software wgd (Zwaenepoel and Van de Peer, 2019), and then summed the results and the distribution of the *Ks* values was plotted

(**Supplementary Figure 12**). The results of the *Ks* distribution were consistent with the 4DTv values, and showed a main peak at approximately 0.38, which indicated that a recent WGD event occurred in the *P. notoginseng* genome. Then we calculated the time of WGD event of *P. notoginseng* according to the method reported in the literature (Qin et al., 2014), and summarized the WGD events of each published genome for centralized display (Iorizzo et al., 2016; Tu et al., 2020; Vanneste et al., 2014). The WGD event occurred approximately 29.6 MYA in *P. notoginseng* genome.

Supplementary Section S5 - Analysis of genes related to terpenes biosynthesis pathway

5.1 Identification and phylogenetic analysis of genes

The biosynthesis pathways of terpenoids in plants have been comprehensively explained, and research on *Panax* L. plants has attracted extensive interest from researchers. To identify the terpenoid biosynthesis-related genes in the *P. notoginseng* genome, we used two methods to analyze the genes in 8 species. For genes with corresponding domains in Pfam database such as *CYP450*, *DXR*, *DXS*, *HDR*, *HDS*, *HMGR*, *HMGS*, *MCS*, *MCT*, *MDD*, *PMK*, *SE*, *SS*, *UGT*, we used HMMER (3.1b1) to annotation and searched for each species to obtain copies of genes in different species. For genes where the corresponding domain in the Pfam database was not found, such as *AACT*, *CMK*, *DS*, *FPS*, *GGPPS*, *GPS*, *IPI*, *MVK*, we first downloaded the homologous sequences of genes in different species from NCBI and then compared the sequences by blast (2.2.28) (setting parameter: e value: $1e^{-5}$, coveraged > 50%, identity > 50%; **Supplementary URLs**) to obtain gene copies (**Supplementary Table 17**). After obtaining the gene sequences, we constructed the phylogenetic tree with each gene using the protein sequences in 8 species using MEGA-X (**Supplementary Figure 13-14**), the genetic relationship among the three species *P. notoginseng*, *P. ginseng* and *D. carota* was relatively close.

5.2 Analysis of the duplication of homologous gene pairs

After counting the genes in the terpenoid biosynthetic pathway, we found that most of the genes had multiple copies, so we analyzed the replication time of these multicopy gene pairs. We used the default parameter of wgd software to calculate the Ka value and Ks value of gene pairs, and then converted the Ks value to years (**Supplementary Table 18**). Finally, the results were presented in the form of pictures using Adobe Illustrator (**Supplementary Figure 15**).

Supplementary Section S6 - Transcriptomic analysis and transcription factor regulation

6.1 Sample collections and RNA isolation of tissue transcriptome

One- to four-year-old *P. notoginseng* plants were collected from Wenshan County, Yunnan Province, China. After harvested, we subdivided the plant into different tissue parts, including root (xylem), stem, leaf, flower, rhizome, fibril, periderm, phloem and tubercle (**Supplementary Figure 16-17**). All collected samples were transported by dry ice, washed with ultrapure water three times, immediately frozen on liquid nitrogen and stored at -80 °C prior to RNA extraction. Total RNA for each tissue was extracted using Trizol method. Generally, three biological replicates from each tissue were collected.

6.2 RNA sequencing and assembly

The RNA purity was checked using the kaiaoK5500®Spectrophotometer (Kaiao, Beijing, China) and the RNA integrity and concentration was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Then, the integrate RNA was used in cDNA library construction and Illumina sequencing. The cDNA library was constructed using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB), following the manufacturer's recommendations. After cluster generation, the libraries were sequenced on an Illumina novaseq S2 platform and 150 bp paired-end reads were generated.

In order to guarantee the data quality which was used to analysis, Raw data was filter (**Supplementary Table 19**) with following steps: trim primer sequence from the reads; remove the contaminated reads for adapters; remove the low quality reads; remove the reads whose N base more than 5% for total bases. Bowtie2 v2.2.3 was used for building the genome index, and Clean Data was then aligned to the reference genome using HISAT2 v2.1.0 (**Supplementary Figure 18-19**). The filtered sequences were mapped on the *P. notoginseng* genome and the mapping rate ranged from 90%-96%, indicating a high quality of our genome.

6.3 Transcriptome analysis

ASprofile software was used to analyze and count the alternative splicing events of each sample in this study and rMats to classify and count the alternative splicing events in different groups (**Supplementary Figure 20 and Supplementary Table 20**). We also used Cuffcompare to detect new transcription and discovered some new unknown genes and laid the foundation for a more comprehensive analysis of transcript information. SNP and InDel were detected by Samtools (**Supplementary Figure 21 and Supplementary Table 21**).

Reads Count for each gene in each sample was counted by HTSeq v0.6.0, and FPKM (Fragments Per Kilobase Million Mapped Reads) was then calculated (**Supplementary Figure 22 and 23**). To explore the gene-level regulation of the formation of the root morphological characteristics, we conducted a comparative analysis between the different root groups to screen for differentially expressed genes (DEGs). DESeq2 was employed for differential gene expression analysis between two samples with biological replicates. Genes with $q \leq 0.05$ and $|\log2\_ratio| \geq 1$ were identified as DEGs. The GO and KEGG enrichment of differentially expressed genes were performed and considered to be significantly enriched with $q < 0.05$. After screening the DEGs between the periderm group and tubercle group, and GO enrichment analysis results showed that DEGs, which were highly expressed in tubercle group, were mainly enriched in secondary root formation, terpene catabolic process, shoot axis formation, strigolactone biosynthetic process (**Supplementary Figure 24**), etc. By annotating these DEGs, we found a series of genes related to the biosynthesis of phytohormone (**Supplementary Table 22**), such as the carotenoid cleavage dioxygenase 7 (CCD7) and CCD8 genes involved in the biosynthesis of strigolactone, hydroxylase and dehydrogenase genes related to cytokinin, expansin related genes, etc.

6.4 Analysis of saponin biosynthesis and regulation mechanism

Through comparison with the PlnatTFDB database, we identified a total of 2150 transcription factor genes from the *P. notoginseng* genome, which were classified into 57 subfamilies. Among these subfamilies, bHLH transcription factor, ERF transcription

factor, NAC transcription factor, MYB transcription factor, C2H2 transcription factor, MYB-related transcription factor contained a large number of gene copies (**Supplementary Table 23**). To investigate the role of transcription factors in terpenoid biosynthesis pathway, we studied the correlation between transcription factors and key enzyme genes. We first used R to calculate the Pearson correlation coefficient between transcription factors and genes in batches (set a significant correlation parameter $p<0.05$). Then, we selected the strong correlation gene pair whose correlation coefficient is greater than 0.7 and used Cytoscape software to draw the correlation map (**Figure 2D**). From the correlation map, several subfamilies had a strong correlation include bHLH transcription factor, ERF transcription factor, MYB transcription factor, WRKY transcription factor, indicating that these genes may participate in terpenoid biosynthesis process by regulating the expression of key enzyme genes. In addition, we also studied the temporal (**Figure 3**) and spatial (**Supplementary Figure 25**) expression profiles of saponin pathway genes during the growth and development of *P. notoginseng*, with a view to more fully revealing the production and development of saponins in *P. notoginseng* plants.

Supplementary Section S7 - Analyzing key enzyme genes of ginsenosides pathway and functional verification of UGT genes

7.1 Phylogenetic analysis of UGT and CYP450 genes

By comparison with the Pfam database, we identified 336 CYP450 genes and 158 UGT genes from the *P. notoginseng* genome. Then we downloaded the gene sequences of each subfamily from NCBI (**Supplementary Table 25-26**), used MEGA-X software to construct the phylogenetic tree, and modified the evolution trees on the online website iTOL (**Supplementary Figure 26 and Figure 3A**) (Letunic and Bork, 2019).

7.2 Function verification of candidate UGT genes

7.2.1 Gene cloning and expression vector construction

After sampling, *P. notoginseng* plants were frozen immediately in liquid nitrogen and

ground into powder for isolation of total RNA using Trizol (Invitrogen, Carlsbad, CA, USA) as the manufacture's instruction and then converted into cDNA using PrimeScript® RT reagent Kit with gDNA eraser (Takara, Dalian, China). Among the gene sequences obtained through systematic evolution and homologous alignment, we designed primers and cloned 32 open reading frames of UGT genes (**Supplementary Figure 27 and Supplementary Table 27**), the cloned open reading frames (ORFs) of UGT genes were inserted into *pEASY*®-Blunt Cloning Vector (TransGen Biotech, Beijing, China) independently. After the cloned gene was sequenced successfully, we connected them to the expression vector HIS-MBP-PreSc-pET28a (Li et al., 2018b) using Seamless Cloning Kit (Beyotime, Shanghai, China) as the manufacture's instruction.

7.2.2 Induced protein expression and functional verification

After successful construction, the expression vector was transformed into *E. coli* BL21 (DE3) (Transgen Biotech, Beijing, Chain), and the recombinant *E. coli* BL21 (DE3) strain was cultured in LB medium (with 50μg/mL kanamycin) at 37 ℃ at 200 rpm until the OD_{600} reached 0.6-0.8. Cool the bacterial solution on ice and add IPTG to a final concentration of 50 μM. After incubation at 16 ℃ at 120 rpm for 16h, the cells were harvested by centrifugation at 4 ℃ and suspended in 100 mM phosphate buffer (pH 8.0), 1mM PMSF. The resuspension solution was disrupted by ultrasonication and the mixture was centrifuged at 4 ℃ at 12000 *g* for 20 min, so that protein and cell debris were successfully separated. The supernatant was used for enzymatic assays. The pET28a-transformed *E. coli* BL21 (DE3) cells were treated in parallel as a control. Next, we checked whether the vectors expressed protein by SDS-PAGE protein electrophoresis, and used the crude enzyme to carry out the enzymatic reaction of glycosylation. Generally, the reaction was carried out in a 100 μL volume containing 100 mM crude enzyme buffer (pH 8.0), 1mM UDP-glucose, 0.1 mM acceptor substrate for 2h in a 35 ℃ water bath and was terminated by adding 100 μL methanol. At first, we used PPD and PPT as substrates and then use monoglucosides such as Rh_2, F_1 as substrates to verify whether UGT genes have catalytic functions. The mixed solution

was allowed to stand overnight at 4 ℃. The extraction was passed through a 0.22 μM organic filter membrane, and the resulting solution was tested by UPLC/Q-TOF-MS (ultrahigh-performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry) (**Supplementary Figure 28-31**). The following Q-TOF-MS parameters were used: the experiment was performed in the ESI (-) ionization mode; scan range, 50-1500 Da; scan time, 0.2 s; cone voltage, 40 V; source temperature, 100 ℃; dissolved gas temperature, 450 ℃; cone gas flow rate, 50 L/h; desolvation flow rate, 900 L/h; collision energy, 20-50V. The mass accuracy was corrected by a lock spray with leucine enkephalin (200 pg/μL, 10 μL/min) as the reference (m/z 556.2766 ESI (+) and 554.2620 ESI (–)).

The UPLC separation was performed using an Agilent Technologies 1290 Infinity II system (Agilent Technologies, Santa Clara, CA, USA) with a Waters ACQUITY UPLC HSS T3 analytical column (2.1 mm x 100 mm, 1.8 μm) kept at 35 ℃. The mobile phases were a mixture of 0.1% (v/v) acetic acid in water (A) and acetonitrile (B), and the flow rate was 0.3 mL/min. The gradient elution was programmed as follows: 0–2.0 min, 20–28% B; 2.0–3.0 min, 28–36% B; 3.0–10.0 min, 36–40% B; 10.0–15.0 min, 40–64% B; 15.0–17.0 min, 64–90% B; 17.0–22.0 min, 90–20% B. The injection volume was 1 μL for each sample.

7.3 Screen for candidate UGT genes involved in the saponin biosynthetic pathway

7.3.1 WGCNA analysis of UGT genes

We used the WGCNA software package in the R to perform the analysis on the genes annotated as glucosyltransferases or glycosyltransferases and saponins pathway genes in *P. notoginseng* genome (**Supplementary Figure 33**). First of all, we sorted and filtered the genes expression data, genes with a variance of 0 in the expression between different samples were filtered out. In addition, genes with a gene expression level of 0 that exceeded 10% of the total number of samples were also filtered out. Based on the filtered data, the hierarchical clustering was used to draw the sample tree, and the relationship between different samples could be seen from the dendrogram. Then we used the pickSoftThreshold function to calculate the soft threshold (β value). From the

result graph, we can see that when β value was 10, the correlation threshold was the highest. Finally, we used the function blockwiseModules to construct the modules present in these UGT genes, and obtained a total of 7 different modules **(Supplementary Table 28)**. In addition, the genes in the saponin biosynthesis pathway were mainly concentrated in the four modules: green, turquoise, red and brown.

7.3.2 Identification and expression profiling of genes related to terpenoids biosynthesis

Through the transcriptome data, we obtained the expression levels of the genes related to terpenoid biosynthesis in each group of samples, and used MEV software to draw the heat maps (**Supplementary Figure 34**). According to their expression patterns in different samples, these genes could be divided into three categories: most genes were highly expressed in flowers; some gens were highly expressed in various parts of roots; only a small part of genes were highly expressed in leaves. On this basis, we also screened a series of candidate UGT genes by comparing the expression patterns of pathway genes and annotated UGT genes **(Supplementary Figure 35 and Supplementary Table 29)**.

Supplementary URLs

PacBio Sequel System: https://www.pacb.com/products-and-services/pacbio-systems/sequel/

Canu: https://github.com/marbl/canu

Blasr: https://github.com/ PacificBiosciences/blasr

Smrt Link: https://downloads.pacbcloud.com/public/software/installers/smrtlink_5.0.1.9585.zip

Pilon: https://github.com/broadinstitute/pilon

LACHESIS: http://shendurelab.github.io/LACHESIS/

BUSCO: http://busco.ezlab.org/

RepeatModeler: http://www.repeatmasker.org/RepeatModeler/

RepeatScout: http://www.repeatmasker.org/

RepBase: https://www.girinst.org/server/RepBase/index.php

RepeatMasker: http://www.repeatmasker.org/

RepeatProteinMask: http://www.repeatmasker.org/

Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi

Genewise: http://www.ebi.ac.uk/~birney/wise2

Augustus: http://augustus.gobics.de/

SNAP: https://github.com/KorfLab/SNAP

GeneMark: http://exon.gatech.edu/GeneMark/

PASA: http://pasa.sourceforge.net/

EVidenceModeler: http://evidencemodeler.github.io/

Swissprot: https://web.expay.org/docs/swiss-prot_guideline.html

NT: https://www.ncbi.nlm.nih.gov/nucleotide/

NR: ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

PFAM: http://xfam.org/

eggNOG: http://eggnogdb.embl.de/

GO: http://geneontology.org/page/go-database

KEGG: http://www.kegg.jp/

Rfam: http://rfam.xfam.org/

552 tRNAsacn-SE: http://lowelab.ucsc.edu/tRNAscan-SE/

553 Mummer 4.0: https://github.com/mummer4/mummer

554 MUSCLE: http://www.drive5.com/muscle/

555 ENSEMBL database: http://www.ensembl.org/index.html

556 PlantTFDB: planttfdb.cbi.pku.edu.cn/ iTOL: https://itol.embl.de/itol.cgi

557 MCMCtree: http://abacus.gene.ucl.ac.uk/software/paml.html

558 TimeTree: http://www.time.org/

559

References

Aoki, K., Yano, K., Suzuki, A., Kawamura, S., Sakurai, N., Suda, K., Kurabayashi, A., Suzuki, T., Tsugane, T., Watanabe, M., et al. (2010). Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. BMC GENOMICS 11:210.

Augustin, M.M., Ruzicka, D.R., Shukla, A.K., Augustin, J.M., Starks, C.M., O'Neil-Johnson, M., McKain, M.R., Evans, B.S., Barrett, M.D., Smithson, A., et al. (2015). Elucidating steroid alkaloid biosynthesis in *Veratrum californicum*: production of verazine in Sf9 cells. Plant J 82:991-1003.

Bak, S., Kahn, R., Nielsen, H., Moller, B., and Halkier, B. (1998). Cloning of three A-type cytochromes p450, CYP71E1, CYP98, and CYP99 from *Sorghum bicolor* (L.) Moench by a PCR approach and identification by expression in *Escherichia coli* of CYP71E1 as a multifunctional cytochrome p450 in the biosynthesis of the cyanogenic glucoside dhurrin. Plant Mol Biol 36:393-405.

Benton, M.J., and Donoghue, P.C. (2007). Paleontological evidence to date the tree of life. Mol Biol Evol 24:26-53.

BJ, W., T, A., T, S., M, P., A, A., S, S., A., C.C., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9:e112963.

Blanc, G., and Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678.

Brazier-Hicks, M., and Edwards, R. (2005). Functional importance of the family 1 glucosyltransferase UGT72B1 in the metabolism of xenobiotics in Arabidopsis thaliana. Plant J 42:556-566.

Chakraborty, M., Baldwin-Brown, J.G., Long, A.D., and Emerson, J.J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res 44:e147.

Chen, B., Chen, J., Du, Q., Zhou, D., Wang, L., Xie, J., Li, Y., and Zhang, D. (2018). Genetic variants in microRNA biogenesis genes as novel indicators for secondary growth in Populus. New Phytol 219:1263-1282.

Chen, H., Wu, B., Nelson, D.R., Wu, K., and Liu, C. (2014). Computational Identification and Systematic Classification of Novel Cytochrome P450 Genes in *Salvia miltiorrhiza*. PLOS ONE 9:e115149.

Chen, H.Y., and Li, X. (2017). Identification of a residue responsible for UDP-sugar donor selectivity of a dihydroxybenzoic acid glycosyltransferase from Arabidopsis natural accessions. Plant J 89:195-203.

Courtney, K.J., Percival, F.W., Hallaban, D.L., Christoffersen, R.E., and Hallahan, D.L. (1996). The Electronic Plant Gene Register. Plant Physiol 112:445-446.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269-1271.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Eljounaidi, K., Cankar, K., Comino, C., Moglia, A., Hehn, A., Bourgaud, F.,

Bouwmeester, H., Menin, B., Lanteri, S., and Beekwilder, J. (2014). Cytochrome P450s from Cynara cardunculus L. CYP71AV9 and CYP71BL5, catalyze distinct hydroxylations in the sesquiterpene lactone biosynthetic pathway. Plant Science 223:59-68.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., et al. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7:e47768.

Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J., Liu, Y., Hu, X., et al. (2002). Sequence and analysis of rice chromosome 4. Nature 420:316-320.

Fukushima, E.O., Seki, H., Ohyama, K., Ono, E., Umemoto, N., Mizutani, M., Saito, K., and Muranaka, T. (2011). CYP716A Subfamily Members are Multifunctional Oxidases in Triterpenoid Biosynthesis. Plant Cell Physiol 52:2050-2061.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-321.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696-704.

Guo, J., Ma, X., Cai, Y., Ma, Y., Zhan, Z., Zhou, Y.J., Liu, W., Guan, M., Yang, J., Cui, G., et al. (2016a). Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. New Phytol 210:525-534.

Guo, J., Zhou, Y.J., Hillwig, M.L., Shen, Y., Yang, L., Wang, Y., Zhang, X., Liu, W., Peters, R.J., Chen, X., et al. (2013). CYP76AH1 catalyzes turnover of miltiradiene in tanshinones biosynthesis and enables heterologous production of ferruginol in yeasts. Proc Natl Acad Sci U S A 110:12108-12113.

Guo, L., Yang, R., and Gu, Z. (2016b). Cloning of genes related to aliphatic glucosinolate metabolism and the mechanism of sulforaphane accumulation in broccoli sprouts under jasmonic acid treatment. J Sci Food Agr 96:4329-4336.

Hackl, T., Hedrich, R., Schultz, J., and Forster, F. (2014). proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30:3004-3011.

Han, J.Y., Hwang, H.S., Choi, S.W., Kim, H.J., and Choi, Y.E. (2012). Cytochrome P450 CYP716A53v2 catalyzes the formation of protopanaxatriol from protopanaxadiol during ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Physiol 53:1535-1545.

Han, J.Y., Kim, H.J., Kwon, Y.S., and Choi, Y.E. (2011). The Cyt P450 enzyme CYP716A47 catalyzes the formation of protopanaxadiol from dammarenediol-II during ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Physiol 52:2062-2073.

Hansen, E., Osmani, S., Kristensen, C., Moller, B., and Hansen, J. (2009). Substrate specificities of family 1 UGTs gained by domain swapping. Phytochemistry 70:473-482.

Hou, B., Lim, E., Higgins, G., and Bowles, D.J. (2004). N-glucosylation of cytokinins

by glycosyltransferases of Arabidopsis thaliana. J Biol Chem 279:47822-47832.
Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476-481.
Huang, L., Li, J., Ye, H., Li, C., Wang, H., Liu, B., and Zhang, Y. (2012). Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in *Catharanthus roseus*. Planta 236:1571-1581.
Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., et al. (2009). The genome of the cucumber, *Cucumis sativus* L. Nature Genetics 41:1275-1281.
International Brachypodium, I. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763-768.
International Rice Genome Sequencing, P. (2005). The map-based sequence of the rice genome. Nature 436:793-800.
Iorizzo, M., Ellison, S., Senalik, D., Zeng, P., Satapoomin, P., Huang, J., Bowman, M., Iovene, M., Sanseverino, W., Cavagnaro, P., et al. (2016). A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. Nat Genet 48:657-666.
Jia, H.M., Jia, H.J., Cai, Q.L., Wang, Y., Zhao, H.B., Yang, W.F., Wang, G.Y., Li, Y.H., Zhan, D.L., Shen, Y.T., et al. (2019). The red bayberry genome and genetic basis of sex determination. Plant Biotechnol J 17:397-409.
Jones, P., Messner, B., Nakajima, J.-I., Schäffner, A.R., and Saito, K. (2003). UGT73C6 and UGT78D1, Glycosyltransferases Involved in Flavonol Glycoside Biosynthesis in *Arabidopsis thaliana*. J Biol Chem 278:43910-43918.
Jung, S.-C., Kim, W., Park, S.C., Jeong, J., Park, M.K., Lim, S., Lee, Y., Im, W.-T., Lee, J.H., Choi, G., et al. (2014). Two Ginseng UDP-Glycosyltransferases Synthesize Ginsenoside Rg3 and Rd. Plant Cell Physiol 55:2177-2188.
Kahn, R., Bak, S., Svendsen, I., Halkier, B., and Moller, B. (1997). Isolation and reconstitution of cytochrome P450ox and in vitro reconstitution of the entire biosynthetic pathway of the cyanogenic glucoside dhurrin from sorghum. Plant Physiol 115:1661-1670.
Kim, B.-G., Sung, S.H., and Ahn, J.-H. (2012). Biological synthesis of quercetin 3-O-N-acetylglucosamine conjugate using engineered *Escherichia coli* expressing UGT78D2. Appl Microbiol Biot 93:2447-2453.
Knoch, E., Sugawara, S., Mori, T., Nakabayashi, R., Saito, K., and Yonekura-Sakakibara, K. (2017). UGT79B31 is responsible for the final modification step of pollen-specific flavonoid biosynthesis in *Petunia hybrida*. Planta 247:779-790.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722-736.
Kovinich, N., Saleem, A., Arnason, J.T., and Miki, B. (2010). Functional characterization of a UDP-glucose:flavonoid 3-O-glucosyltransferase from the

692 seed coat of black soybean (*Glycine max* (L.) Merr.). Phytochemistry 71:1253-
693 1263.
694 KRAUS, P., and KUTCHAN, T. (1995). Molecular-cloning and heterologous
695 expression of a cdna-encoding berbamunine synthase, a c-o phenol-coupling
696 cytochrome-p450 from the higher-plant berberis-stolonifera. P Natl Acad Sci
697 USA 92:2071-2075.
698 Krokida, A., Delis, C., Geisler, K., Garagounis, C., Tsikou, D., Peña-Rodríguez, L.M.,
699 Katsarou, D., Field, B., Osbourn, A.E., and Papadopoulou, K.K. (2013). A
700 metabolic gene cluster in *Lotus japonicusd* iscloses novel enzyme functions and
701 products in triterpene biosynthesis. New Phytol 200:675-690.
702 Lanot, A., Hodge, D., Jackson, R.G., George, G.L., Elias, L., Lim, E.K., Vaistij, F.E.,
703 and Bowles, D.J. (2006). The glucosyltransferase UGT72E2 is responsible for
704 monolignol 4-O-glucoside production in *Arabidopsis thaliana*. Plant J 48:286-
705 295.
706 Lanot, A., Hodge, D., Lim, E.K., Vaistij, F.E., and Bowles, D.J. (2008). Redirection of
707 flux through the phenylpropanoid pathway by increased glucosylation of
708 soluble intermediates. Planta 228:609-616.
709 Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and
710 new developments. Nucleic Acids Res 47:W256-W259.
711 Li, L., Cheng, H., Gai, J., and Yu, D. (2007). Genome-wide identification and
712 characterization of putative cytochrome P450 genes in the model legume
713 *Medicago truncatula*. Planta 226:109-123.
714 Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog
715 groups for eukaryotic genomes. Genome Res 13:2178-2189.
716 Li, Q., Yu, H.M., Meng, X.F., Lin, J.S., Li, Y.J., and Hou, B.K. (2018a). Ectopic
717 expression of glycosyltransferase UGT76E11 increases flavonoid accumulation
718 and enhances abiotic stress tolerance in Arabidopsis. Plant Biol (Stuttg) 20:10-
719 19.
720 Li, Q.S., Lin, X.M., Qiao, R.Y., Zheng, X.Q., Lu, J.L., Ye, J.H., and Liang, Y.R. (2017).
721 Effect of fluoride treatment on gene expression in tea plant (*Camellia sinensis*).
722 Sci Rep 7:9847.
723 Li, X.J., Chen, X.J., Guo, X., Yin, L.L., Ahammed, G.J., Xu, C.J., Chen, K.S., Liu, C.C.,
724 Xia, X.J., Shi, K., et al. (2016). DWARF overexpression induces alteration in
725 phytohormone homeostasis, development, architecture and carotenoid
726 accumulation in tomato. Plant Biotechnol J 14:1021-1033.
727 Li, Y., Lin, H.X., Wang, J., Yang, J., Lai, C.J., Wang, X., Ma, B.W., Tang, J.F., Li, Y.,
728 Li, X.L., et al. (2018b). Glucosyltransferase Capable of Catalyzing the Last Step
729 in Neoandrographolide Biosynthesis. Org Lett 20:5999-6002.
730 Lim, C.E., Ahn, J.-H., and Lim, J. (2006). Molecular genetic analysis of tandernly
731 located glycosyltransferase genes, UGT73B1, UGT73B2, and UGT73B3, in
732 Arabidopsis thaliana. J Plant Biol 49:309-314.
733 Lim, E., Higgins, G., Li, Y., and Bowles, D.J. (2003). Regioselectivity of glucosylation
734 of caffeic acid by a UDP-glucose : glucosyltransferase is maintained in planta.
735 Biochem J 373:987-992.

Lim, E.K., Jackson, R.G., and Bowles, D.J. (2005). Identification and characterisation of Arabidopsis glycosyltransferases capable of glucosylating coniferyl aldehyde and sinapyl aldehyde. FEBS Lett 579:2802-2806.

Lin, J.S., Huang, X.X., Li, Q., Cao, Y., Bao, Y., Meng, X.F., Li, Y.J., Fu, C., and Hou, B.K. (2016). UDP-glycosyltransferase 72B1 catalyzes the glucose conjugation of monolignols and is essential for the normal cell wall lignification in *Arabidopsis thaliana*. Plant J 88:26-42.

Lin, X., Kaul, S., Rounsley, S., Shea, T., Benito, M., Town, C., Fujii, C., Mason, T., Bowman, C., Barnstead, M., et al. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761-+.

Liu, X., Liu, Y., Huang, P., Ma, Y., Qing, Z., Tang, Q., Cao, H., Cheng, P., Zheng, Y., Yuan, Z., et al. (2017). The Genome of Medicinal Plant *Macleaya cordata* Provides New Insights into Benzylisoquinoline Alkaloids Metabolism. Molecular Plant 10:975-989.

Mao, H., Liu, J., Ren, F., Peters, R.J., and Wang, Q. (2016). Characterization of CYP71Z18 indicates a role in maize zealexin biosynthesis. Phytochemistry 121:4-10.

Mayer, K., and Schuller, C., and Wambutt, R., and Murphy, G., and Volckaert, G., and Pohl, T., and Dusterhoft, A., and Stiekema, W., and Entian, K., and Terryn, N., et al. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature 402:769-+.

Mazel, A., and Levine, A. (2002). Induction of glucosyltransferase transcription and activity during superoxide-dependent cell death in Arabidopsis plants. Plant Physiol Bioch 40:133-140.

Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., Woittiez, L., van der Krol, S., Lugan, R., Ilc, T., et al. (2014). The seco-iridoid pathway from *Catharanthus roseus*. Nat Commun 5:3606.

Morant, M., Schoch, G.A., Ullmann, P., Ertunc, T., Little, D., Olsen, C.E., Petersen, M., Negrel, J., and Werck-Reichhart, D. (2007). Catalytic activity, duplication and evolution of the CYP98 cytochrome P450 family in wheat. Plant Mol Biol 63:1-19.

Moses, T., Pollier, J., Almagro, L., Buyst, D., Van Montagu, M., Pedreno, M.A., Martins, J.C., Thevelein, J.M., and Goossens, A. (2014). Combinatorial biosynthesis of sapogenins and saponins in *Saccharomyces cerevisiae* using a C-16 alpha hydroxylase from *Bupleurum falcatum*. Proc Natl Acad Sci U S A 111:1634-1639.

Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., III, Cornejo, O., Findley, S.D., Zheng, P., Utro, F., Royaert, S., et al. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. GENOME BIOLOGY 14:r53.

Nomura, T., Kushiro, T., Yokota, T., Kamiya, Y., Bishop, G., and Yamaguchi, S. (2005). The last reaction producing brassinolide is catalyzed by cytochrome P-450s, CYP85A3 in tomato and CYP85A2 in Arabidopsis. J Biol Chem 280:17873-17879.

Ohnishi, T., Nomura, T., Watanabe, B., Ohta, D., Yokota, T., Miyagawa, H., Sakata, K., and Mizutani, M. (2006). Tomato cytochrome P450CYP734A7 functions in brassinosteroid catabolism. Phytochemistry 67:1895-1906.

Ono, E., Ruike, M., Iwashina, T., Nomoto, K., and Fukui, Y. (2010). Co-pigmentation and flavonoid glycosyltransferases in blue *Veronica persica* flowers. Phytochemistry 71:726-735.

Oudin, A., Hamdi, S.d., Ouélhazi, L., Chénieux, J.-C., Rideau, M., and Clastre, M. (1999). Induction of a novel cytochrome P450 (CYP96 family) in periwinkle (*Catharanthus roseus*) cells induced for terpenoid indole alkaloid production. Plant Science 149:105-113.

Overkamp, S., Hein, F., and Barz, W. (2000). Cloning and characterization of eight cytochrome P450 cDNAs from chickpea (*Cicer arietinum* L.) cell suspension cultures. Plant Science 155:101-108.

Pham, T., Chen, H., Dai, L., and Vu, T.Q.T. (2016). Isolation a P450 gene in *Pinus armandi* and its expression after inoculation of Leptographium qinlingensis and treatment with methyl jasmonate. Russ J Plant Physl+ 63:111-118.

Priest, D.M., Ambrose, S.J., Vaistij, F.E., Elias, L., Higgins, G.S., Ross, A.R., Abrams, S.R., and Bowles, D.J. (2006). Use of the glucosyltransferase UGT71B6 to disturb abscisic acid homeostasis in *Arabidopsis thaliana*. Plant J 46:492-502.

Qi, X., Bakht, S., Qin, B., Leggett, M., Hemmings, A., Mellon, F., Eagles, J., Werck-Reichhart, D., Schaller, H., Lesot, A., et al. (2006). A different function for a member of an ancient and highly conserved cytochrome P450 family: From essential sterols to plant defense. P Natl Acad Sci USA 103:18848-18853.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y., et al. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proc Natl Acad Sci U S A 111:5135-5140.

Ro, D., Arimura, G., Lau, S., Piers, E., and Bohlmann, J. (2005). Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. P Natl Acad Sci USA 102:8060-8065.

Roach, M.J., Johnson, D.L., Bohlmann, J., van Vuuren, H.J.J., Jones, S.J.M., Pretorius, I.S., Schmidt, S.A., and Borneman, A.R. (2018). Population sequencing reveals clonal diversity and ancestral inbreeding in the grapevine cultivar Chardonnay. PLoS Genet 14:e1007807.

Rojas Rodas, F., Rodriguez, T.O., Murai, Y., Iwashina, T., Sugawara, S., Suzuki, M., Nakabayashi, R., Yonekura-Sakakibara, K., Saito, K., Kitajima, J., et al. (2014). Linkage mapping, molecular cloning and functional analysis of soybean gene Fg2 encoding flavonol 3-O-glucoside (1 --> 6) rhamnosyltransferase. Plant Mol Biol 84:287-300.

Saito, S., Hirai, N., Matsumoto, C., Ohigashi, H., Ohta, D., Sakata, K., and Mizutani, M. (2004). Arabidopsis CYP707As encode (+)-abscisic acid 8'-hydroxylase, a key enzyme in the oxidative catabolism of abscisic acid. Plant Physiol 134:1439-1449.

Salanoubat, M., and Lemcke, K., and Rieger, M., and Ansorge, W., and Unseld, M., and Fartmann, B., and Valle, G., and Blocker, H., and Perez-Alonso, M., and Obermaier, B., et al. (2000). Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. Nature 408:820-822.

Salim, V., Wiens, B., Masada-Atsumi, S., Yu, F., and De Luca, V. (2014). 7-Deoxyloganetic acid synthase catalyzes a key 3 step oxidation to form 7-deoxyloganetic acid in *Catharanthus roseus* iridoid biosynthesis. Phytochemistry 101:23-31.

Sanderson, M. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301-302.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature 463:178-183.

Schopfer, C., and Ebel, J. (1998). Identification of elicitor-induced cytochrome P450s of soybean (*Glycine max* L.) using differential display of mRNA. Molecular and general genetics 258:315-322.

Seki, H., Ohyama, K., Sawai, S., Mizutani, M., Ohnishi, T., Sudo, H., Akashi, T., Aoki, T., Saito, K., and Muranaka, T. (2008). Licorice beta-amyrin 11-oxidase, a cytochrome P450 with a key role in the biosynthesis of the triterpene sweetener glycyrrhizin. P Natl Acad Sci USA 105:14204-14209.

Seki, H., Sawai, S., Ohyama, K., Mizutani, M., Ohnishi, T., Sudo, H., Fukushima, E.O., Akashi, T., Aoki, T., Saito, K., et al. (2011). Triterpene functional genomics in licorice for identification of CYP72A154 involved in the biosynthesis of glycyrrhizin. Plant Cell 23:4112-4123.

Shaokang Di, F.Y., Felipe Rojas Rodas, Tito O Rodriguez, Yoshinori Murai, Tsukasa Iwashina, Satoko Sugawara, Tetsuya Mori, Ryo Nakabayashi, Keiko Yonekura-Sakakibara, Kazuki Saito, Ryoji Takahashi. (2015). Linkage mapping, molecular cloning and functional analysis of soybean gene Fg3 encoding flavonol 3-O-glucoside/galactoside (1 → 2) glucosyltransferase. BMC Plant Biol 15:126.

Shi, P., Fu, X., Shen, Q., Liu, M., Pan, Q., Tang, Y., Jiang, W., Lv, Z., Yan, T., Ma, Y., et al. (2018). The roles ofAaMIXTA1in regulating the initiation of glandular trichomes and cuticle biosynthesis in *Artemisia annua*. New Phytol 217:261-276.

Shibuya, M., Nishimura, K., Yasuyama, N., and Ebizuka, Y. (2010). Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in *Glycine max*. FEBS Lett 584:2258-2264.

Simao, 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 31:3210-3212.

Steele, C., Gijzen, M., Qutob, D., and Dixon, R. (1999). Molecular characterization of the enzyme catalyzing the aryl migration reaction of isoflavonoid biosynthesis in soybean. Arch Biochem Biophys 367:146-150.

Stigliani, A.L., Giorio, G., and D'Ambrosio, C. (2011). Characterization of P450 Carotenoid β- and ε-Hydroxylases of Tomato and Transcriptional Regulation of Xanthophyll Biosynthesis in Root, Leaf, Petal and Fruit. Plant Cell Physiol 52:851-865.

Sun, Y., Chen, Z., Li, J., Li, J., Lv, H., Yang, J., Li, W., Xie, D., Xiong, Z., Zhang, P., et al. (2018). Diterpenoid UDP-Glycosyltransferases from Chinese Sweet Tea and Ashitaba Complete the Biosynthesis of Rubusoside. Mol Plant 11:1308-1311.

Sun, Y., Ji, K., Liang, B., Du, Y., Jiang, L., Wang, J., Kai, W., Zhang, Y., Zhai, X., Chen, P., et al. (2017). Suppressing ABA uridine diphosphate glucosyltransferase (SlUGT75C1) alters fruit ripening and the stress response in tomato. Plant J 91:574-589.

Tabata, S., and Kaneko, T., and Nakamura, Y., and Kotani, H., and Kato, T., and Asamizu, E., and Miyajima, N., and Sasamoto, S., and Kimura, T., and Hosouchi, T., et al. (2000). Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature 408:823-826.

Tamura, K., Seki, H., Suzuki, H., Kojoma, M., Saito, K., and Muranaka, T. (2016). CYP716A179 functions as a triterpene C-28 oxidase in tissue-cultured stolons of *Glycyrrhiza uralensis*. Plant Cell Rep 36:437-445.

Tamura, K., Teranishi, Y., Ueda, S., Suzuki, H., Kawano, N., Yoshimatsu, K., Saito, K., Kawahara, N., Muranaka, T., and Seki, H. (2017). Cytochrome P450 Monooxygenase CYP716A141 is a Unique beta-Amyrin C-16beta Oxidase Involved in Triterpenoid Saponin Biosynthesis in *Platycodon grandiflorus*. Plant Cell Physiol 58:874-884.

Tanaka, K., Hayashi, K.-i., Natsume, M., Kamiya, Y., Sakakibara, H., Kawaide, H., and Kasahara, H. (2014). UGT74D1 Catalyzes the Glucosylation of 2-Oxindole-3-Acetic Acid in the Auxin Metabolic Pathway in Arabidopsis. Plant Cell Physiol 55:218-228.

Theologis, A., Ecker, J., Palm, C., Federspiel, N., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C., et al. (2000). Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. Nature 408:816-820.

Trikka, F.A., Nikolaidis, A., Ignea, C., Tsaballa, A., Tziveleka, L.-A., Ioannou, E., Roussis, V., Stea, E.A., Božić, D., Argiriou, A., et al. (2015). Combined metabolome and transcriptome profiling provides new insights into diterpene biosynthesis in *S. pomifera* glandular trichomes. BMC Genomics 16.

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T., Zhou, J., Zhang, Y., Zhao, Y., Liu, Y., et al. (2020). Genome of *Tripterygium wilfordii* and identification of cytochrome P450 involved in triptolide biosynthesis. Nat Commun 11:971.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res 24:1334-1347.

Vasav, A.P., and Barvkar, V.T. (2019). Phylogenomic analysis of cytochrome P450 multigene family and their differential expression analysis in *Solanum lycopersicum* L. suggested tissue specific promoters. BMC Genomics 20:116.

Wang, P., Wei, Y., Fan, Y., Liu, Q., Wei, W., Yang, C., Zhang, L., Zhao, G., Yue, J., Yan,

X., et al. (2015). Production of bioactive ginsenosides Rh2 and Rg3 by metabolically engineered yeasts. Metab Eng 29:97-105.

Wang, S., Wang, R., Liu, T., Lv, C., Liang, J., Kang, C., Zhou, L., Guo, J., Cui, G., Zhang, Y., et al. (2019). CYP76B74 Catalyzes the 3 "-Hydroxylation of Geranylhydroquinone in Shikonin Biosynthesis. Plant Physiol 179:402-414.

Wei, W., Wang, P., Wei, Y., Liu, Q., Yang, C., Zhao, G., Yue, J., Yan, X., and Zhou, Z. (2015). Characterization of *Panax ginseng* UDP-Glycosyltransferases Catalyzing Protopanaxatriol and Biosyntheses of Bioactive Ginsenosides F1 and Rh1 in Metabolically Engineered Yeasts. Mol Plant 8:1412-1424.

Weng, J.-K., Li, Y., Mo, H., and Chapple, C. (2012). Assembly of an Evolutionarily New Pathway for alpha-Pyrone Biosynthesis in Arabidopsis. Science 337:960-964.

Wilson, A.E., Wu, S., and Tian, L. (2019). PgUGT95B2 preferentially metabolizes flavones/flavonols and has evolved independently from flavone/flavonol UGTs identified in *Arabidopsis thaliana*. Phytochemistry 157:184-193.

Witte, S., Moco, S., Vervoort, J., Matern, U., and Martens, S. (2009). Recombinant expression and functional characterisation of regiospecific flavonoid glucosyltransferases from *Hieracium pilosella* L. Planta 229:1135-1146.

Wolfgang Schweiger, J.B., Sanghyun Shin, Brigitte Poppenberger, Franz Berthiller, Marc Lemmens, Gary J Muehlbauer, Gerhard Adam. (2012). Validation of a Candidate Deoxynivalenol-Inactivating UDP-glucosyltransferase From Barley by Heterologous Expression in Yeast. Mol Plant Microbe In 23:977-986.

Xie, F., Xiao, P., Chen, D., Xu, L., and Zhang, B. (2012). miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. Plant Mol Biol 80:75-84.

Yan, X., Fan, Y., Wei, W., Wang, P., Liu, Q., Wei, Y., Zhang, L., Zhao, G., Yue, J., and Zhou, Z. (2014). Production of bioactive ginsenoside compound K in metabolically engineered yeast. Cell Res 24:770-773.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., Watanabe-Takahashi, A., Inoue, E., and Saito, K. (2008). Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis. Plant Cell 20:2160-2176.

Yonekura-Sakakibara, K., Tohge, T., Niida, R., and Saito, K. (2007). Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. J Biol Chem 282:14932-14941.

Yu, H.S., Ma, L.Q., Zhang, J.X., Shi, G.L., Hu, Y.H., and Wang, Y.N. (2011). Characterization of glycosyltransferases responsible for salidroside biosynthesis in *Rhodiola sachalinensis*. Phytochemistry 72:862-870.

Zhang, Y., Xie, K., Liu, A., Chen, R., Chen, D., Yang, L., and Dai, J. (2019). Enzymatic biosynthesis of benzylisoquinoline alkaloid glycosides via promiscuous

glycosyltransferases from *Carthamus tinctorius*. Chinese Chemical Letters 30:443-446.

Zwaenepoel, A., and Van de Peer, Y. (2019). wgd-simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35:2153-2155.

Supplementary Figure Legends

Supplementary Figure 1. *K-mer* analysis for estimating the genome size of *P. notoginseng*. (A) The distribution of numbers of *K-mer* individuals. The red dashed line marks the main peak with depth = 63. **(B)** The distribution of numbers of *K-mer* species.

Supplementary Figure 2. Genome-wide Hi-C map of *P. notoginseng*. Interaction frequency distribution of Hi-C links among chromosomes shows in color key of heatmap ranging from light yellow to dark brown indicated the frequency of Hi-C interaction links from low to high (0−5).

Supplementary Figure 3. Analysis of repetitive sequences in the *P. notoginseng* genome. (A) LTR retrotransposons prediction using different methods; **(B)** Summary of repetitive sequences in the *P. notoginseng* genome; **(C)** Analysis of the divergence of various types of repetitive sequences.

Supplementary Figure 4. Characteristic distribution of *P. notoginseng* annotated genes and comparison with related species. Plant genomes for comparison include *A. thaliana*, *D. carota*, *P. ginseng* and *P. notoginseng*-published.

Supplementary Figure 5. Cluster analysis of gene families of eight species.

Supplementary Figure 6. Venn diagram of distribution of gene families among the seven species. The histogram on the left represents the number of gene families of each species. The red dots and lines on the right represent the gene family's classification between the designated species, and the number of each group is shown by the bar graph above.

Supplementary Figure 7. Enrichment analysis of GO and KEGG in *P. notoginseng*-specific gene families. GO enrichment analysis includes three parts: biological process **(A)**, cellular component **(B)** and molecular function **(C)**. **(D)** KEGG enrichment analysis of *P. notoginseng*-specific genes. The larger the bubble radius, the higher the rich-ratio value and the redder the color of bubble, the higher the degree of enrichment.

Supplementary Figure 8. Evolution analysis of *P. notoginseng* genome. (A) Estimated time of divergence between the eight species in the evolutionary process. **(B)**

Analysis of gene family expansion and contraction between the eight plant genomes.

Supplementary Figure 9. GO enrichment analysis of expanded and contracted gene families of *P. notoginseng* genome. (A) GO enrichment analysis of expanded gene families; **(B)** GO enrichment analysis of contracted gene families; **(C)** Summary of gene numbers expanding and contracting in different categories of GO enrichment analysis.

Supplementary Figure 10. Summary of the syntenic analysis between *P. notoginseng* and *V. vinifera* (n=1 biologically independent samples).

Supplementary Figure 11. Collinear analysis among *D. carota*, *P. notoginseng* and *V. vinifera* genome. The red lines in the genomes of *P. notoginseng* and *V. vinifera* indicate that the 1:2 correspondence between the two collinear regions.

Supplementary Figure 12. Synonymous substitution rate (*Ks*) distributions of syntenic blocks in *P. notoginseng* and comparison with *P. ginseng* and *V. vinifera* genome.

Supplementary Figure 13. Phylogenetic tree of key enzyme genes in terpenoid biosynthetic pathway in 8 species including *P. notoginseng*, *P. ginseng*, *D. carota*, *V. vinifera*, *O. sativa*, *A. thaliana*, *G. uralensis* and *C. annuum* (1). Each phylogenetic tree of terpenoid biosynthetic genes was constructed by using MEGA X with the neighbor-joining method.

Supplementary Figure 14. Phylogenetic trees of key enzyme genes involved in terpenoid biosynthetic pathway in 8 species including *P. notoginseng*, *P. ginseng*, *D. carota*, *V. vinifera*, *O. sativa*, *A. thaliana*, *G. uralensis* and *C. annuum* (2). Each phylogenetic tree of terpenoid biosynthetic genes was constructed by using MEGA X with the neighbor-joining method.

Supplementary Figure 15. Evolution of ginsenoside-associated genes in *P. notoginseng*. (A) Genome duplication in *P. notoginseng*. The calculated Ks value was converted to the divergence time according to T=Ks/2r, where r represents a substitution rate of 6.5×10-9 mutations per site per year for eudicots (n=1 biologically independent samples). **(B)** Duplication event(s) for each gene pair is(are) shown along the timeline from 0 to 150 million years ago with different colors.

Supplementary Figure 16. Overview of clustering of transcriptome samples.

Supplementary Figure 17. Pearson correlation analysis of transcriptome samples. The R^2 value between two random transcripts were indicated in the box, and ranging from white to blue indicted from low to high (0-1).

Supplementary Figure 18. The proportion distribution of various reads before filtering in all samples.

Supplementary Figure 19. The coverage distribution of gene regions mapping on genome in each transcript.

Supplementary Figure 20. Statistics of alternative splicing events. (A) events in each sample; **(B)** different types of alternative splicing events in the comparison groups. TSS: Transcription Start Site; TTS: Transcription Terminal Site; SKIP: Skipped exon; XSKIP: Approximate SKIP; MSKIP: Multi-exon SKIP; XMSKIP: Approximate MSKIP; IR: Intron retention; XIR: Approximate IR; MIR: Multi-IR; XMIR: Approximate MIR; AE: Alternative exon ends (5', 3' or both); XAE: Approximate AE (5' or 3'); A3SS: Alternative 3' splice site; A5SS: Alternative 5' splice site.

Supplementary Figure 21. Variation analysis of each sample. (A) distribution of each variant type; **(B)** according to the detected SNP loci, the frequency of each mutation type has been counted, taking the data results of One1Leaf, Two1Leaf, Two2Flower, Tri3Stem, Tri1Xylem, Fou2Perid as examples; **(C)** according to the detected InDel loci, the frequency of each InDel length has been counted, taking the data results of One1Leaf, Two1Leaf, Two2Flower, Tri3Stem, Tri1Xylem, Fou2Perid as examples.

Supplementary Figure 22. Density distribution diagram of gene expression in each transcriptome sample.

Supplementary Figure 23. Box plot of the overall distribution of gene expression in each transcriptome sample.

Supplementary Figure 24. The exploration of the molecular mechanism of the formation of *P. notoginseng*'s tubercles. (**A**) the display of root morphology of *P. notoginseng*, and the red arrow points to the tubercles. (**B**) GO enrichment analysis of DEGs between the periderm and tubercle group. (**C**) the Directed Acyclic Graph (DAG)

of GO enrichment analysis, the darker color indicates the more significant enrichment and the red is the most significant. The larger the bubble radius, the higher the rich-ratio value and the redder the color of bubble, the higher the degree of enrichment.

Supplementary Figure 25. Spatial expression profile of key enzyme genes in saponin biosynthesis pathway. The genes in red font are the functional UGT cloned in this study.

Supplementary Figure 26. Phylogenetic analysis of CYP450 genes in *P. notoginseng* using MEGA-X.

Supplementary Figure 27. Heat map of the expression of the cloned UGT genes in different transcript samples. The genes marked by five-pointed stars are those with catalytic function identified in this study. In the heat map, the relative expression level from high to low (-2 to 2) is represented by the range from blue to red.

Supplementary Figure 28 The blank control experiments of protein catalytic reaction in this study. (A) Analysis results of *E. coli* no-load control with PPD as the catalytic substrate. **(B)** Analysis results of *E. coli* no-load control with PPT as the catalytic substrate. **(C)** Analysis results of *E. coli* no-load control with ginsenoside F1 as the catalytic substrate. **(D)** Analysis results of *E. coli* no-load control with Rh2 as the catalytic substrate. The molecular ion peaks with 667.4465, 683.4406, 845.496 and 784.00 were extracted respectively, and the mass spectrum in the green box did not match with any corresponding glycoside product, indicating no product was formed.

Supplementary Figure 29. UPLC/Q-TOF analysis results of PnUGT3 protein catalytic reaction. (A) Chromatograms and mass spectrum of ginsenoside Rh1 standard and PnUGT3 catalytic products using PPT as substrate. **(B)** Chromatograms and mass spectrum of ginsenoside Rg1 standard and PnUGT3 catalytic products using F1 as substrate.

Supplementary Figure 30. UPLC/Q-TOF analysis results of PnUGT1 and PnUGT5 protein catalytic reaction. **(A)** Chromatograms and mass spectrum of ginsenoside F1 standard and PnUGT1 catalytic products using PPT as substrate. **(B)** Chromatograms and mass spectrum of ginsenoside CK standard and PnUGT1 catalytic products using PPD as substrate. **(C)** Chromatograms and mass spectrum of

ginsenoside F2 standard and PnUGT1 catalytic products using Rh2 as substrate. **(D)** Chromatograms and mass spectrum of ginsenoside Rh2 standard and PnUGT5 catalytic products using PPD as substrate.

Supplementary Figure 31. UPLC/Q-TOF analysis results of PnUGT2 and PnUGT4 protein catalytic reaction. Chromatograms and mass spectrum of ginsenoside Rg3 standard and PnUGT2 and PnUGT4 catalytic products using Rh2 as substrate.

Supplementary Figure 32. The structural formulas of various saponins in *P. notoginseng*.

Supplementary Figure 33. WGCAN analysis and characterization of corresponding data. (A) Construction of the sample clustering evolutionary tree of transcriptome to screen out outliers. **(B)** Construction the PCA map of transcriptome samples. **(C)** Analysis of network topology for various soft-thresholding powers. When we set R^2=0.9, the optimal candidate threshold to reach this height is 10. **(D)** Visualization of the eigengene network representing the relationships among the modules. The redder the color in the heat map, the stronger the correlation between the two modules.

Supplementary Figure 34. Expression profile of key enzyme genes in saponin biosynthesis pathway. The right side of the heatmap shows the evolutionary tree of genes, and genes with similar expression patterns are clustered into one group.

Supplementary Figure 35. Heat map of expression of UGT genes and genes in terpenoid biosynthesis pathway. The right side of the heatmap shows the evolutionary tree of genes, and genes with similar expression patterns are clustered into one group.

Supplementary Figure 36. Gene clusters involved in saponins biosynthesis found in *P. notoginseng* genome. (A) Gene clusters on chromosomes 1, 2 and their correspondence. **(B)** Gene clusters on chromosomes 6, 8 and their correspondence. Orange lines indicate copies of genes with the same function, and blue lines indicate the correlation between transcription factors and pathway genes.

Supplementary Table legends

Supplementary Table 1. Estimation of genome size of *P. notoginseng* based on *K-mer* analysis.

Supplementary Table 2. Sequencing data statistics of *P. notoginseng*.

Supplementary Table 3. The Statistics of Pseudomolecule based on Hi-C technique.

Supplementary Table 4. Statistic of DNA base composition in the *P. notoginseng* genome.

Supplementary Table 5. Statistics of consistency assessment of the *P. notoginseng* genome.

Supplementary Table 6. Assessment the gene coverage rate using BUSCO.

Supplementary Table 7. Annotation of repetitive sequences in the *P. notoginseng* genome.

Supplementary Table 8. Summary of repetitive sequences in the *P. notoginseng* genome.

Supplementary Table 9. Basic statistical results of gene structure prediction of *P. notoginseng* genome.

Supplementary Table 10. Basic statistical results of gene structure prediction of *P. notoginseng* and relative species.

Supplementary Table 11. Statistical results of gene function annotation of *P. notoginseng* genome.

Supplementary Table 12. Statistical results of non-coding RNA of *P. notoginseng* genome.

Supplementary Table 13. The Statistics of gene clustering to gene families in various species.

Supplementary Table 14. Enriched GO terms of genes in *P. notoginseng*-specific families.

Supplementary Table 15. Enriched GO terms of genes in expanded gene families.

Supplementary Table 16. Enriched GO terms of genes in contracted gene families.

Supplementary Table 17. Copy number variation of genes involved in the ginsenoside biosynthesis in the *P. notoginseng* and seven other plant species.

Supplementary Table 18. *Ks* values and duplication times of genes involved in ginsenoside biosynthesis in *P. notoginseng*.

Supplementary Table 19. Statistics of the information and grouping of transcriptome samples.

Supplementary Table 20. Statistics of alternative splicing events occurred in *P. notoginseng* genome.

Supplementary Table 21. Statistics of variation events occurred in *P. notoginseng* genome.

Supplementary Table 22. Representative genes which are highly expressed in tubercle group.

Supplementary Table 23. Statistics of transcription factors in *P. notoginseng* genome.

Supplementary Table 24. Statistics of FPKM expression in different tissues of some key enzyme genes in the terpene biosynthesis pathway.

Supplementary Table 25. The CYP450 genes used to construct phylogenetic tree in this research.

Supplementary Table 26. The UGT genes used to construct phylogenetic tree in this research.

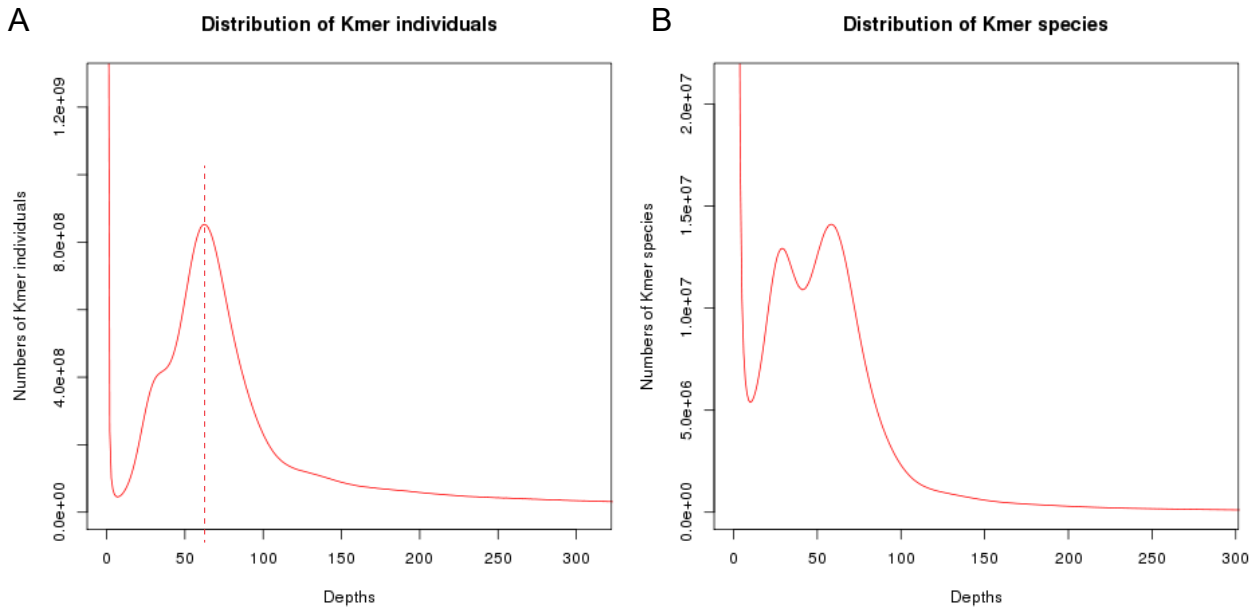
Supplementary Table 27. Primers for cloning UGT genes in *P. notoginseng* genome.

Supplementary Table 28. Annotation and GO enrichment of candidate UGT genes selected by WGCNA analysis.

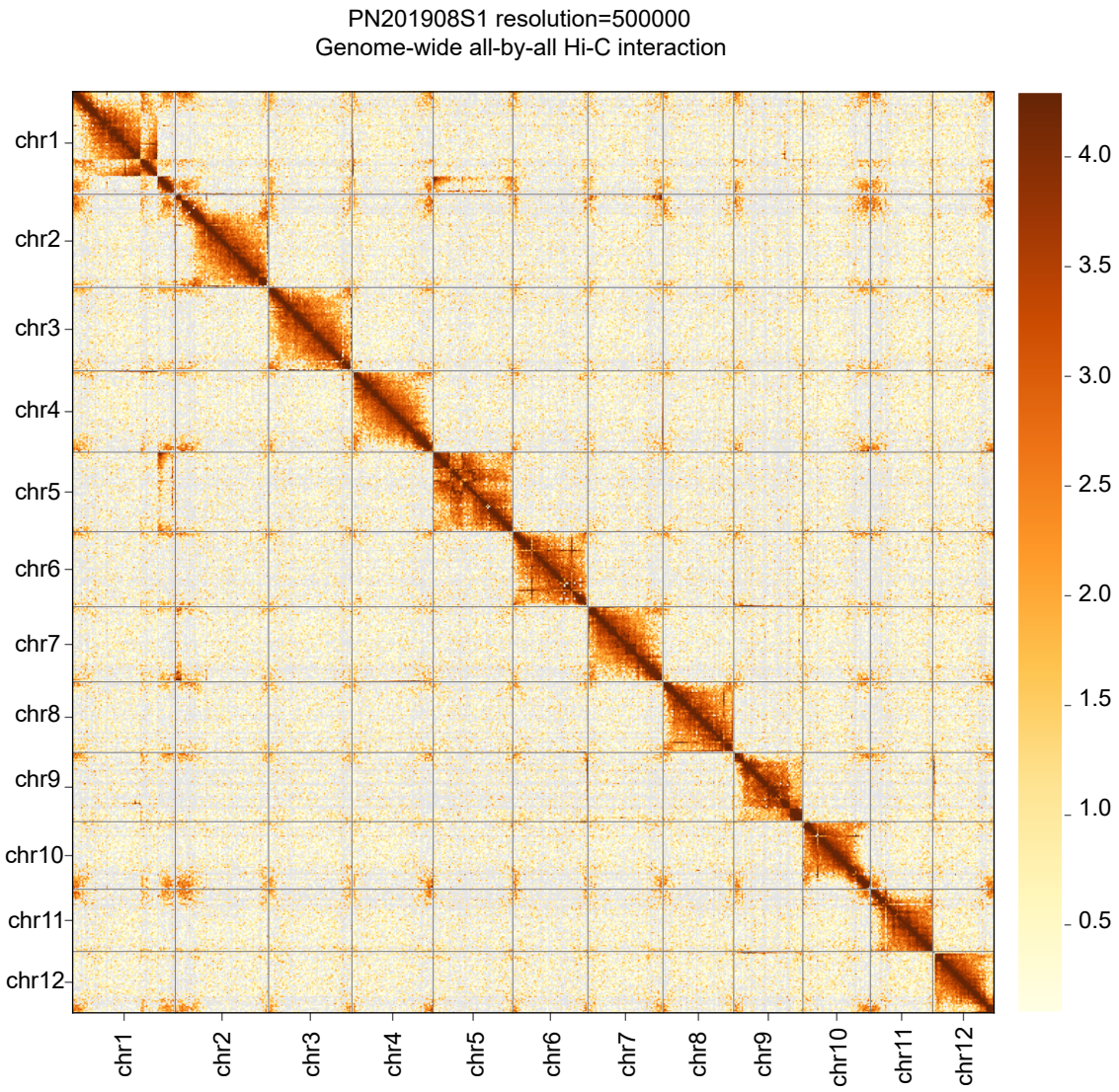
Supplementary Table 29. Annotation and GO enrichment of candidate UGT genes screened from the gene expression patterns.

(**Note:** Among these tables, Supplementary Table 19, 20, 21, 24, 28, 29 are placed in a separate excel sheet due to the large content.)

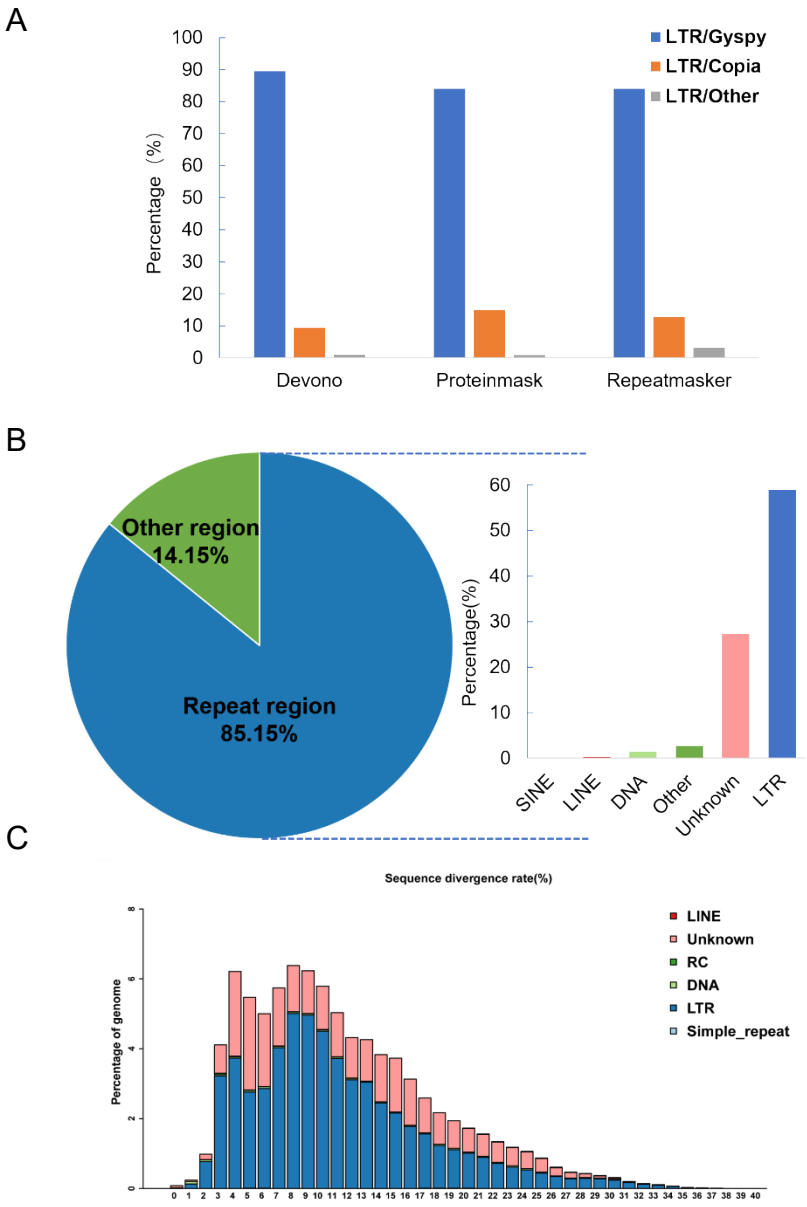
Supplementary Figures



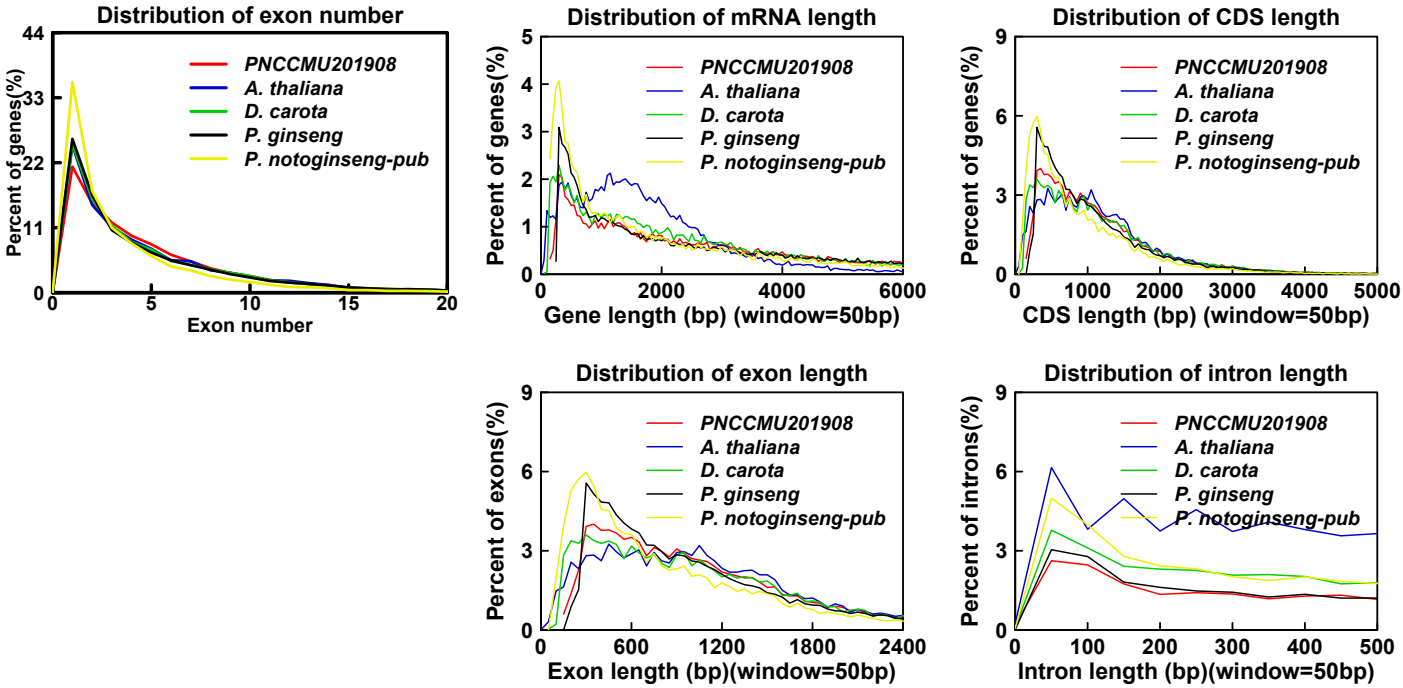
Supplementary Figure 1. *K-mer* analysis for estimating the genome size of *P. notoginseng*. **(A)** The distribution of numbers of *K-mer* individuals. The red dashed line marks the main peak with depth = 63. **(B)** The distribution of numbers of *K-mer* species.



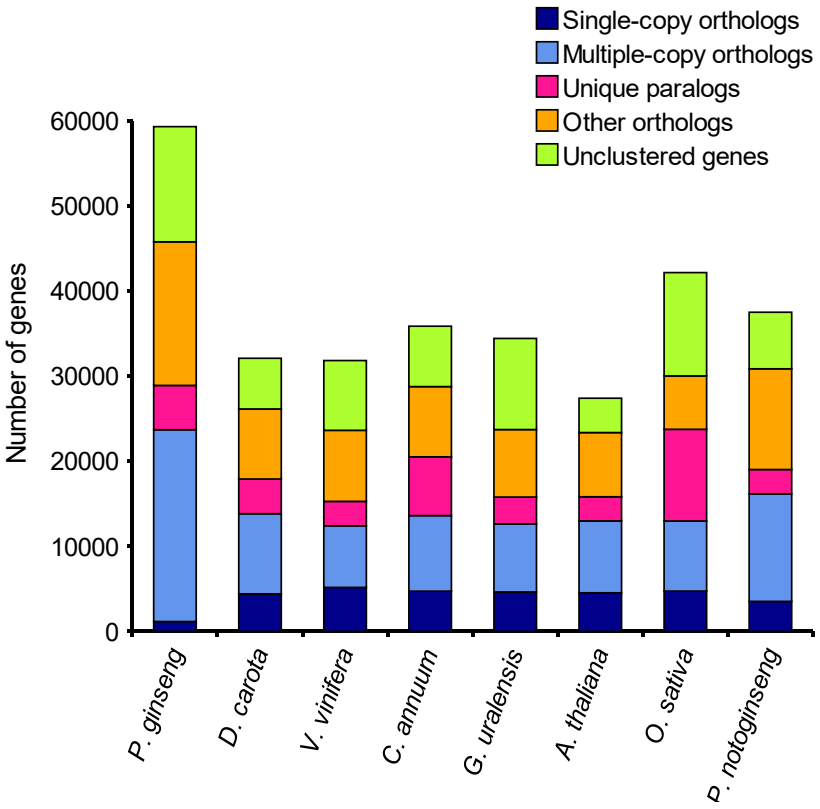
Supplemental Figure 2. Genome-wide Hi-C map of *P. notoginseng*. Interaction frequency distribution of Hi-C links among chromosomes shows in color key of heatmap ranging from light yellow to dark brown indicated the frequency of Hi-C interaction links from low to high (0-5).



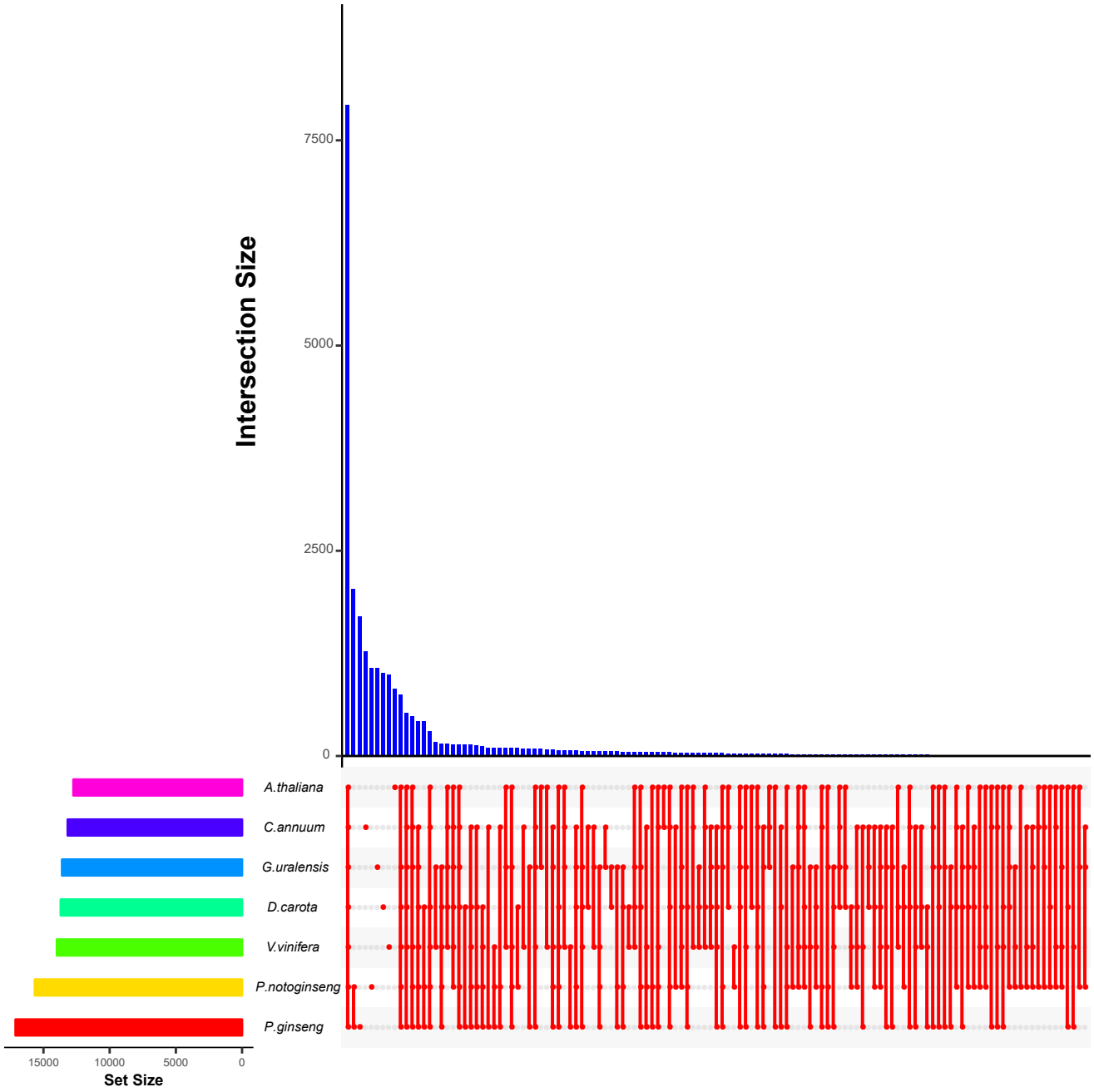
Supplementary Figure 3. Analysis of repetitive sequences in the *P. notoginseng* genome. (**A**) LTR retrotransposons prediction using different methods. (**B**) Summary of repetitive sequences in the *P. notoginseng* genome. (**C**) Analysis of the divergence of various types of repetitive sequences.



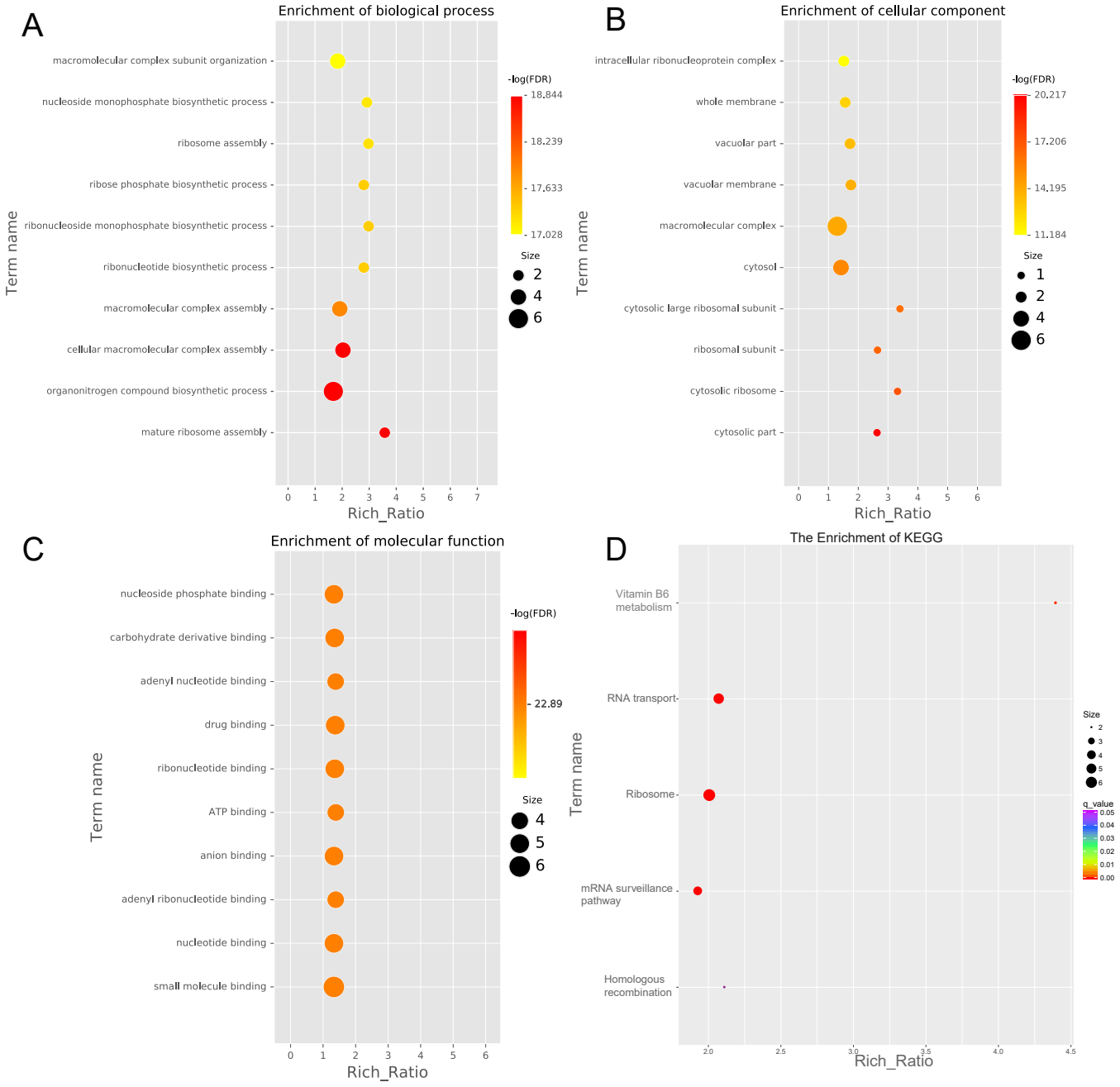
Supplementary Figure 4. Characteristic distribution of *P. notoginseng* annotated genes and comparison with related species. Plant genomes for comparison include *A. thaliana*, *D. carota*, *P. ginseng* and *P. notoginseng*-published.



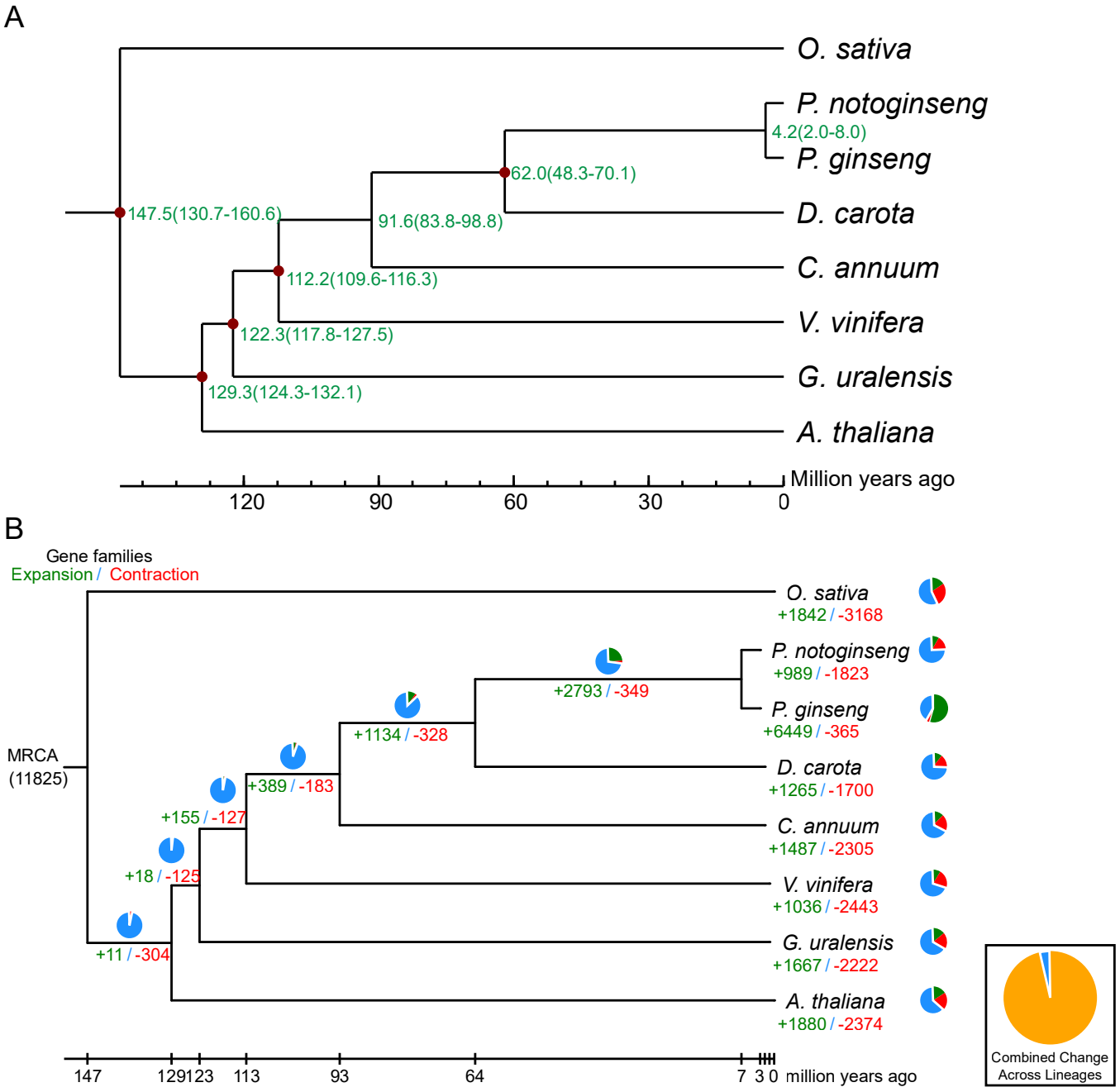
Supplementary Figure 5. Cluster analysis of gene families of eight species.



Supplementary Figure 6. Venn diagram of distribution of gene families among the seven species.
The histogram on the left represents the number of gene families of each species. The red dots and lines on the right represent the gene family's classification between the designated species, and the number of each group is shown by the bar graph above.



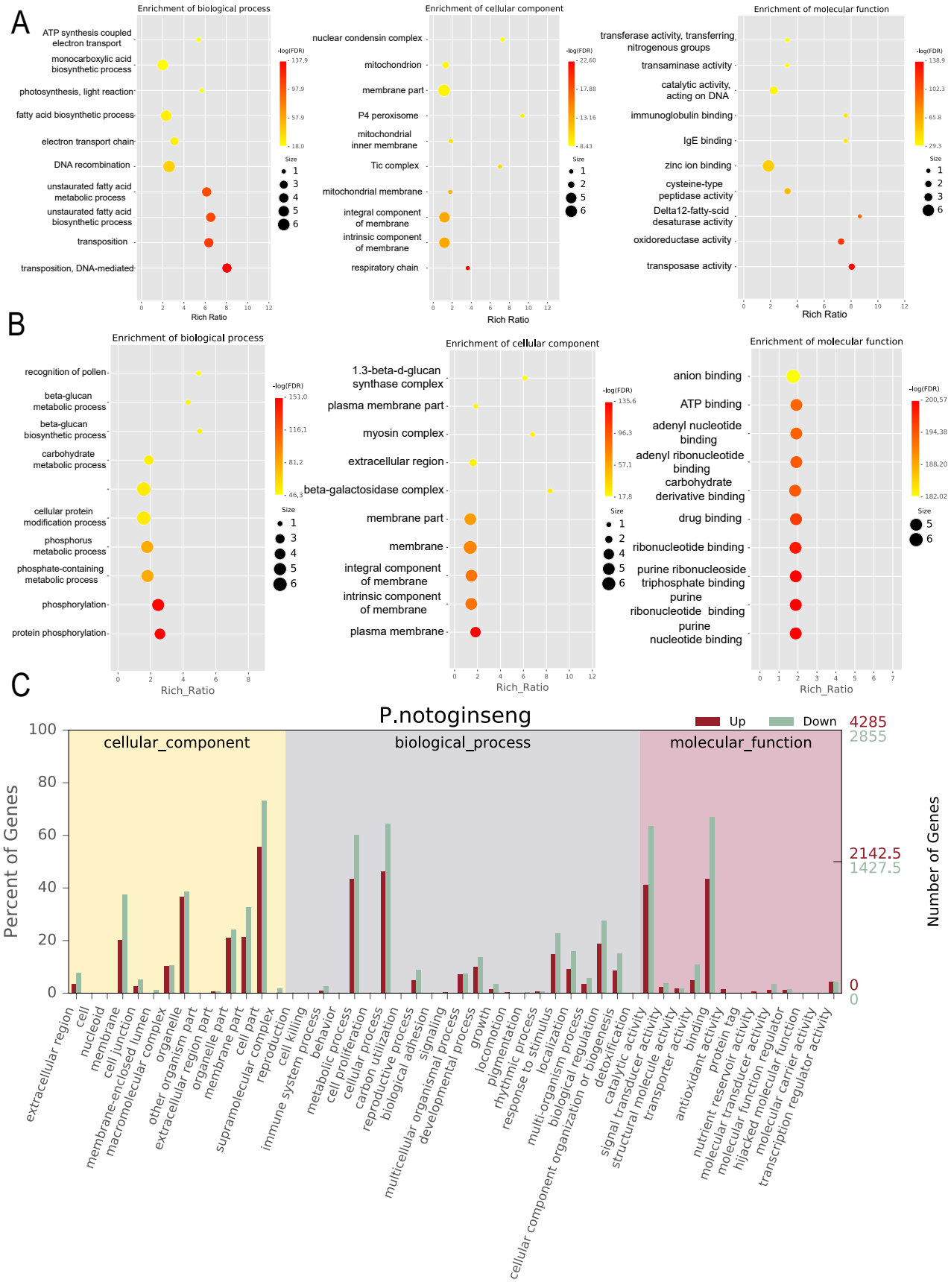
Supplementary Figure 7. Enrichment analysis of GO and KEGG in *P. notoginseng*-specific gene families. GO enrichment analysis includes three parts: biological process **(A)**, cellular component **(B)** and molecular function **(C)**. **(D)** KEGG enrichment analysis of *P. notoginseng*-specific genes. The larger the bubble radius, the higher the rich-ratio value and the redder the color of bubble, the higher the degree of enrichment.



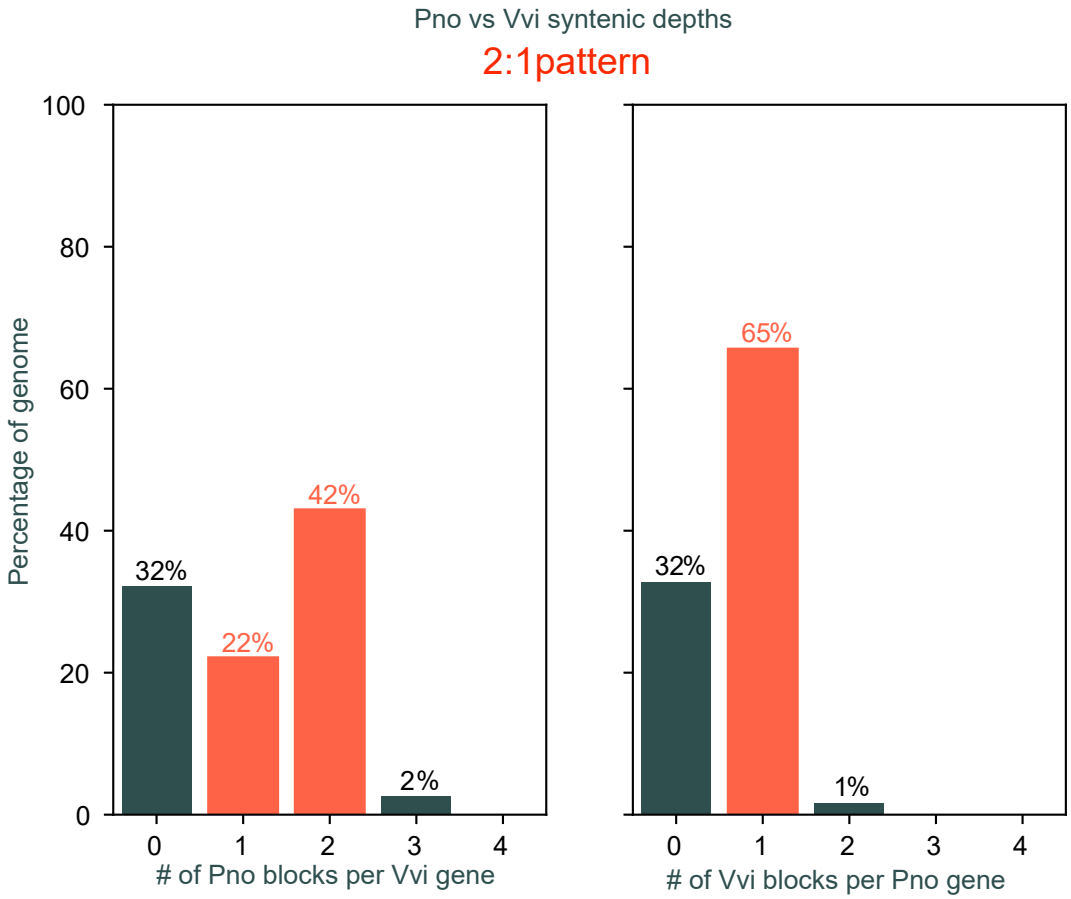
Supplementary Figure 8. Evolution analysis of *P. notoginseng* genome.

(A) Estimated time of divergence between the eight species in the evolutionary process.

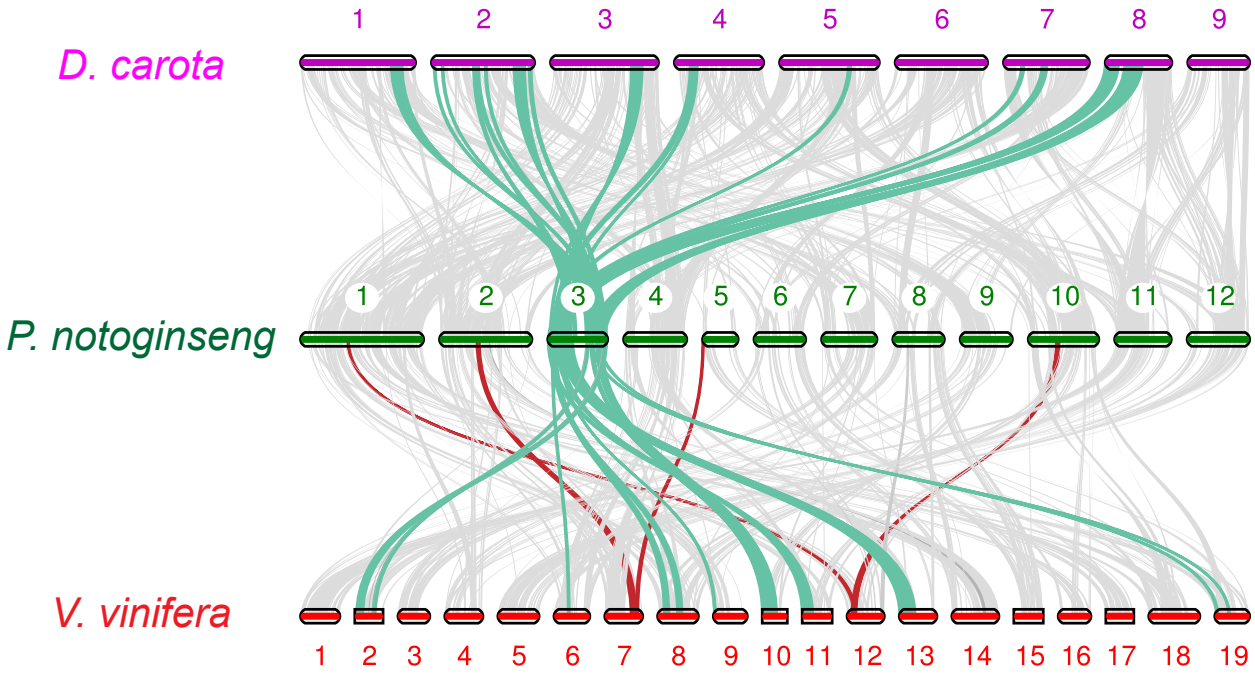
(B) Analysis of gene family expansion and contraction between the eight plant genomes.



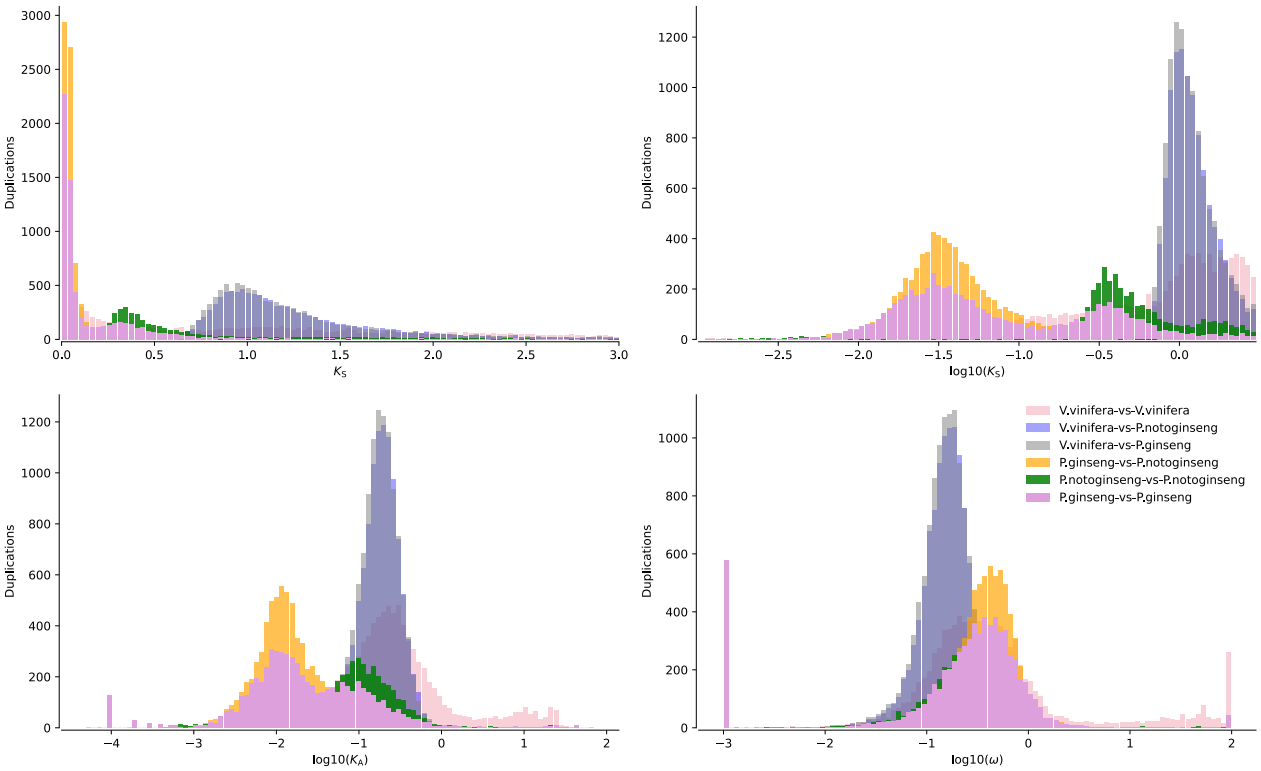
Supplemental Figure 9. GO enrichment analysis of expanded and contracted gene families of *P. notoginseng* genome. (A) GO enrichment analysis of expanded gene families; **(B)** GO enrichment analysis of contracted gene families; **(C)** Summary of gene numbers expanding and contracting in different categories of GO enrichment analysis.



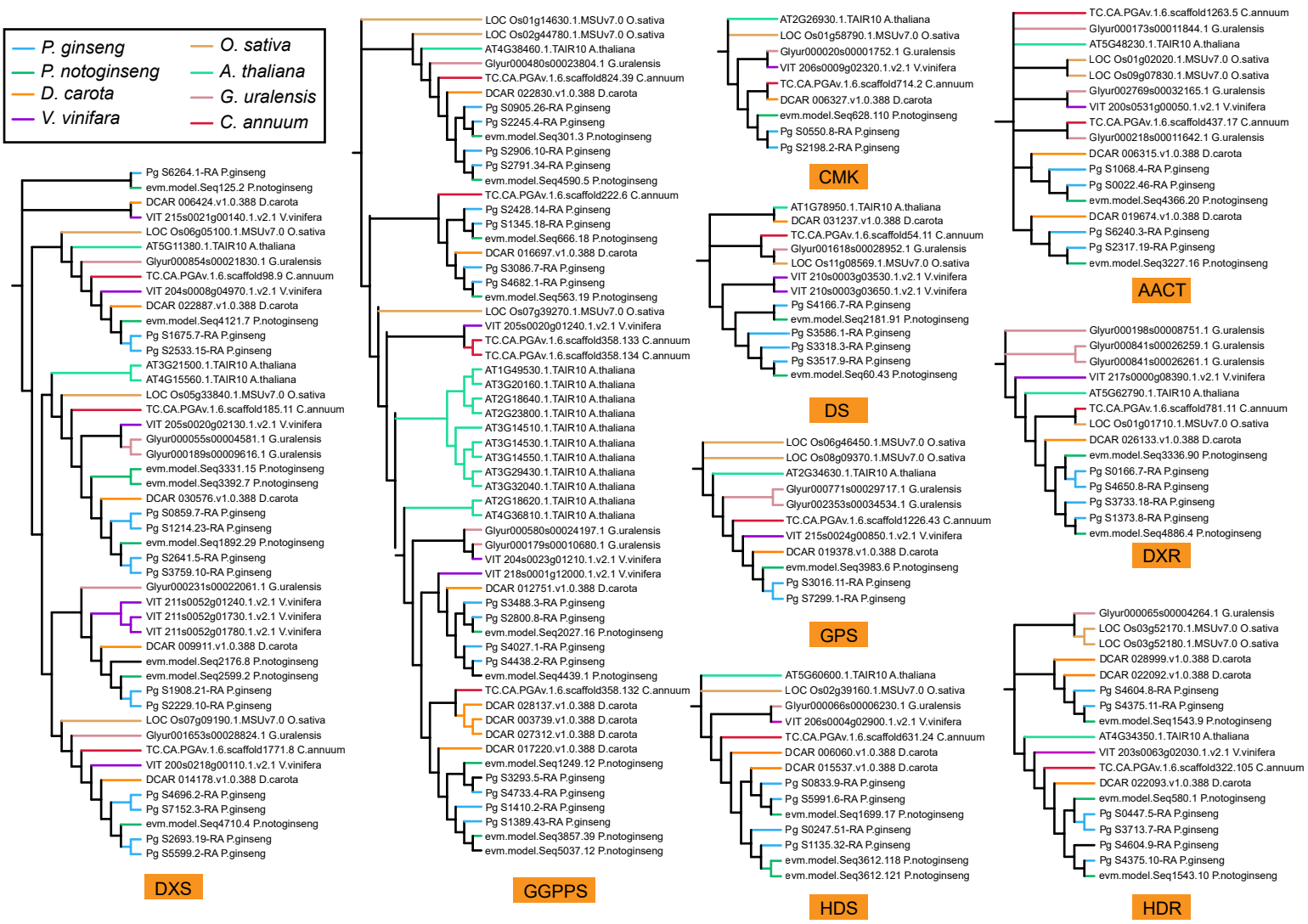
Supplementary Figure 10. Summary of the syntenic analysis between *P. notoginseng* and *V. vinifera* (n=1 biologically independent samples).



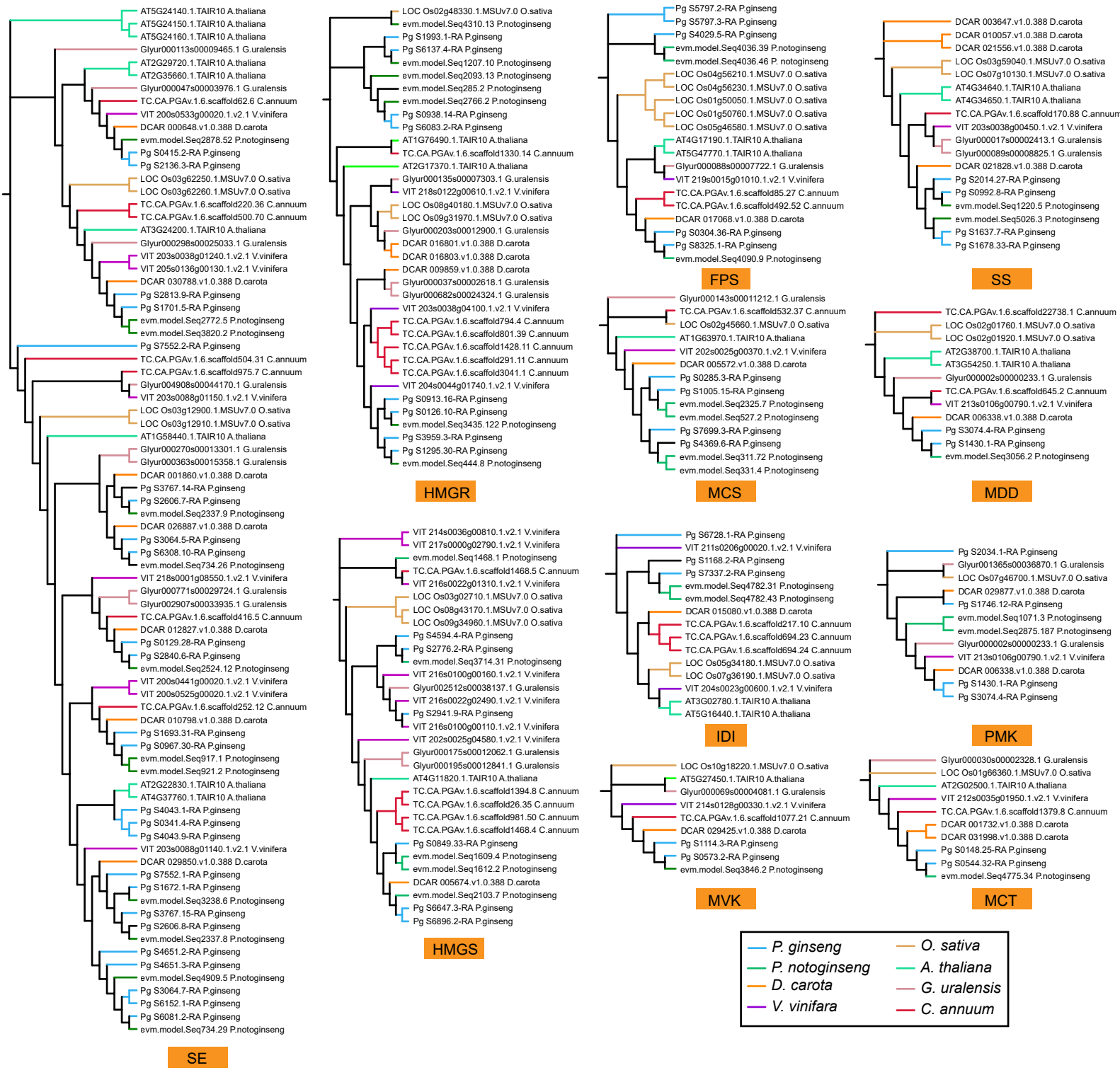
Supplementary Figure 11. Collinear analysis among *D. carota*, *P. notoginseng* and *V. vinifera* genome. The red lines in the genomes of *P. notoginseng* and *V. vinifera* indicate that the 1:2 correspondence between the two collinear regions.



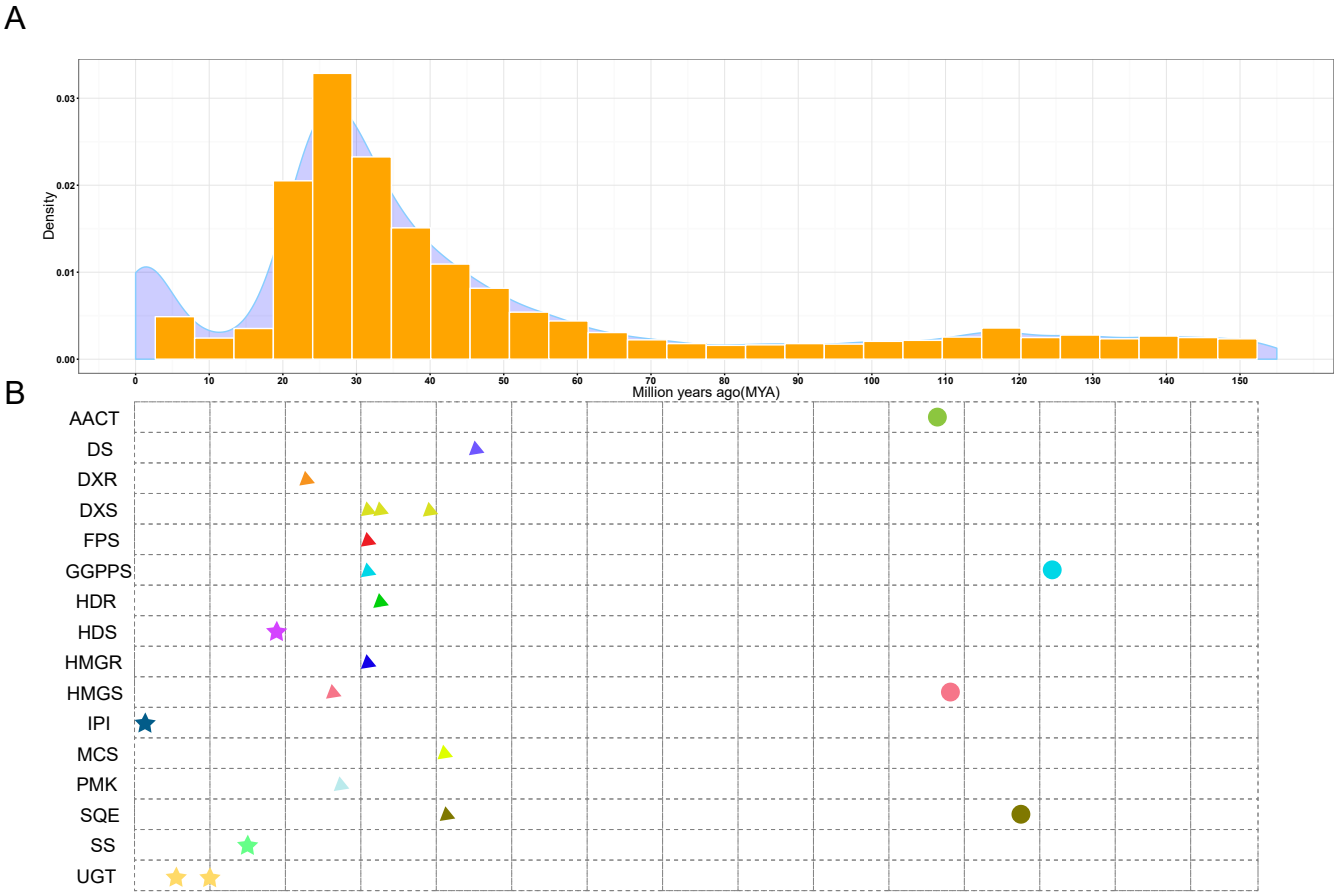
Supplementary Figure 12. Synonymous substitution rate (*Ks*) distributions of syntenic blocks in *P. notoginseng* and comparison with *P. ginseng* and *V. vinifera* genome.



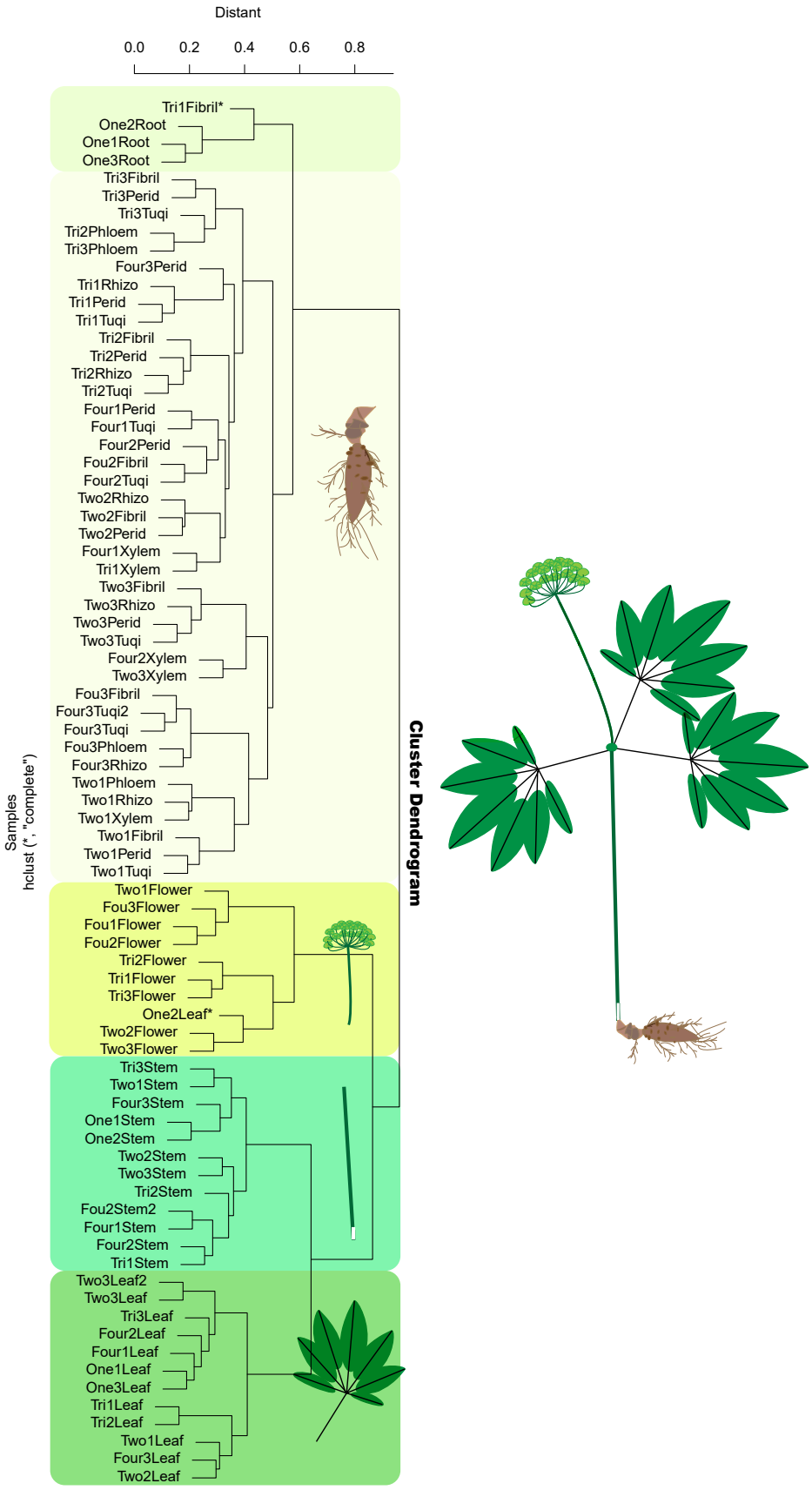
Supplemental Figure 13. Phylogenetic tree of key enzyme genes in terpenoid biosynthetic pathway in 8 species including *P. notoginseng*, *P. ginseng*, *D. carota*, *V. vinifera*, *O. sativa*, *A. thaliana*, *G. uralensis* and *C. annuum* (1). Each phylogenetic tree of terpenoid biosynthetic genes was constructed by using MEGA X with the neighbor-joining method.



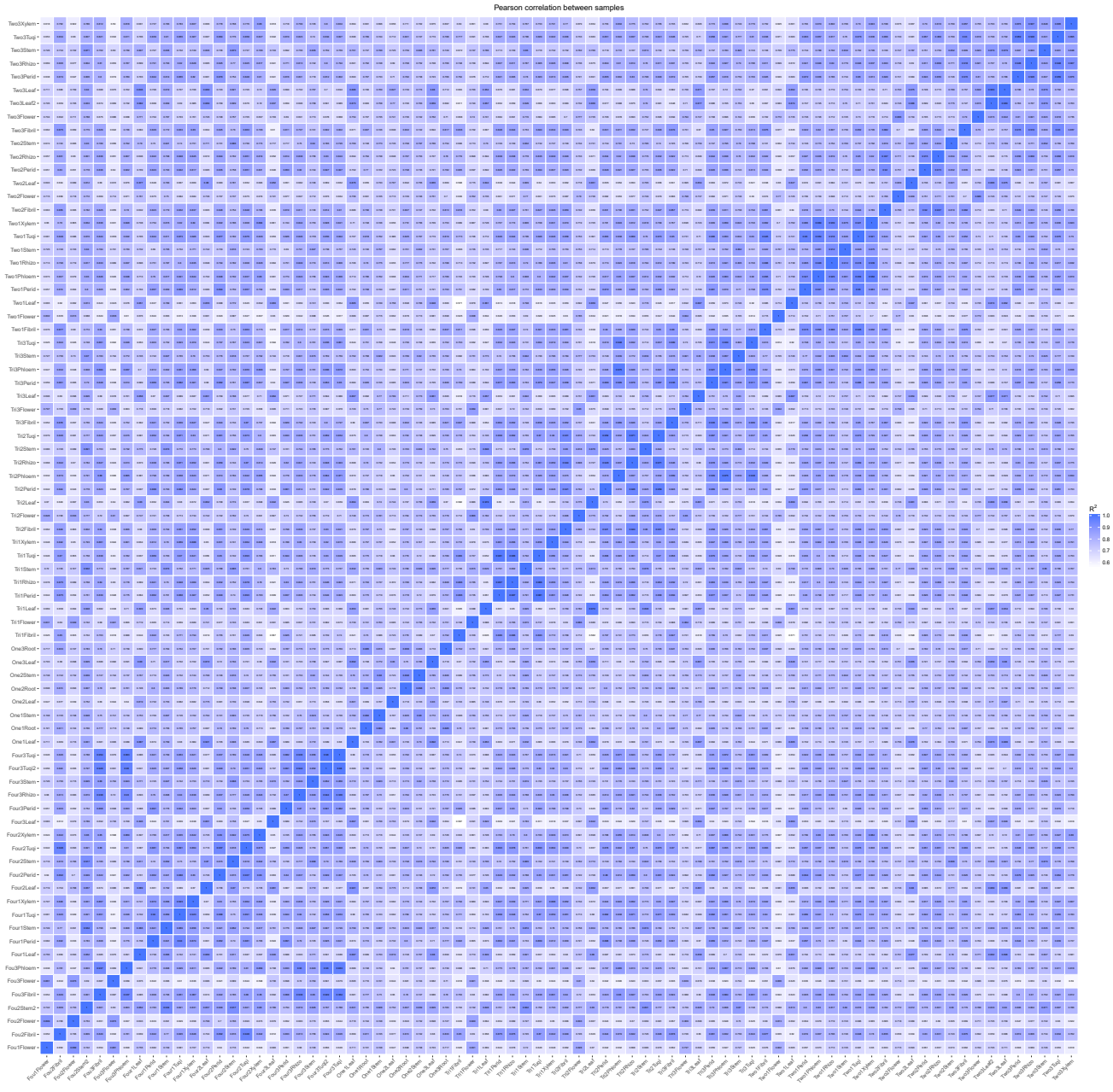
Supplemental Figure 14. Phylogenetic trees of key enzyme genes involved in terpenoid biosynthetic pathway in 8 species including *P. notoginseng*, *P. ginseng*, *D. carota*, *V. vinifera*, *O. sativa*, *A. thaliana*, *G. uralensis* and *C. annuum* (2). Each phylogenetic tree of terpenoid biosynthetic genes was constructed by using MEGA X with the neighbor-joining method.



Supplemental Figure 15. Evolution of ginsenoside-associated genes in *P. notoginseng*. **(A)** Genome duplication in *P. notoginseng*. The calculated *Ks* value was converted to the divergence time according to T=*Ks*/2r, where r represents a substitution rate of 6.5×10-9 mutations per site per year for eudicots (n=1 biologically independent samples). **(B)** Duplication event(s) for each gene pair is(are) shown along the timeline from 0 to 150 million years ago with different colors.

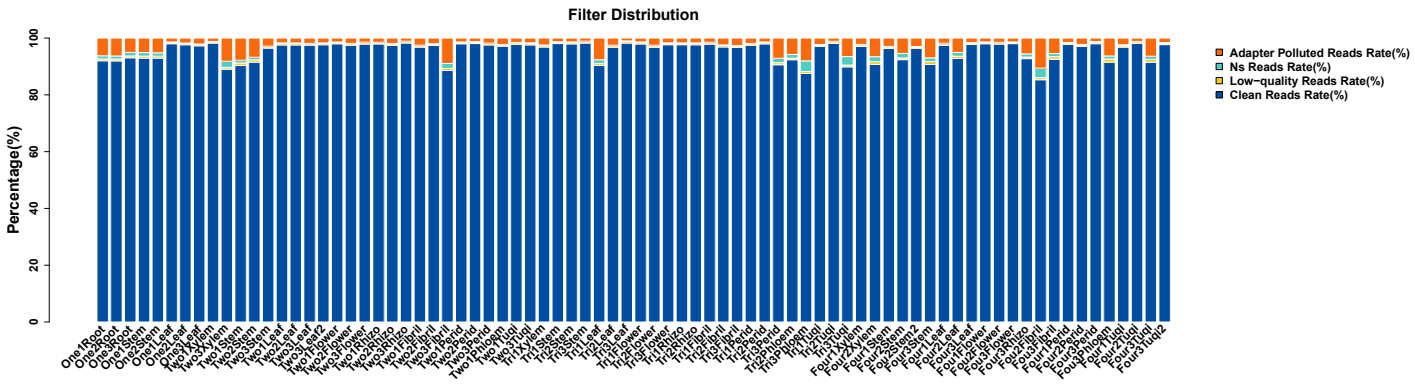


Supplemental Figure 16. Overview of clustering of transcriptome samples.

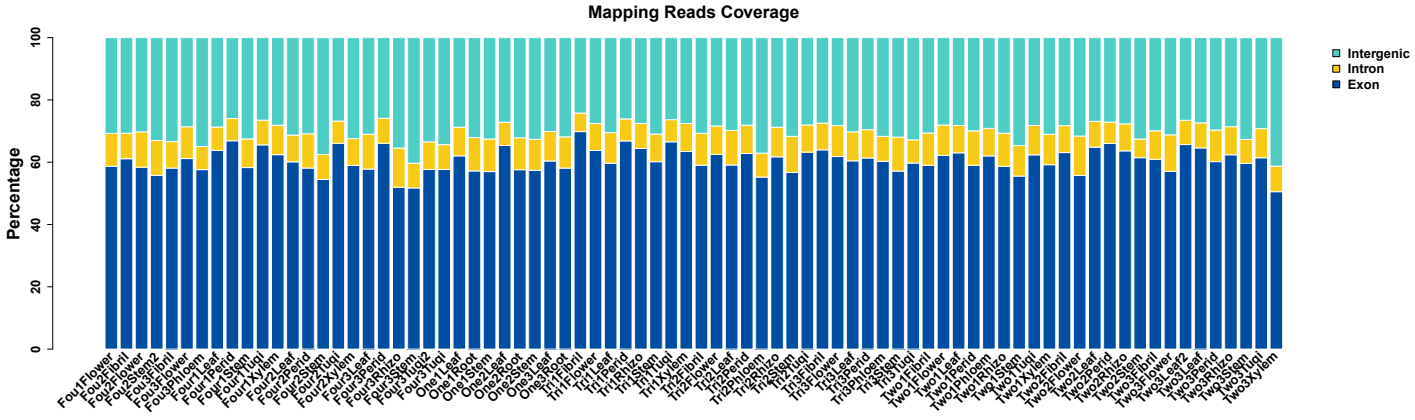


Supplemental Figure 17. Pearson correlation analysis of transcriptome samples.

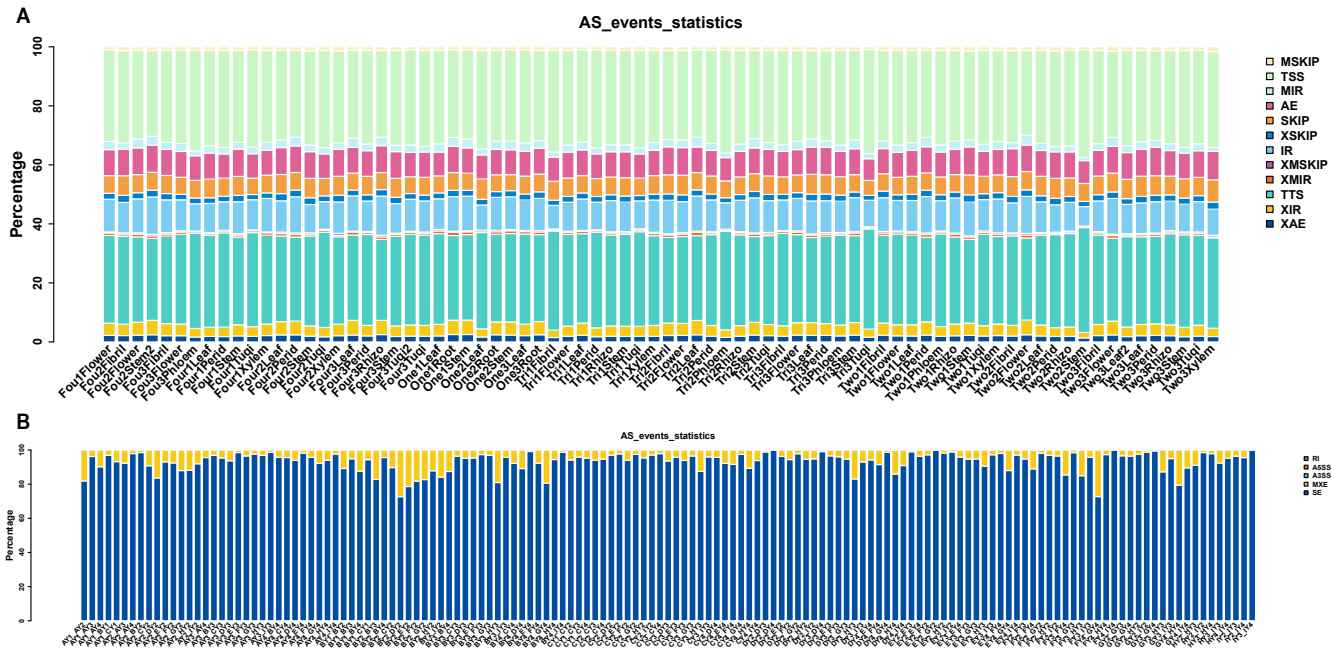
The R2 value between two random transcripts were indicated in the box, and ranging from white to blue indicted from low to high (0-1).



Supplemental Figure 18. The proportion distribution of various reads before filtering in all samples.

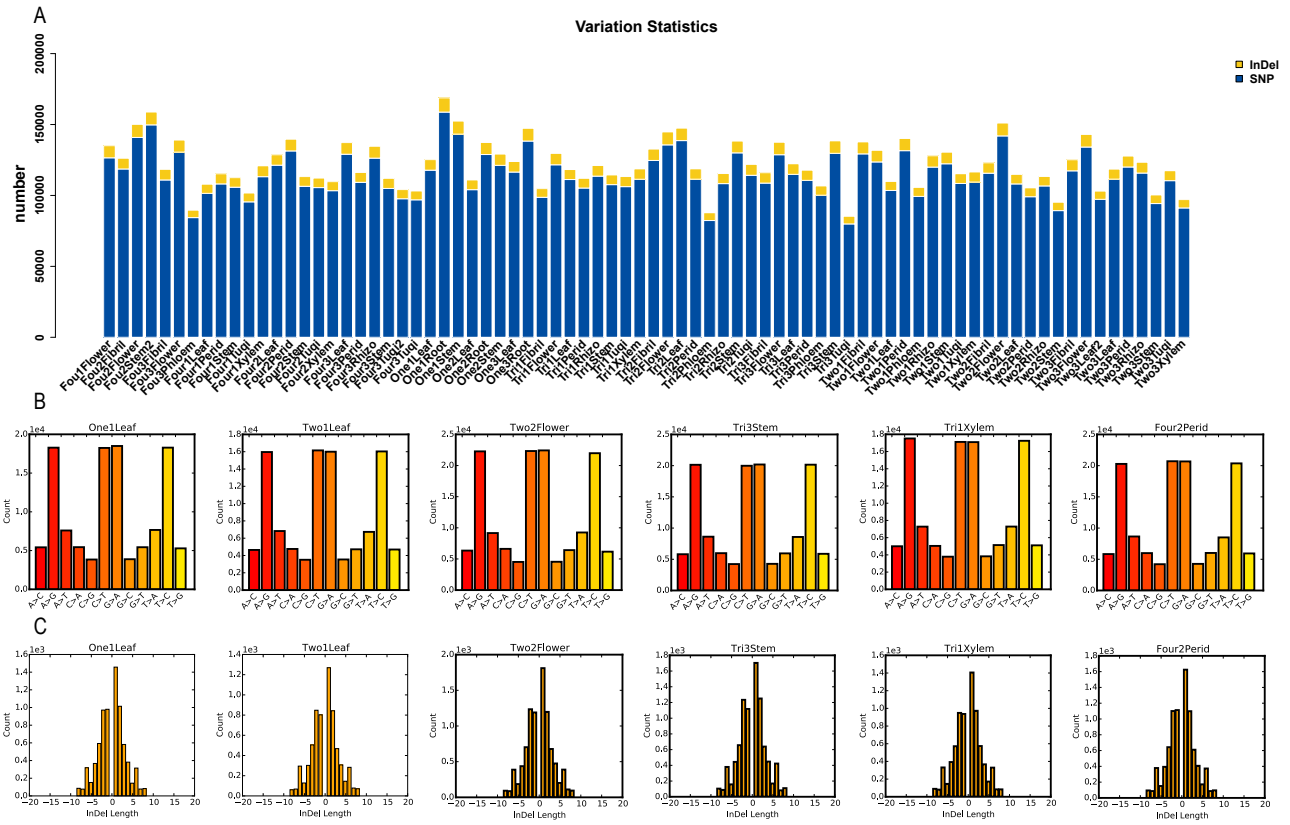


Supplemental Figure 19. The coverage distribution of gene regions mapping on genome in each transcript.

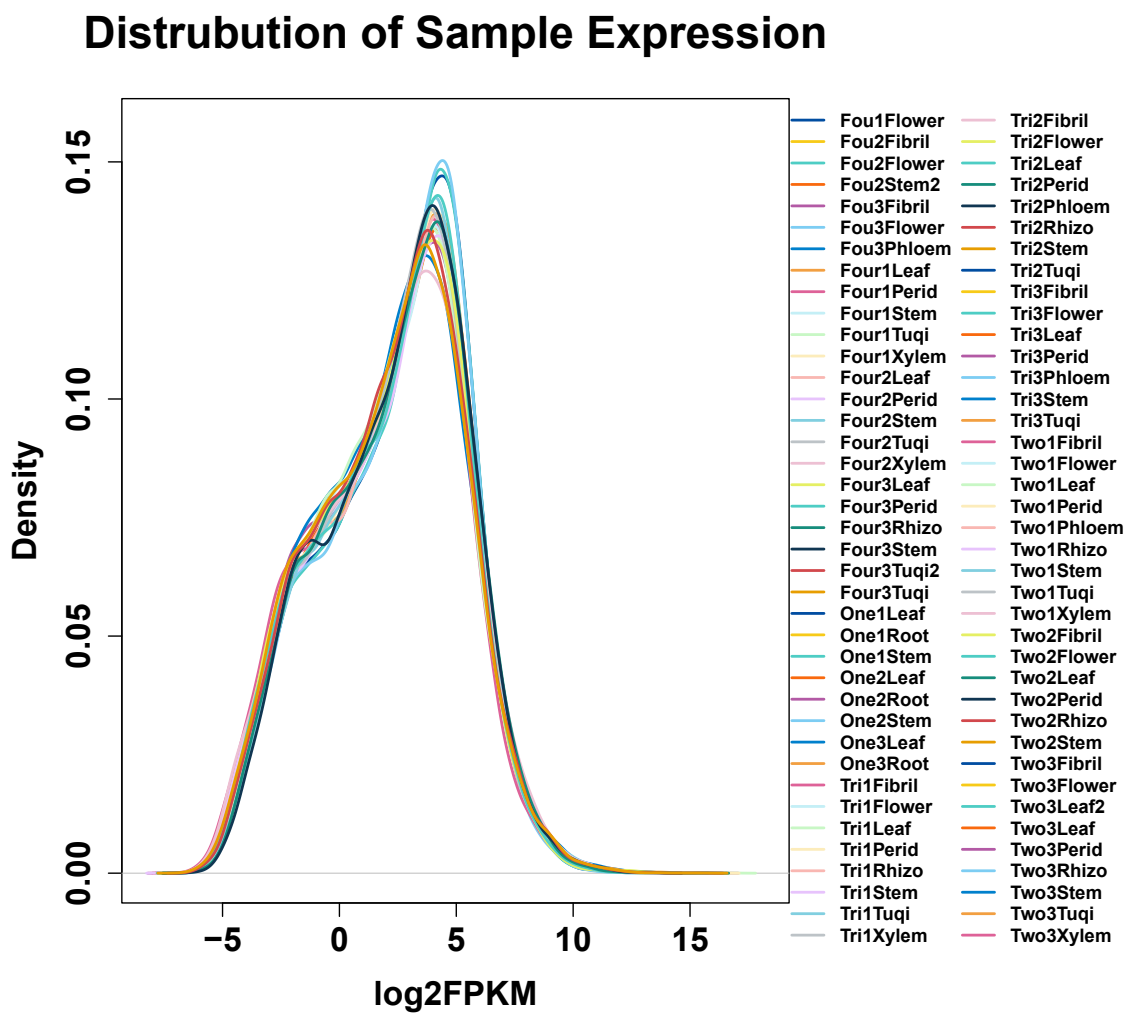


Supplemental Figure 20. Statistics of alternative splicing events. (A) events in each sample; **(B)** different types of alternative splicing events in the comparison groups.

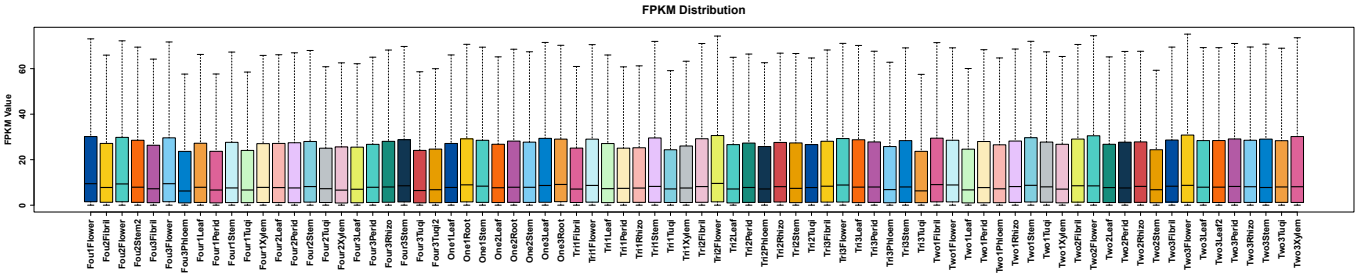
TSS: TranscriptionStart Site; TTS: Transcription Terminal Site; SKIP: Skipped exon; XSKIP: Approximate SKIP; MSKIP: Multi-exon SKIP; XMSKIP: Approximate MSKIP; IR: Intron retention; XIR: Approximate IR; MIR: Multi-IR; XMIR: Approximate MIR; AE: Alternative exon ends (5', 3' or both); XAE: Approximate AE (5' or 3'); A3SS: Alternative 3' splice site; A5SS: Alternative 5' splice site.



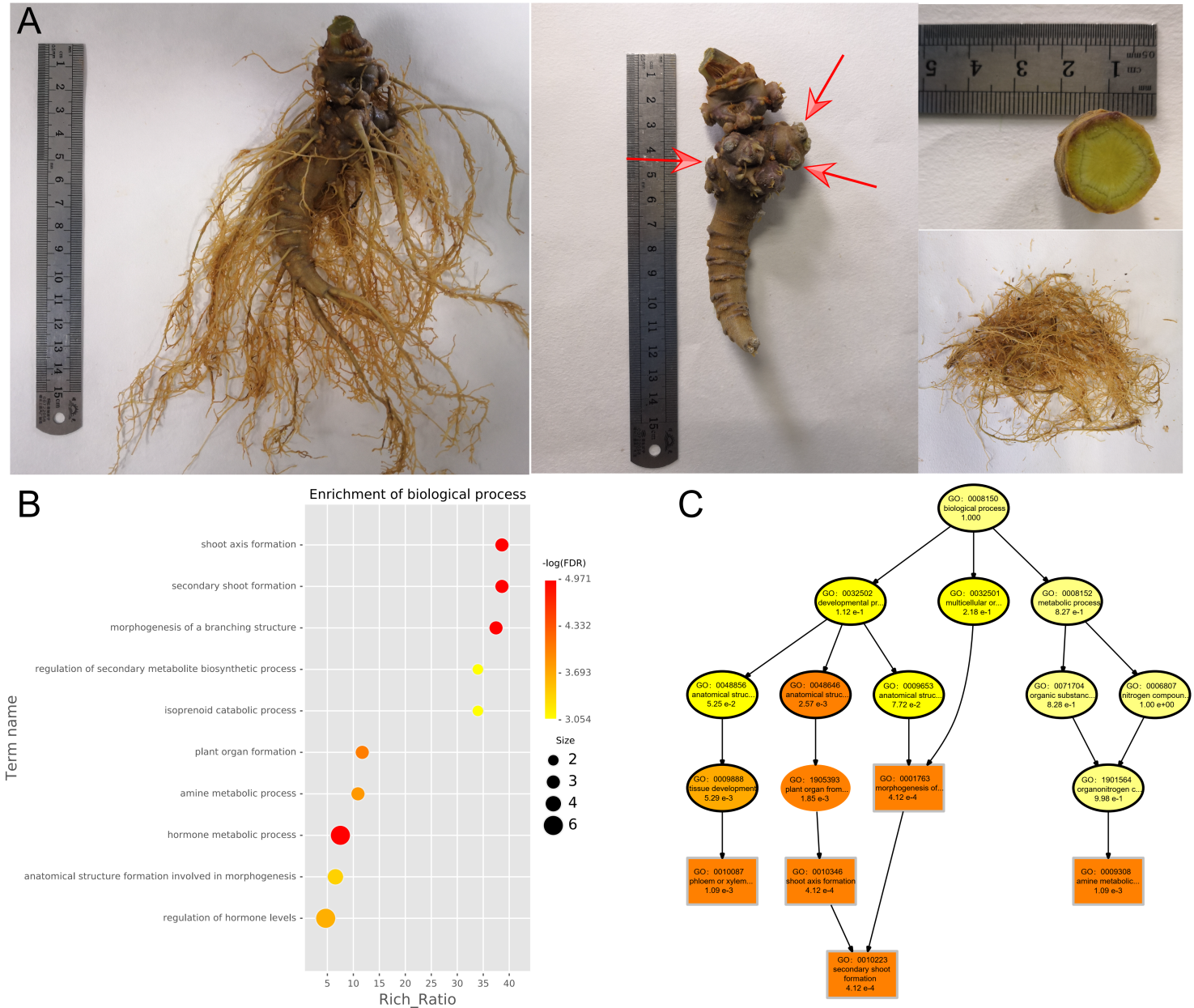
Supplemental Figure 21. Variation analysis of each sample. (A) distribution of each variant type; **(B)** according to the detected SNP loci, the frequency of each mutation type has been counted, taking the data results of One1Leaf, Two1Leaf, Two2Flower, Tri3Stem, Tri1Xylem, Fou2Perid as examples; **(C)** according to the detected InDel loci, the frequency of each InDel length has been counted, taking the data results of One1Leaf, Two1Leaf, Two2Flower, Tri3Stem, Tri1Xylem, Fou2Perid as examples.



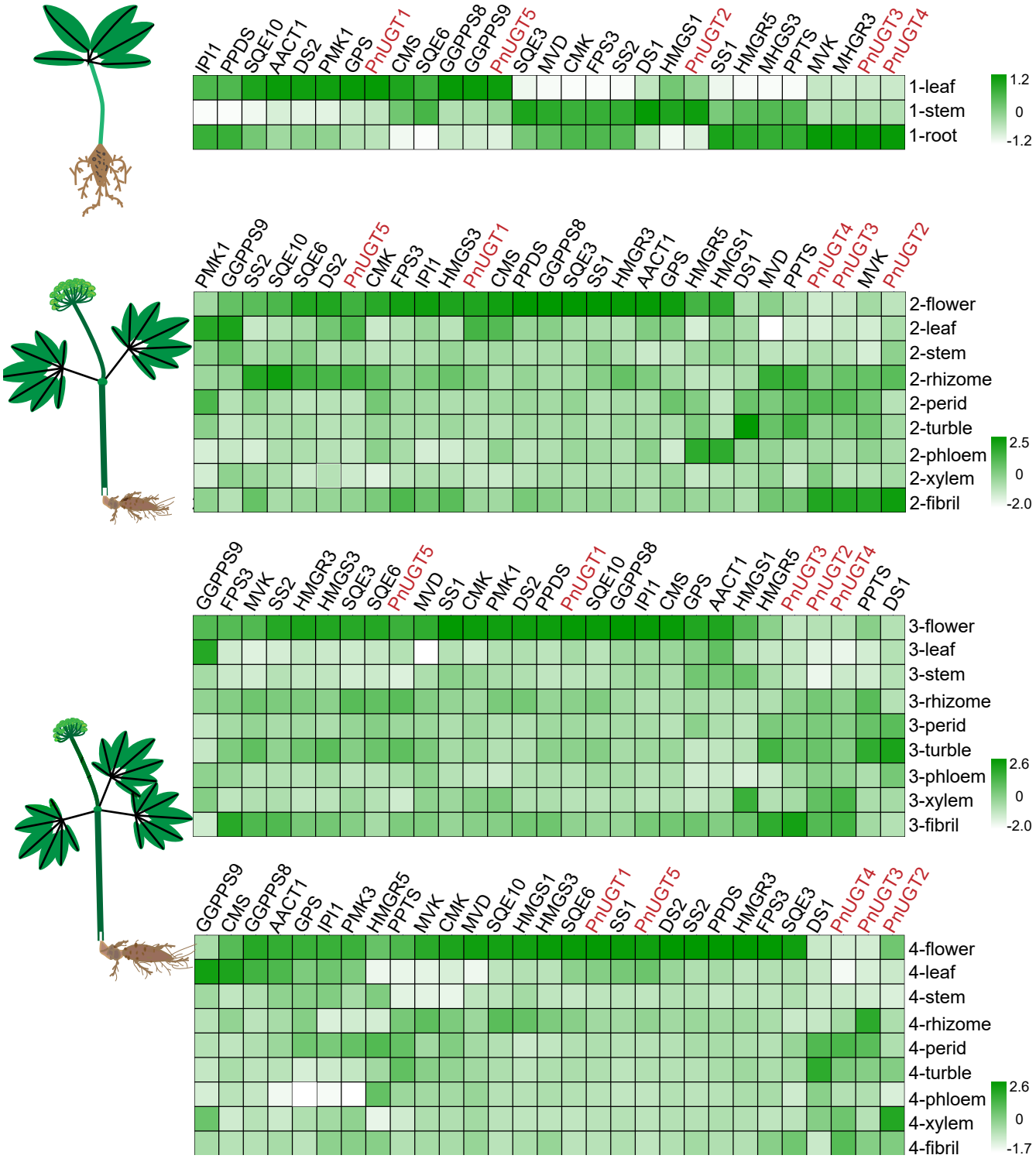
Supplemental Figure 22. Density distribution diagram of gene expression in each transcriptome sample.



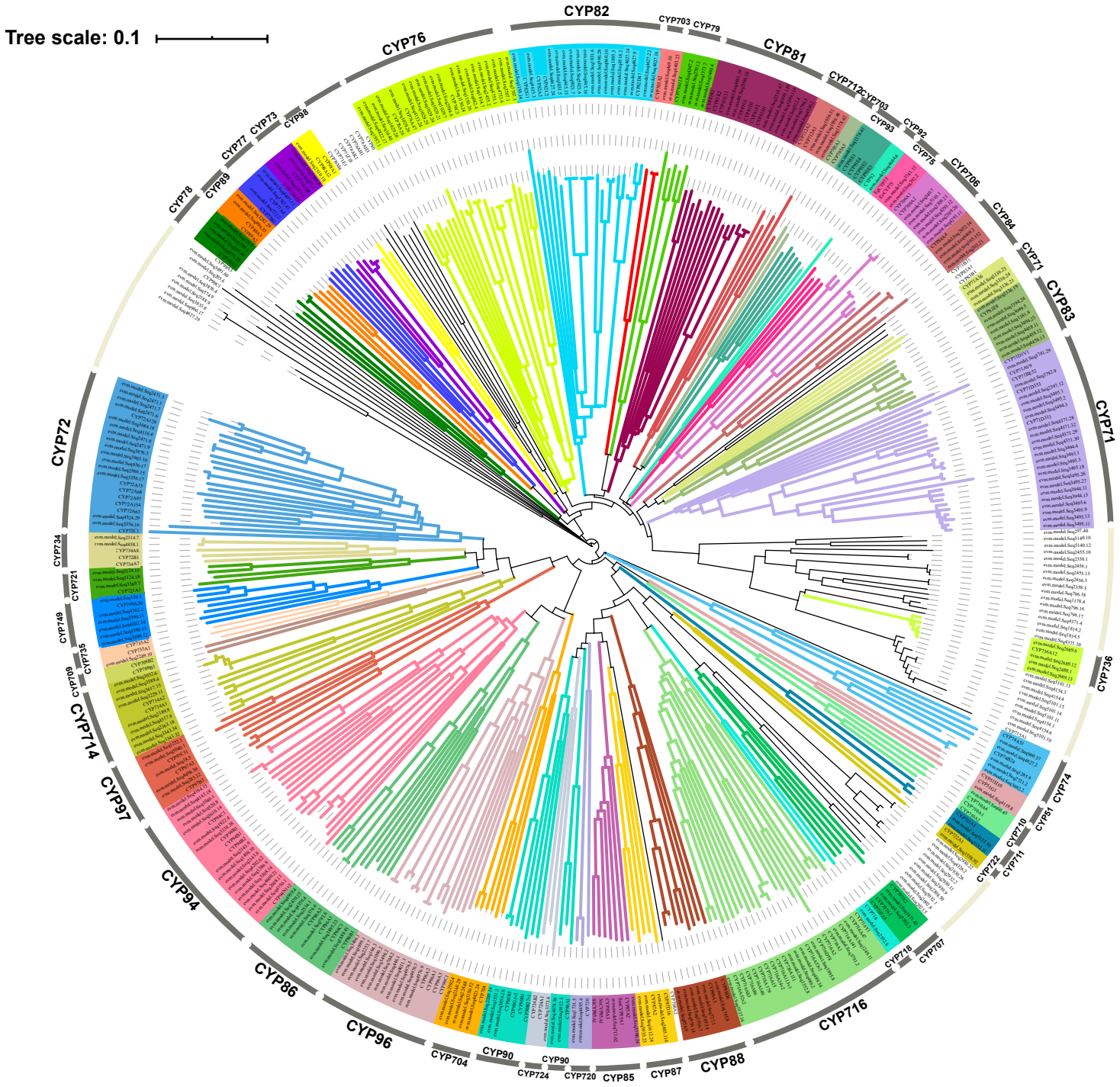
Supplemental Figure 23. Box plot of the overall distribution of gene expression in each transcriptome sample.



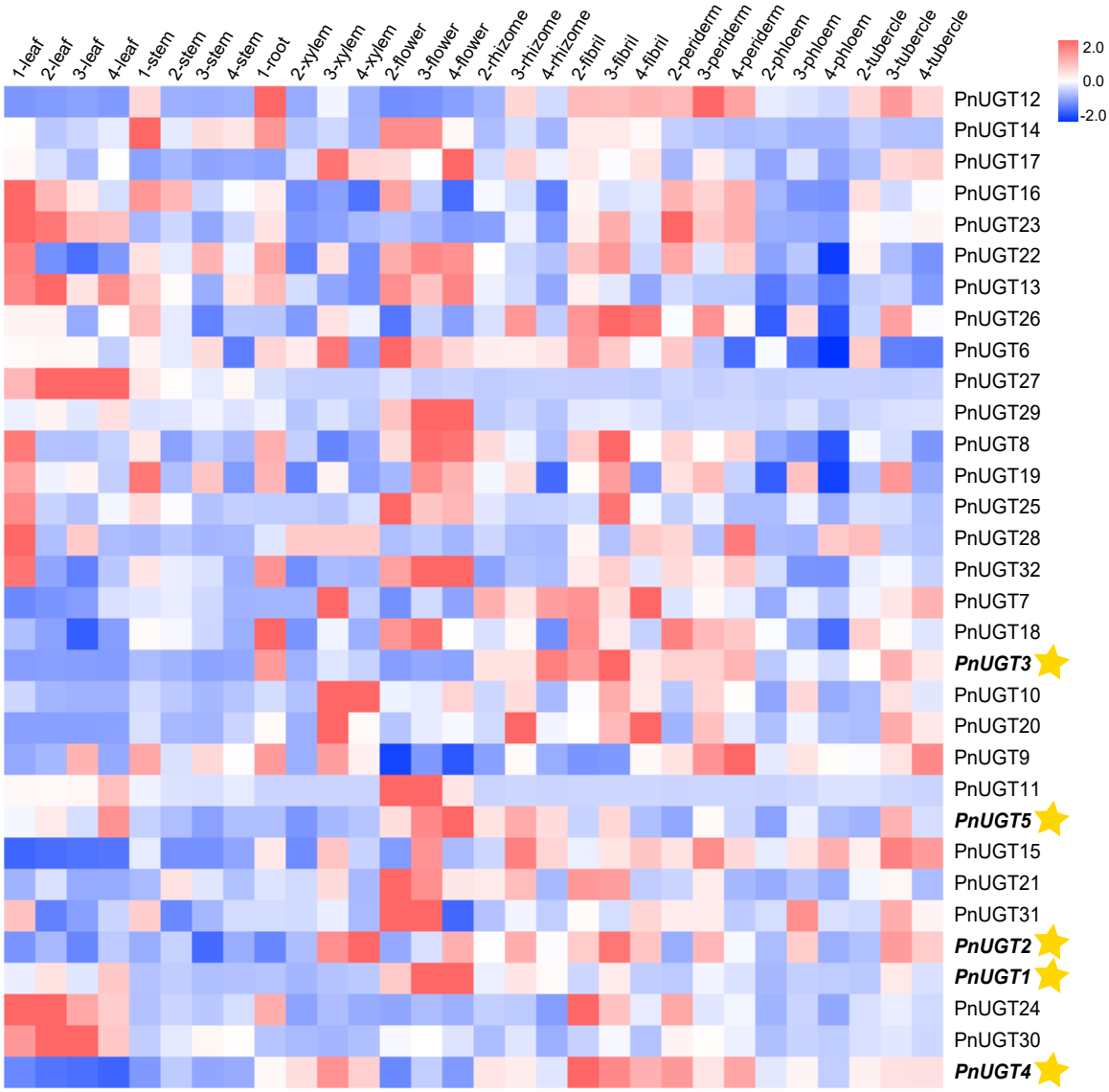
Supplementary Figure 24. The exploration of the molecular mechanism of the formation of *P. notoginseng*'s tubercles. (A) the display of root morphology of *P. notoginseng*, and the red arrow points to the tubercles. **(B)** GO enrichment analysis of DEGs between the periderm and tubercle group. **(C)** the Directed Acyclic Graph (DAG) of GO enrichment analysis, the darker color indicates the more significant enrichment and the red is the most significant. The larger the bubble radius, the higher the rich-ratio value and the redder the color of bubble, the higher the degree of enrichment.



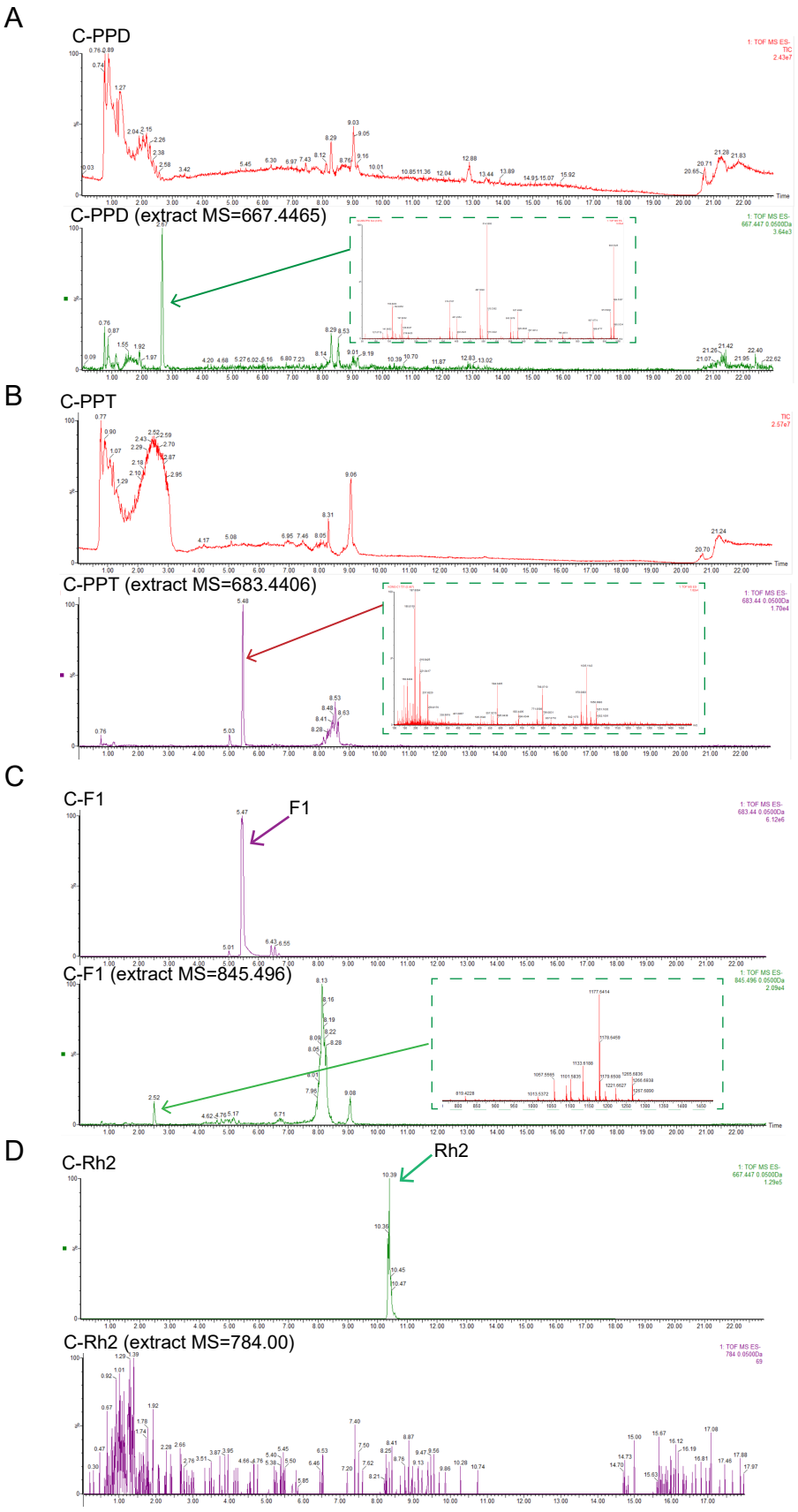
Supplementary Figure 25. Spatial expression profile of key enzyme genes in saponin biosynthesis pathway. The genes in red font are the functional UGT cloned in this study.



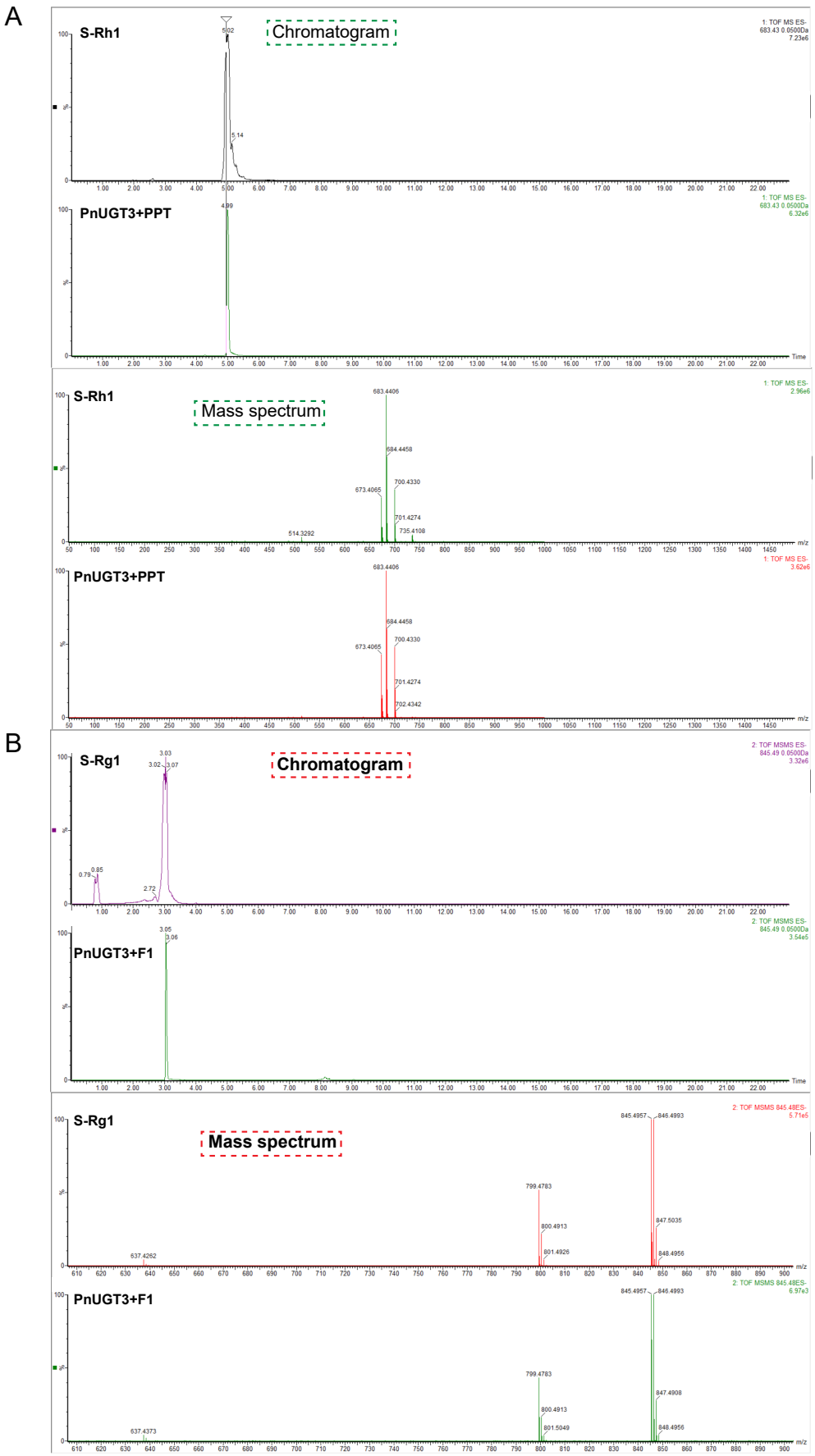
Supplementary Figure 26. Phylogenetic analysis of CYP450 genes in *P. notoginseng* using MEGA-X.



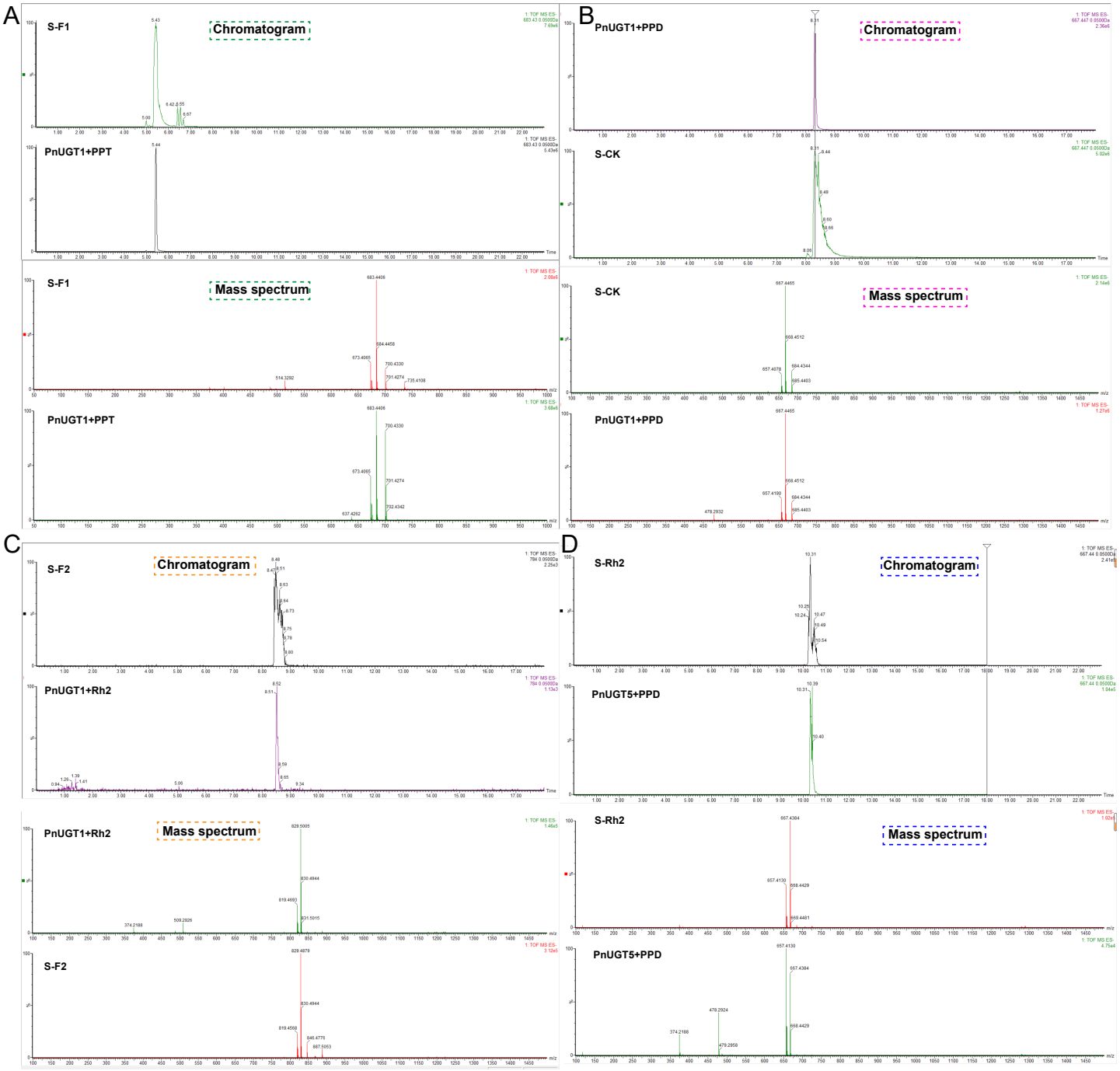
Supplemental Figure 27. Heat map of the expression of the cloned UGT genes in different transcript samples. The genes marked by five-pointed stars are those with catalytic function identified in this study. In the heat map, the relative expression level from high to low (-2 to 2) is represented by the range from blue to red.



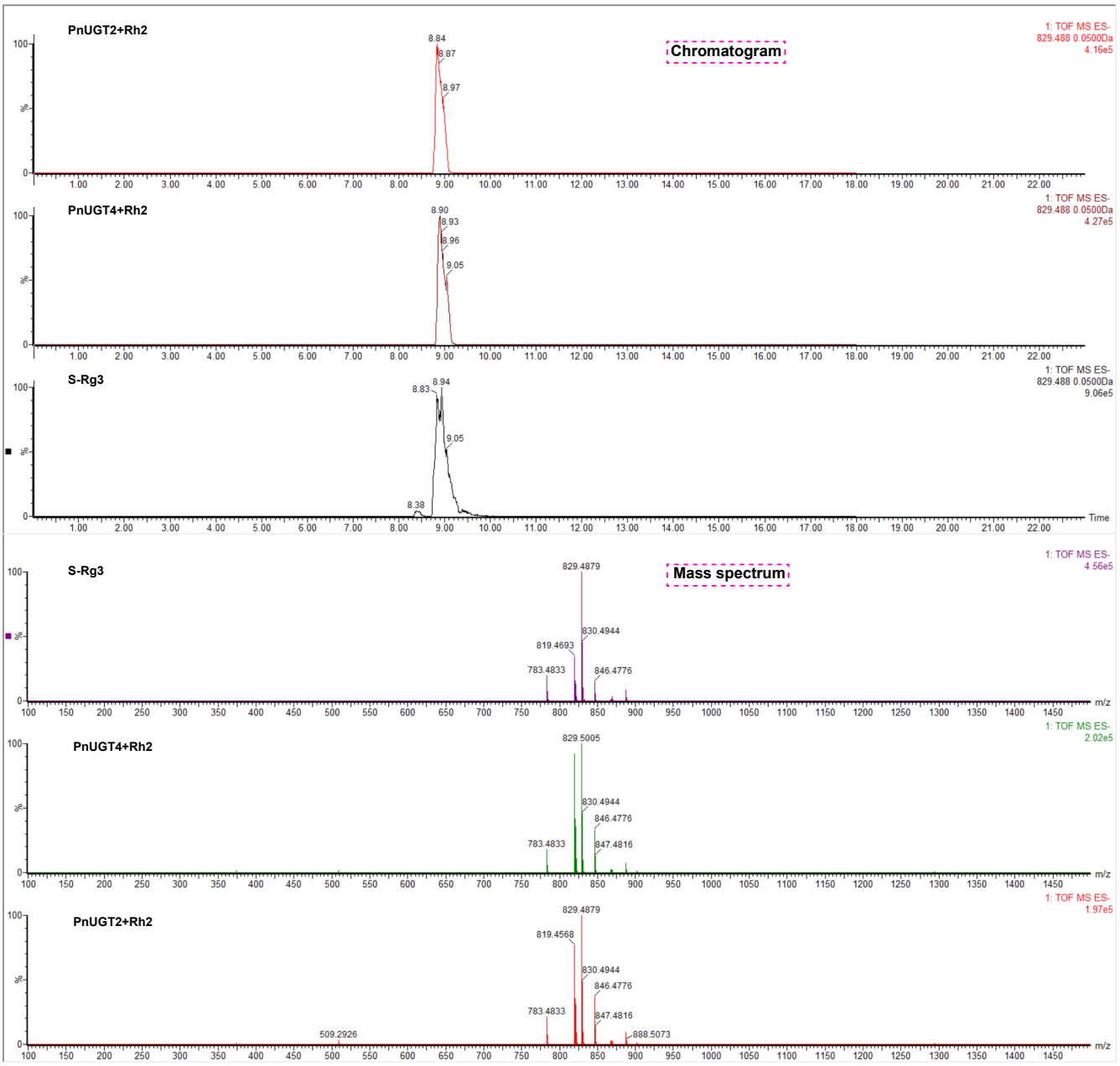
Supplementary Figure 28. The blank control experiments of protein catalytic reaction in this study. **(A)** Analysis results of *E. coli* no-load control with PPD as the catalytic substrate. **(B)** Analysis results of *E. coli* no-load control with PPT as the catalytic substrate. **(C)** Analysis results of *E. coli* no-load control with ginsenoside F1 as the catalytic substrate. **(D)**Analysis results of *E. coli* no-load control with Rh2 as the catalytic substrate. The molecular ion peaks with 667.4465, 683.4406, 845.496 and 784.00 were extracted respectively, and the mass spectrum in the green box did not match with any corresponding glycoside product, indicating no product was formed.



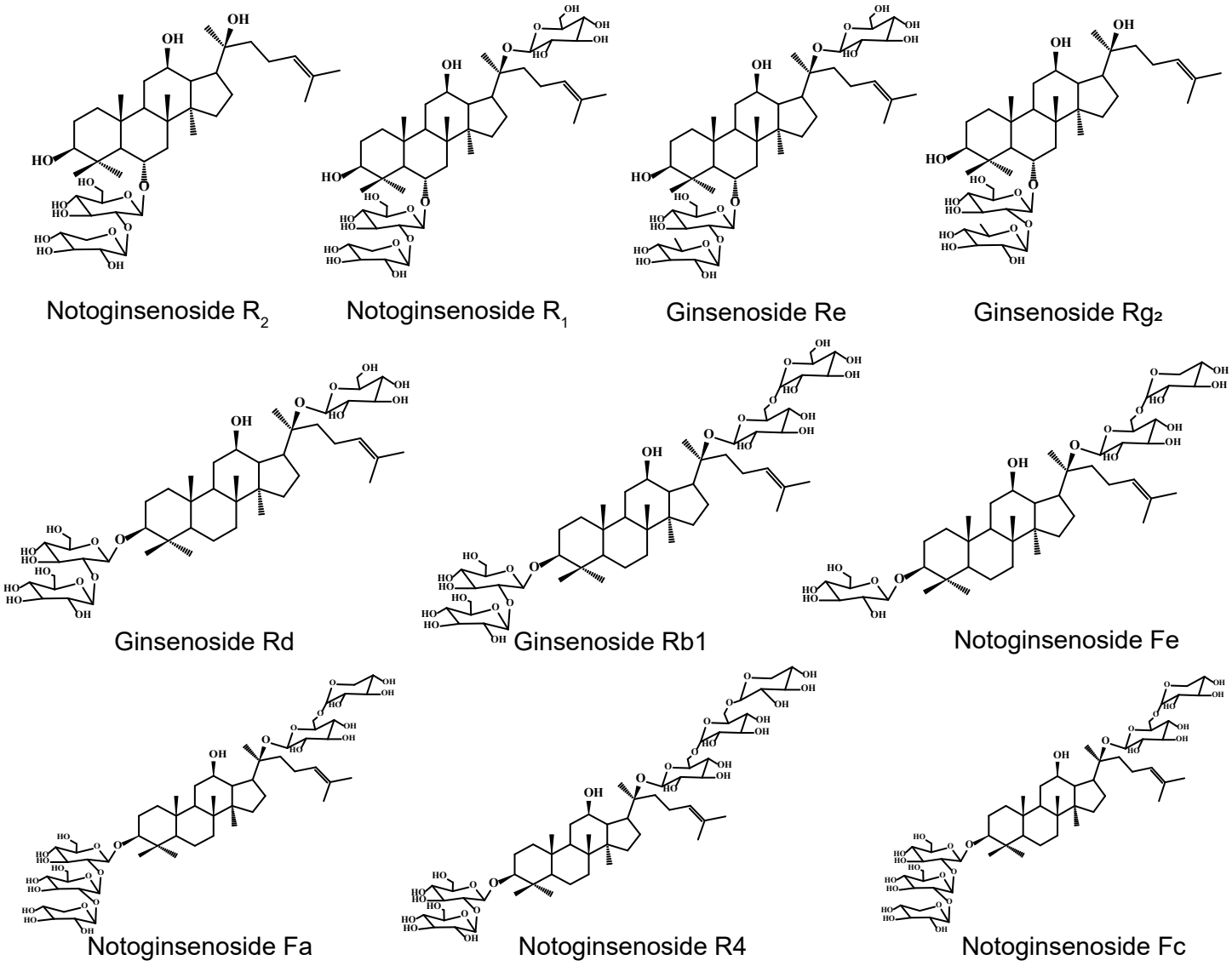
Supplementary Figure 29. UPLC/Q-TOF analysis results of PnUGT3 protein catalytic reaction. **(A)** Chromatograms and mass spectrum of ginsenoside Rh1 standard and PnUGT3 catalytic products using PPT as substrate. **(B)** Chromatograms and mass spectrum of ginsenoside Rg1 standard and PnUGT3 catalytic products using F1 as substrate.



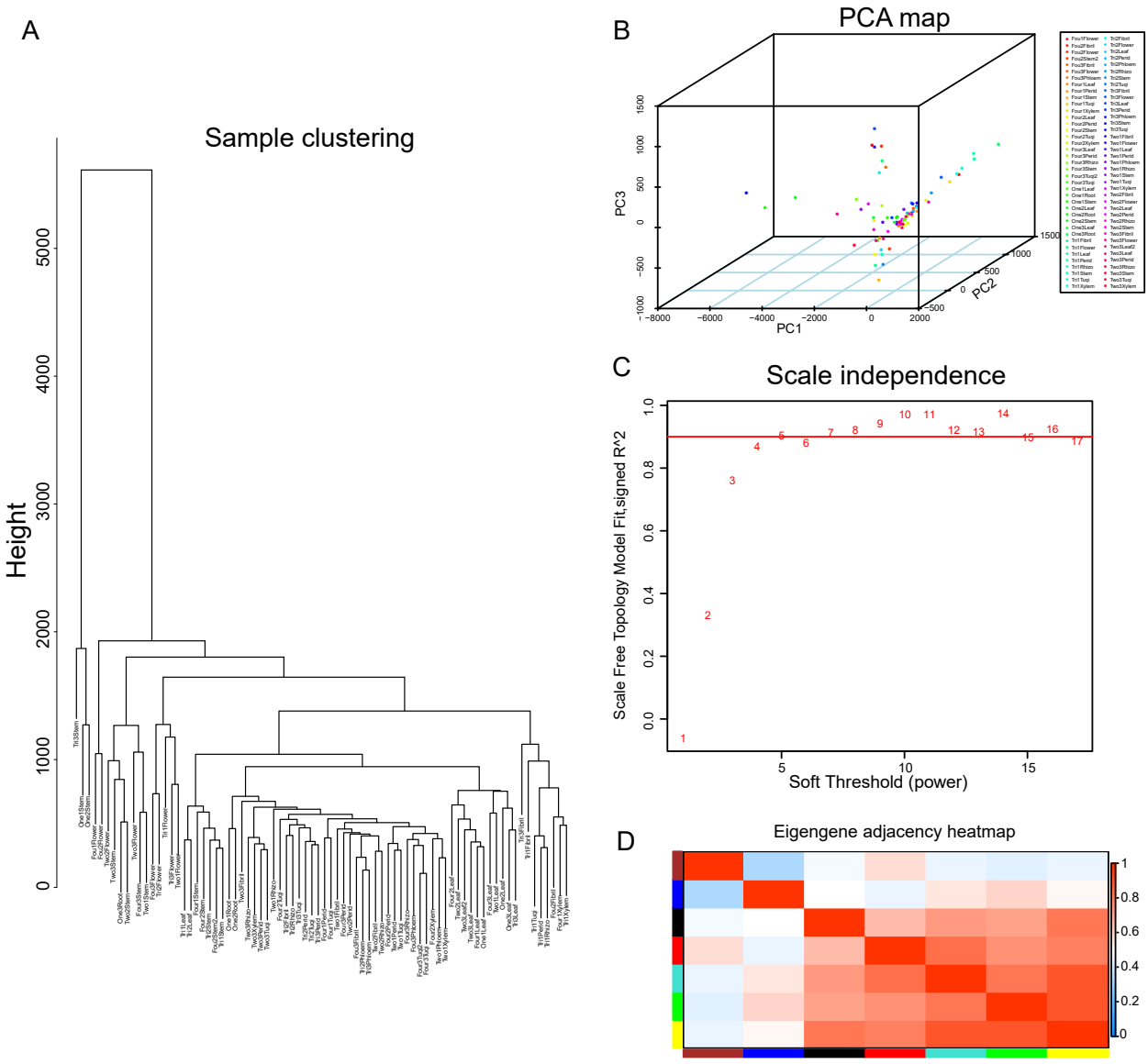
Supplementary Figure 30. UPLC/Q-TOF analysis results of PnUGT1 and PnUGT5 protein catalytic reaction. (A) Chromatograms and mass spectrum of ginsenoside F1 standard and PnUGT1 catalytic products using PPT as substrate. **(B)** Chromatograms and mass spectrum of ginsenoside CK standard and PnUGT1 catalytic products using PPD as substrate. **(C)** Chromatograms and mass spectrum of ginsenoside F2 standard and PnUGT1 catalytic products using Rh2 as substrate. **(D)** Chromatograms and mass spectrum of ginsenoside Rh2 standard and PnUGT5 catalytic products using PPD as substrate.



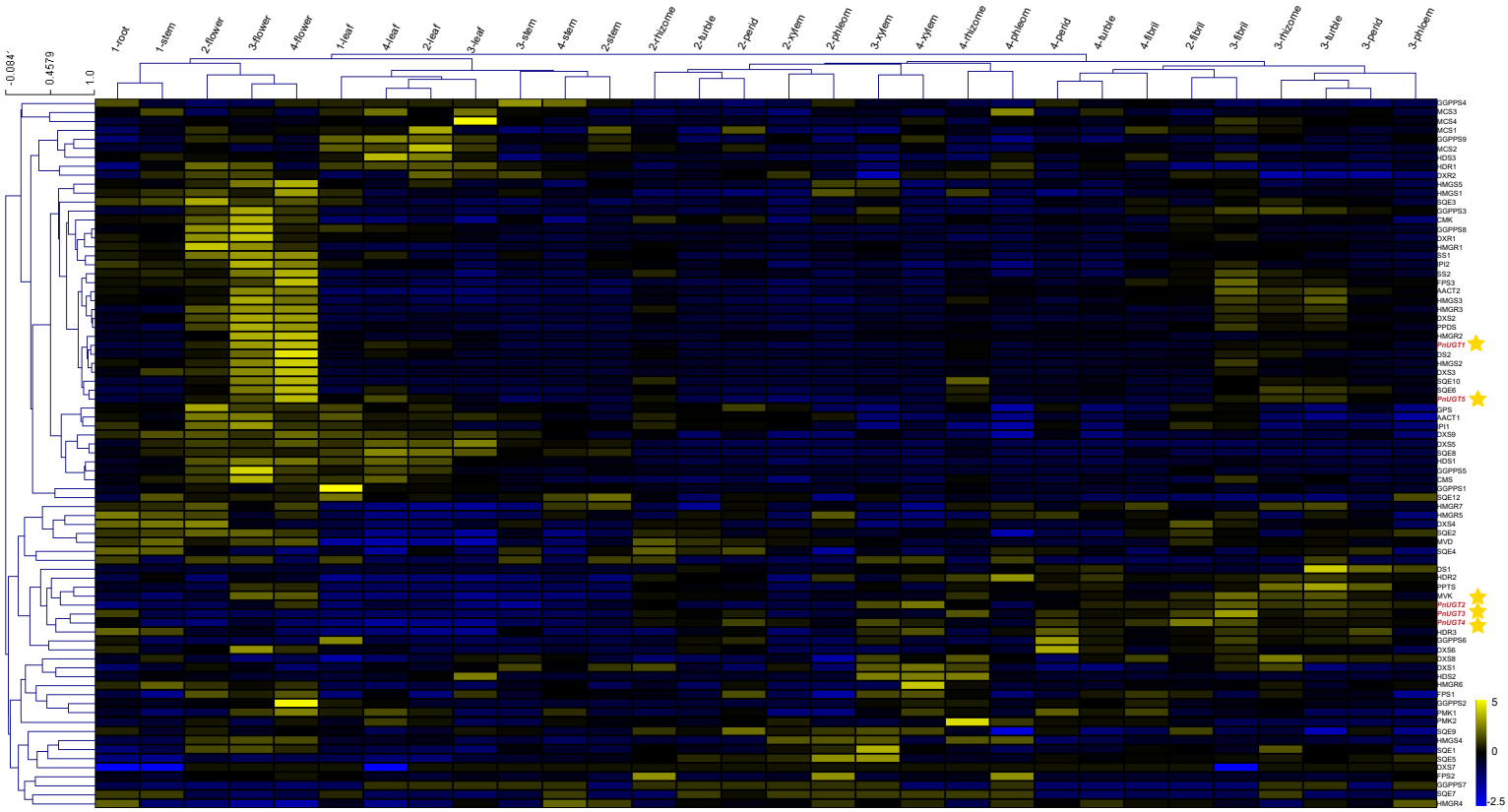
Supplemental Figure 31. UPLC/Q-TOF analysis results of PnUGT2 and PnUGT4 protein catalytic reaction. Chromatograms and mass spectrum of ginsenoside Rg3 standard and PnUGT2 and PnUGT4 catalytic products using Rh2 as substrate.



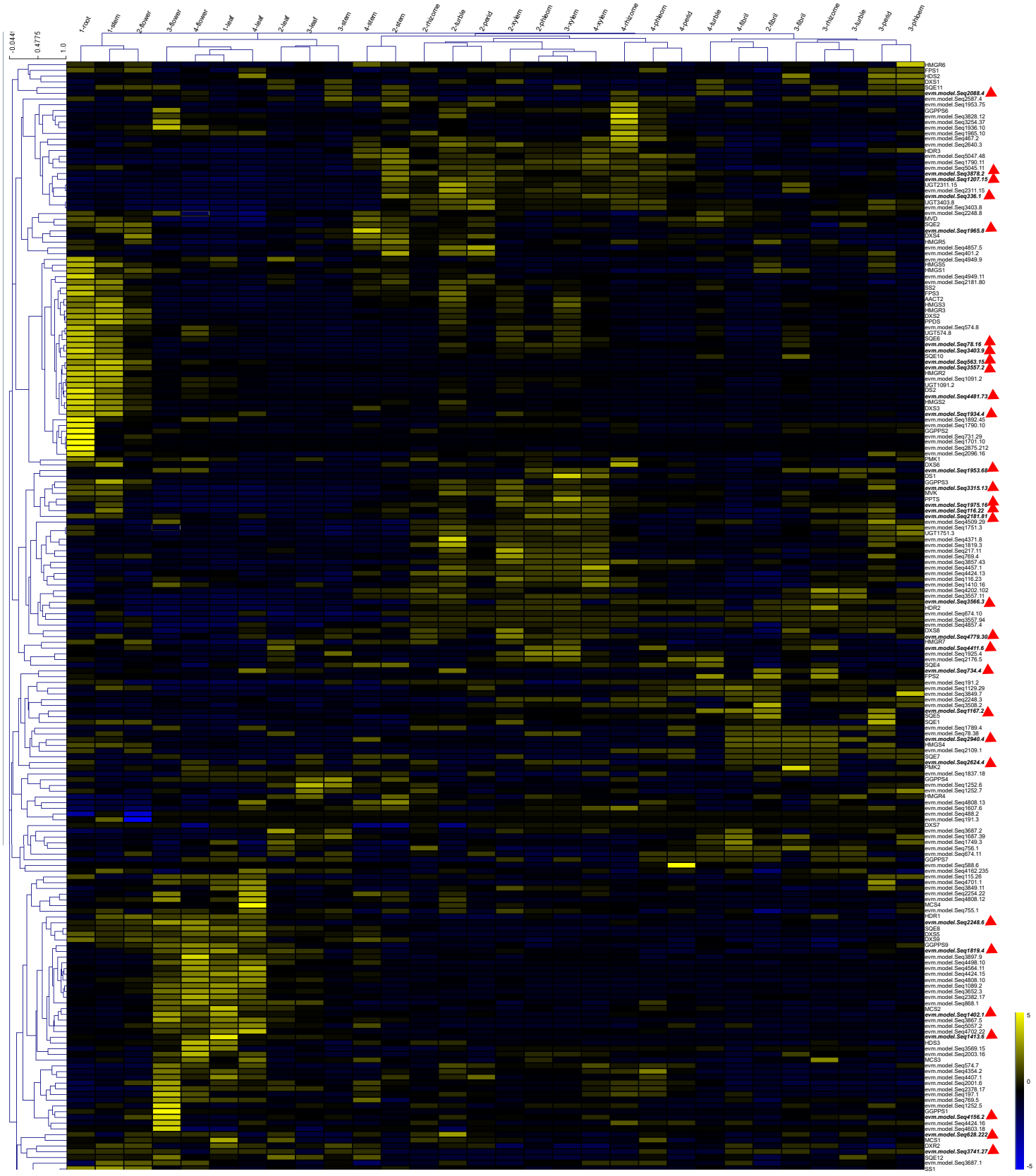
Supplementary Figure 32. The structural formulas of various saponins in *P. notoginseng*.



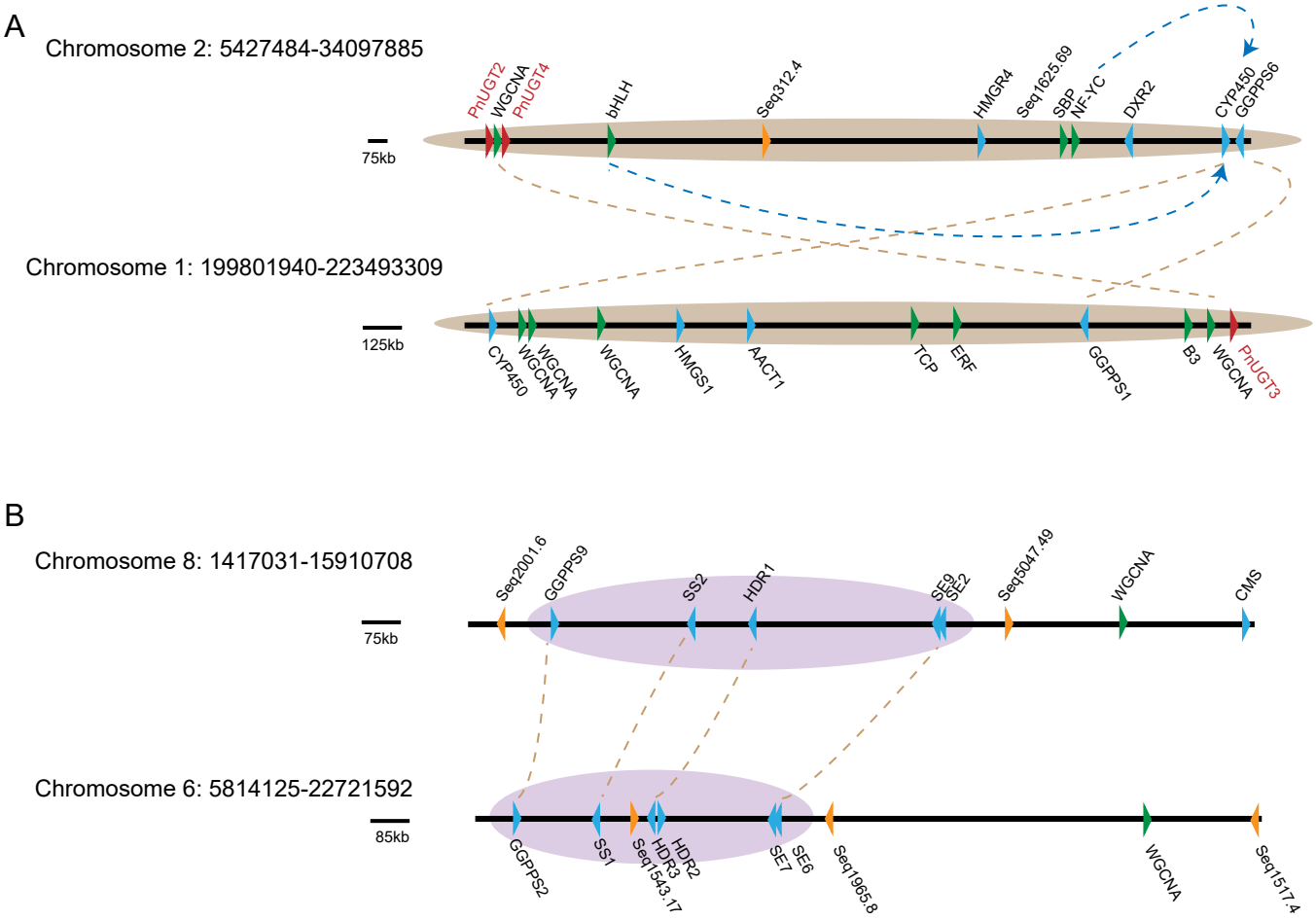
Supplemental Figure 33. WGCAN analysis and characterization of corresponding data.
(A) Construction of the sample clustering evolutionary tree of transcriptome to screen out outliers. **(B)** Construction the PCA map of transcriptome samples. **(C)** Analysis of network topology for various soft-thresholding powers. When we set R^2=0.9, the optimal candidate threshold to reach this height is 10. **(D)** Visualization of the eigengene network representing the relationships among the modules. The redder the color in the heat map, the stronger the correlation between the two modules.



Supplemental Figure 34. Expression profile of key enzyme genes in saponin biosynthesis pathway. The right side of the heatmap shows the evolutionary tree of genes, and genes with similar expression patterns are clustered into one group.



Supplemental Figure 35. Heat map of expression of UGT genes and genes in terpenoid biosynthesis pathway. The right side of the heatmap shows the evolutionary tree of genes, and genes with similar expression patterns are clustered into one group.



Supplemental Figure 36. Gene clusters involved in saponins biosynthesis found in *P. notoginseng* genome. (A) Gene clusters on chromosomes 1, 2 and their correspondence. **(B)** Gene clusters on chromosomes 6, 8 and their correspondence. Orange lines indicate copies of genes with the same function, and blue lines indicate the correlation between transcription factors and pathway genes.

Supplementary Tables

Supplementary Table 1. Estimation of genome size of *P. notoginseng* based on *K-mer* analysis.

Version	V1.0
Total base (Gb)	231.06
K	31
K-mer number	136,285,293,064
K-mer depth	58
Genome size (Gb)	2.38
Revised genome size (Gb)	2.35
Heterozygous ratio (%)	0.58
Repeat (%)	69.05

Supplementary Table 2. Sequencing data statistics of *P. notoginseng*.

Pair-end libraries	Insert size	Total data (Gb)	Reads length (bp)	Sequence coverage (X)
Illumina reads	350 bp	240.22	150	90.31
Pacbio reads	30 kb	284.07	-	106.79
Hi-C	-	340.83	150	128.13
Total	-	865.12	-	325.23

Supplementary Table 3. The Statistics of Pseudomolecule based on Hi-C technique.

Pseudomolecule	Contig Num	Length
chr1	528	295,554,597
chr2	567	268,893,176
chr3	412	240,203,249
chr4	406	234,363,521
chr5	550	229,146,523
chr6	584	216,472,261
chr7	383	216,178,389
chr8	462	203,820,015
chr9	500	199,351,208
chr10	340	194,535,185
chr11	305	179,501,441
chr12	279	176,575,628
Total anchored	5316	2,654,595,193
Unanchored	44	2,772,262

Supplementary Table 4. Statistic of DNA base composition in the *P. notoginseng* genome.

Iterms	Number(bp)	Percent (%)
A	871,740,949	32.76
T	871,858,137	32.77
C	457,971,493	17.21
G	458,576,639	17.24
N	530,400	0.02
GC	916,548,132	34.45
Total	2,660,147,218	100

Supplementary Table 5. Statistics of consistency assessment of the *P. notoginseng* genome.

Sample	PNCCMU201908
Clean Reads	668,000,000
Clean Bases	100,200,000,000
Mapped Reads	666,778,003
Mapped Reads Rate (%)	99.82
Mapped Bases	99,793,880,477
Mapped Bases Rate (%)	99.59
Mean Depth	37.87
Coverage Rate (%)	97.97

Supplementary Table 6. Assessment the gene coverage rate using BUSCO.

Iterms	Number	Percent (%)
Complete BUSCOs (C)	2,049	96.6
Complete and single-copy BUSCOs (S)	1,495	70.5
Complete and duplicated BUSCOs (D)	554	26.1
Fragmented BUSCOs (F)	23	1.1
Missing BUSCOs (M)	49	2.3
Total BUSCO groups searched	2,121	100

Supplementary Table 7. Annotation of repetitive sequences in the *P. notoginseng* genome.

Type	Repeat length (bp)	% of genome
RepeatMasker	482,030,502	18.12
ProteinMask	597,728,373	22.47
Denovo	2,228,207,907	83.76
TRF	145,037,356	5.45
Total	2,283,667,181	85.85

Supplementary Table 8. Summary of repetitive sequences in the *P. notoginseng* genome.

Class	RepeatMasker TEs		RepeatProteinMasker TEs		RepeatModeler TEs		Combined TEs	
Type	**Length(bp)**	**% in genome**	**Length(bp)**	**% in genome**	**Length(bp)**	**% in genome**	**Length(bp)**	**% in genome**
DNA RT	16,404,105	0.62	6,987,822	0.26	28,157,924	1.06	38,999,478	1.47
LINE	2,256,979	0.08	998,015	0.04	3,718,106	0.14	6,035,132	0.23
SINE	3,185	0	0	0	0	0	3,185	0
LTR RT	463,294,570	17.42	525,365,110	19.75	1,533,286,234	57.64	1,566,296,092	58.88
Unknown	41,143	0	0	0	725,955,721	27.29	725,996,018	27.29
Other	1,185,919	0.07	65,048,999	2.45	7,539,129	0.28	71,825,002	2.7
Total	482,030,502	18.12	597,728,373	22.47	2,228,207,907	83.76	2,271,609,517	85.39

DNA RT：DNA retrotransposons; LINE：long interspersed nuclear elements; SINE：short interspersed nuclear elements; LTR RT：long terminal repeat retrotransposons; Unknown：which have been included in the database, but not classified; Other：Which can be classified by RepeatMasker, but don’t belong to the above categories

Supplementary Table 9. Basic statistical results of gene structure prediction of *P. notoginseng* genome.

Gene set		Number of proteins	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	49,549	4,535.36	1,043.08	4.38	238.04	1,033.62
	Genebank	55,883	4,452.17	1,119.06	5.1	219.37	813.69
Homolog	*A. thaliana*	28,931	4379.27	1069.89	4.77	224.13	878.00
	D. carota	32,038	4905.04	1074.87	4.74	226.99	1026.40
	P. ginseng	43,292	4653.92	1000.35	4.51	222.00	1043.04
	P. notoginseng-pub	41238	3031.43	960.98	3.76	255.82	752.13
RNA-seq	cDNA	58,196	5,834.68	925.12	5.09	351.88	990.1
EVM		37,606	5059.63	1202.85	5.21	231.00	917.71

Supplementary Table 10. Basic statistical results of gene structure prediction of *P. notoginseng* and relative species.

Species	Number of proteins	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
P. notoginseng	37,606	5059.63	1202.85	5.21	231.00	917.71
P. ginseng	59,352	4213.49	1120.12	5.01	223.53	772.23
D. carota	32,112	3104.08	1187.75	5.01	237.23	479.28
A. thaliana	27,416	1855.01	1209.65	5.08	238.02	159.09
P. notoginseng-pub	34,369	2705	957	3.80	251.39	622.24

Supplementary Table 11. Statistical results of gene function annotation of *P. notoginseng* genome.

Database	Count	Percentage (%)
BLASTP	29,662	78.88
BLASTX	29,510	78.47
GO	29,931	79.59
KO	11,657	31.00
Map	7,459	19.83
NR	35,926	95.53
NT	26,037	69.24
PFAM	28,457	75.67
eggNOG	25,058	66.63
Total_anno	36,154	96.14
Total_unigene	37,606	100

Supplementary Table 12. Statistical results of non-coding RNA of *P. notoginseng* genome.

Class	Type	Copy	Average length (bp)	Total length (bp)	% of genome
miRNA	miRNA	14430	235.19	3393818	0.12758
tRNA	tRNA	1513	73.99	111954	0.00421
rRNA	18S	165	867.15	143079	0.00538
	28S	314	136.11	42737	0.00161
	5.8S	80	148.46	11877	0.00045
	5S	2459	118.23	290726	0.01093
snRNA	CD-box	7216	105.46	760992	0.02861
	HACA-box	272	146.74	39913	0.0015
	splicing	686	154.29	105842	0.00398

Supplementary Table 13. The Statistics of gene clustering to gene families in various species.

Species	Genes number	Genes in families	Unclustered genes	Family number	Unique families	Average genes per family
P. ginseng	59,352	45,808	13,544	17,112	1,658	2.68
D. carota	32,113	26,151	5,962	13,697	972	1.91
V. vinifera	31,845	23,639	8,206	13,987	936	1.69
C. annuum	35,884	28,781	7,103	13,172	1,243	2.19
G. uralensis	34,445	23,729	10,716	13,597	1,029	1.75
A. thaliana	27,416	23,374	4,042	12,739	777	1.83
O. sativa	42,189	30,038	12,151	13,183	2,239	2.28
P. notoginseng	37,525	30,874	6,651	15,655	1,059	1.97

Supplementary Table 14. Enriched GO terms of genes in *P. notoginseng*-specific families.

Accession	Ontology	Term name	p-value	FDR
GO:0042256	BP	mature ribosome assembly	8.69E-12	6.55E-09
GO:1901566	BP	organonitrogen compound biosynthetic process	8.65E-12	6.55E-09
GO:0034622	BP	cellular macromolecular complex assembly	3.60E-12	6.55E-09
GO:0065003	BP	macromolecular complex assembly	3.02E-11	1.71E-08
GO:0009260	BP	ribonucleotide biosynthetic process	8.70E-11	2.81E-08
GO:0009156	BP	ribonucleoside monophosphate biosynthetic process	6.66E-11	2.81E-08
GO:0046390	BP	ribose phosphate biosynthetic process	8.70E-11	2.81E-08
GO:0042255	BP	ribosome assembly	1.15E-10	3.26E-08
GO:0009124	BP	nucleoside monophosphate biosynthetic process	1.35E-10	3.39E-08
GO:0043933	BP	macromolecular complex subunit organization	1.78E-10	4.03E-08
GO:0044445	CC	cytosolic part	4.10E-12	1.66E-09
GO:0022626	CC	cytosolic ribosome	1.56E-10	3.16E-08
GO:0044391	CC	ribosomal subunit	4.19E-10	5.66E-08
GO:0022625	CC	cytosolic large ribosomal subunit	8.47E-10	8.58E-08
GO:0005829	CC	cytosol	2.56E-09	2.08E-07
GO:0032991	CC	macromolecular complex	9.62E-09	6.49E-07
GO:0005774	CC	vacuolar membrane	1.40E-08	8.11E-07
GO:0044437	CC	vacuolar part	2.61E-08	1.32E-06
GO:0098805	CC	whole membrane	6.13E-08	2.76E-06
GO:1990904	CC	ribonucleoprotein complex	3.78E-07	1.39E-05
GO:0036094	MF	small molecule binding	1.97E-13	1.15E-10
GO:0000166	MF	nucleotide binding	9.22E-13	1.15E-10
GO:0032559	MF	adenyl ribonucleotide binding	5.37E-13	1.15E-10
GO:0043168	MF	anion binding	2.64E-13	1.15E-10
GO:0005524	MF	ATP binding	4.75E-13	1.15E-10
GO:0032553	MF	ribonucleotide binding	1.11E-12	1.15E-10
GO:0008144	MF	drug binding	1.03E-12	1.15E-10
GO:0030554	MF	adenyl nucleotide binding	6.48E-13	1.15E-10
GO:0097367	MF	carbohydrate derivative binding	1.21E-12	1.15E-10
GO:1901265	MF	nucleoside phosphate binding	9.22E-13	1.15E-10

Supplementary Table 15. Enriched GO terms of genes in expanded gene families.

Accession	Ontology	Term name	FDR
GO:0006313	BP	transposition, DNA-mediated	1.32E-60
GO:0032196	BP	transposition	8.91E-49
GO:0006636	BP	unsaturated fatty acid biosynthetic process	1.41E-45
GO:0033559	BP	unsaturated fatty acid metabolic process	5.30E-45
GO:0006310	BP	DNA recombination	1.96E-18
GO:0022900	BP	electron transport chain	7.47E-12
GO:0006633	BP	fatty acid biosynthetic process	8.70E-12
GO:0019684	BP	photosynthesis, light reaction	1.62E-09
GO:0072330	BP	monocarboxylic acid biosynthetic process	5.02E-09
GO:0042773	BP	ATP synthesis coupled electron transport	1.57E-08
GO:0070469	CC	respiratory chain	1.53E-10
GO:0031224	CC	intrinsic component of membrane	1.56E-06
GO:0016021	CC	integral component of membrane	1.61E-06
GO:0031966	CC	mitochondrial membrane	2.59E-06
GO:0031897	CC	Tic complex	2.72E-05
GO:0005743	CC	mitochondrial inner membrane	3.62E-05
GO:0019822	CC	P4 peroxisome	9.83E-05
GO:0044425	CC	membrane part	0.000117
GO:0005739	CC	mitochondrion	0.000153
GO:0000799	CC	nuclear condensin complex	0.000217
GO:0004803	MF	transposase activity	4.81E-61
GO:0016717	MF	oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water	1.29E-51
GO:0102985	MF	Delta12-fatty-acid desaturase activity	1.08E-40
GO:0008234	MF	cysteine-type peptidase activity	8.78E-27
GO:0008270	MF	zinc ion binding	9.82E-23
GO:0019863	MF	IgE binding	3.20E-18
GO:0019865	MF	immunoglobulin binding	3.20E-18
GO:0140097	MF	catalytic activity, acting on DNA	2.89E-15
GO:0008483	MF	transaminase activity	1.96E-13
GO:0016769	MF	transferase activity, transferring nitrogenous groups	1.96E-13

Supplementary Table 16. Enriched GO terms of genes in contracted gene families.

Accession	Ontology	Term name	FDR
GO:0006468	BP	protein phosphorylation	2.63404E-66
GO:0016310	BP	phosphorylation	1.3388E-65
GO:0006796	BP	phosphate-containing compound metabolic process	5.17223E-36
GO:0006793	BP	phosphorus metabolic process	6.21129E-36
GO:0006464	BP	cellular protein modification process	1.40587E-24
GO:0036211	BP	protein modification process	1.40587E-24
GO:0005975	BP	carbohydrate metabolic process	6.30804E-24
GO:0051274	BP	beta-glucan biosynthetic process	3.40873E-23
GO:0051273	BP	beta-glucan metabolic process	9.58801E-23
GO:0048544	BP	recognition of pollen	7.66426E-21
GO:0005886	CC	plasma membrane	1.35E-59
GO:0031224	CC	intrinsic component of membrane	2.26E-36
GO:0016021	CC	integral component of membrane	2.27E-36
GO:0016020	CC	membrane	5.87E-33
GO:0044425	CC	membrane part	2.47E-28
GO:0009341	CC	beta-galactosidase complex	8.93E-13
GO:0005576	CC	extracellular region	3.69E-11
GO:0016459	CC	myosin complex	3.69E-11
GO:0044459	CC	plasma membrane part	6.17481E-10
GO:0000148	CC	1,3-beta-D-glucan synthase complex	1.82574E-08
GO:0017076	MF	purine nucleotide binding	7.86457E-88
GO:0032555	MF	purine ribonucleotide binding	7.86457E-88
GO:0035639	MF	purine ribonucleoside triphosphate binding	9.1967E-88
GO:0032553	MF	ribonucleotide binding	4.82089E-87
GO:0008144	MF	drug binding	5.31098E-86
GO:0097367	MF	carbohydrate derivative binding	4.96901E-85
GO:0032559	MF	adenyl ribonucleotide binding	4.96901E-85
GO:0030554	MF	adenyl nucleotide binding	8.50134E-85
GO:0005524	MF	ATP binding	1.24308E-84
GO:0043168	MF	anion binding	8.90045E-80

Supplementary Table 17. Copy number variation of genes involved in the ginsenoside biosynthesis in the *P. notoginseng* and seven other plant species.

Gene	PN	PG	DC	CA	VV	GU	AT	OS
AACT	2	4	2	2	2	3	2	2
HMGS	5	8	1	6	7	3	1	3
HMGR	7	17	3	10	3	4	2	3
MVK	1	2	1	1	1	1	1	1
PMK	2	4	2	0	1	2	0	1
MVD	1	2	1	2	1	1	2	2
DXS	9	13	5	3	8	5	3	3
DXR	2	4	1	1	1	3	1	1
CMS	1	2	2	1	1	1	1	1
CMK	1	2	1	1	1	1	1	1
MCS	4	4	1	1	1	1	1	1
HDS	3	4	2	1	1	1	1	1
HDR	3	6	3	3	1	1	1	2
IDI	2	3	1	3	2	0	2	2
GPS	1	2	1	1	1	2	1	2
FPS	3	5	1	2	1	1	2	5
GGPS	9	16	7	5	3	3	12	3
SS	2	4	1	2	1	2	2	2
SQE	12	25	7	7	8	8	9	4
DS	2	4	1	1	2	1	1	1
CYP450	336	482	311	480	410	257	248	360
GT	158	222	117	4	228	91	112	193
Total	566	835	472	537	685	392	406	594

PN: *P. notoginseng*; PG: *P. ginseng*; DC: *D. carota*; CA: *C. annuum*; VV: *V. vinifera*; GU: *G. uralensis*; AT: *A. thaliana*; OS: *O. sativa*

Supplementary Table 18. *Ks* values and duplication times of genes involved in ginsenoside biosynthesis in *P. notoginseng*.

GENE	Paralog1	Paralog2	*Ka*	*Ks*	MYA
AACT	Seq3227.16	Seq4366.20	0.1	1.4065	108.2
DS	Seq60.43	Seq2181.91	0.1218	0.608	46.8
DXR	Seq3336.90	Seq4886.4	0.0477	0.3082	23.7
DXS	Seq1892.29	Seq3392.7	0.1258	0.4184	32.2
DXS	Seq125.2	Seq2176.8	2.0513	0.4374	33.6
DXS	Seq125.2	Seq2599.2	2.0513	0.4374	33.6
DXS	Seq4710.4	Seq4702.20	0.0005	0.464	35.7
DXS	Seq1892.29	Seq3331.15	0.0541	0.5084	39.1
FPS	Seq4090.9	Seq4036.46	0.2527	0.421	32.4
GGPPS	Seq3857.39	Seq1249.12	0.0964	0.402	30.9
GGPPS	Seq2027.16	Seq3857.39	0.1699	1.6357	125.8
HDR	Seq580.1	Seq1543.10	0.0722	0.4323	33.3
HDS	Seq1699.17	Seq3612.118	0.0246	0.2446	18.8
HMGR	Seq1207.10	Seq2093.13	0.0239	0.3989	30.7
HMGS	Seq1609.4	Seq2103.7	0.0185	0.3479	26.8
HMGS	Seq3714.31	Seq1609.4	0.0488	1.4359	110.5
IPI	Seq4782.43	Seq4782.31	0.0277	0.0417	3.2
MCS	Seq311.72	Seq527.2	0.0731	0.5548	42.7
PMK	Seq1071.3	Seq2875.187	0.2405	0.3632	27.9
SQE	Seq734.26	Seq2337.9	0.0447	0.5602	43.1
SQE	Seq734.26	Seq3238.6	0.1236	1.5512	119.3
SS	Seq1220.5	Seq5026.3	0.0469	0.197	15.2
UGT	Seq1751.3	Seq3403.8	0.0454	0.0872	6.7
UGT	Seq2311.15	Seq1091.2	0.064	0.1306	10

Supplementary Table 22. Representative genes which are highly expressed in tubercle group.

Gene ID	P value	Annotation description	GO terms
Seq1082.27	5.9E-06	carotenoid cleavage dioxygenase 7	GO:0016121|carotene catabolic process; GO:0010223|secondary shoot formation; GO:1901601|strigolactone biosynthetic process;
Seq4384.8	0.0002419	carotenoid cleavage dioxygenase 8	GO:0016121|carotene catabolic process; GO:0010223|secondary shoot formation; GO:1901601|strigolactone biosynthetic process;
Seq2249.10	0.013745	Cytokinin hydroxylase	GO:0033466|trans-zeatin biosynthetic process;
Seq2362.18	0.013745	Cytokinin dehydrogenase 6	GO:0009690|cytokinin metabolic process; GO:0010103|stomatal complex morphogenesis;
Seq4266.75	0.002591	Expansin-A4	GO:0009664|plant-type cell wall organization; GO:0006949|syncytium formation;
Seq1722.34	0.001241	Protein WALLS ARE THIN 1	GO:0009851|auxin biosynthetic process; GO:0010315|auxin efflux; GO:0009734|auxin-activated signaling pathway; GO:0071555|cell wall organization;

Supplementary Table 23. Statistics of transcription factors in *P. notoginseng* genome.

TF family	Gene copy number
bHLH	188
ERF	175
NAC	141
MYB	128
C2H2	114
MYB_related	105
bZIP	91
GRAS	91
WRKY	90
G2-like	72
HD-ZIP	71
C3H	57
LBD	57
Trihelix	57
FAR1	52
B3	49
AP2	40
Dof	40
ARF	34
GATA	33
SBP	33
TCP	33
M-type_MADS	28
HSF	25
TALE	25
NF-YB	21
HB-other	19
NF-YC	19
WOX	18
MIKC_MADS	17
Nin-like	17
GRF	15
NF-YA	15
ZF-HD	15
GeBP	14
BES1	13
DBB	13
E2F/DP	13
BBR-BPC	12
CO-like	12
CPP	12

CAMTA	11
ARR-B	10
YABBY	10
HRT-like	8
SRS	7
EIL	6
LSD	5
RAV	5
STAT	3
HB-PHD	2
LFY	2
NF-X1	2
S1Fa-like	2
Whirly	2
SAP	1
VOZ	1
Total	2150

Supplementary Table 25. The CYP450 genes used to construct phylogenetic tree in this research.

Subfamily	Gene name	Genbank number	Reference
CYP51	CYP51G1	DQ335779.1	(Li et al., 2007)
	CYP51H10	DQ680852.1	(Qi et al., 2006)
CYP710	CYP710A1	NM_129002.3	(Lin et al., 1999)
	CYP710A4	NM_128445.3	(Lin et al., 1999)
	CYP710A2	NM_129001.3	(Lin et al., 1999)
CYP711	MAX1(CYP711A1)	NM_179743.2	(Lin et al., 1999)
CYP74	CYP74A51	LC063857.1	unpublished
	CYP74B24	LC063856.1	unpublished
CYP97	CYP97A3	NM_102914.3	(Theologis et al., 2000)
	CYP97C11	EU849604.1	(Stigliani et al., 2011)
	CYP97B3	NM_117600.6	(Mayer et al., 1999)
CYP704	CYP704	AY779540.1	(Ro et al., 2005)
CYP94	CYP94B1	NM_125740.3	(Tabata et al., 2000)
	CYP94B2	NM_111056.3	(Salanoubat et al., 2000)
	CYP94B3	NM_114710.3	(Salanoubat et al., 2000)
	CYP94C1	NM_128328.3	(Lin et al., 1999)
	CYP94N1v2	KJ869255.1	(Augustin et al., 2015)
CYP86	CYP86A1	MF197861.1	(Shi et al., 2018)
	CYP86A2	NM_116260.4	(Mayer et al., 1999)
	CYP86B1	NM_122225.3	(Tabata et al., 2000)
	CYP86C1	NM_102298.3	(Theologis et al., 2000)
CYP96	CYP96A1	NM_127882.3	(Lin et al., 1999)
	CYP96C1	AJ238402.1	(Oudin et al., 1999)
	CYP96A2	NM_119369.4	(Mayer et al., 1999)
	CYP96A3	NM_105208.1	(Theologis et al., 2000)
CYP88	CYP88D6	MG888351.1	unpublished
	CYP88A3	NM_100394.4	(Theologis et al., 2000)
CYP716	CYP716A52v2	JX036032.1	(Han et al., 2012)
	CYP716A83	KU878849.1	unpublished
	CYP716A86	KU878848.1	unpublished
	CYP716A14v2	KF309251.1	unpublished
	CYP716A140	KU878853.1	unpublished
	CYP716A15	AB619802.1	(Fukushima et al., 2011)
	CYP716A179	LC157867.1	(Tamura et al., 2016)
	CYP716A113v1	KU878866.1	unpublished
	CYP716A111	KY047600.1	unpublished
	CYP716A1	NM_123002.2	(Tabata et al., 2000)
	CYP716A2	NM_123005.2	(Tabata et al., 2000)
	CYP716A141	KU878855.1	(Tamura et al., 2017)
	CYP716Y1	KC963423.1	(Moses et al., 2014)

	CYP716A53v2	JX036031.1	(Han et al., 2012)
	CYP716A47	JN604536.1	(Han et al., 2011)
	CYP716D58	LC209201.1	(Tamura et al., 2017)
	CYP90G1v3	KJ869260.1	(Augustin et al., 2015)
	CYP90B27v1	KJ869252.1	(Augustin et al., 2015)
CYP90	CYP90A1	GU326353.1	unpublished
	CYP90B1	KX168703.1	unpublished
	CYP90C1	NM_001342408.1	(Mayer et al., 1999)
	CYP90B3	AB244039.1	unpublished
	CYP707A1	AB122149.1	(Saito et al., 2004)
CYP707	CYP707A2	NM_128466.4	(Lin et al., 1999)
	CYP707A3	AB122150.1	(Saito et al., 2004)
	SmCYP85A1	KP337712.1	(Chen et al., 2014)
	CYP85A2	AB087801.1	(Nomura et al., 2005)
CYP85	CYP85A3	NM_001247591.1	(Nomura et al., 2005)
	NtCYP85A1	DQ649022.1	unpublished
	SlCYP85A1	NM_001329859.1	(Li et al., 2016)
CYP720	CYP720	KJ624415.1	(Pham et al., 2016)
CYP724	CYP724A1	NM_121444.4	(Tabata et al., 2000)
	CYP724B2	AB244038.1	(Aoki et al., 2010)
CYP87	CYP87D16	KF318735.1	unpublished
	CYP87A2	NM_001198045.1	(Theologis et al., 2000)
CYP722	CYP722A1	NM_101819.6	(Theologis et al., 2000)
CYP718	CYP718	NM_129846.3	(Lin et al., 1999)
CYP708	CYP708A2	NM_001344755.1	(Tabata et al., 2000)
	CYP72A63	AB558146.1	(Seki et al., 2011)
	CYP72A154	AB558153.1	(Seki et al., 2011)
	CYP72A15	NM_112330.4	(Salanoubat et al., 2000)
CYP72	CYP72A67	DQ335780.1	(Li et al., 2007)
	CYP72A68	DQ335782.1	(Li et al., 2007)
	CYP72C1	NM_001332275.1	(Theologis et al., 2000)
	CYP72A129	JN604542.1	(Han et al., 2011)
	CYP72B1	NM_128228.4	(Salanoubat et al., 2000)
CYP709	CYP709B1	NM_130264.2	(Lin et al., 1999)
	CYP709B2	MF463434.1	(Chen et al., 2018)
CYP735	CYP735A1	NM_123206.3	(Tabata et al., 2000)
	CYP735A2	NM_105381.5	(Theologis et al., 2000)
CYP734	CYP734A7	NM_001247011.2	(Ohnishi et al., 2006; Vasav and Barvkar, 2019)
	CYP734A8	NM_001247808.2	(Ohnishi et al., 2006)
CYP715	CYP715A1	NC_003076.8	(Tabata et al., 2000)
CYP721	CYP721A1	NM_106169.4	(Theologis et al., 2000)
CYP714	CYP714A1	NM_122400.3	(Tabata et al., 2000)
	CYP714A2	NM_122399.3	(Tabata et al., 2000)

CYP749	CYP749A20	JN604538.1	(Han et al., 2011)
CYP73	CYP73A19	NM_001279222.2	(Overkamp et al., 2000)
	CYP73A100	JN604543.1	(Han et al., 2011)
CYP98	SbCYP98A1	AF029856.1	(Bak et al., 1998)
	SbCYP98A12	AJ583532.1	(Morant et al., 2007)
CYP736	CYP736A12	JN604539.1	(Han et al., 2011)
CYP78	CYP78A5	NM_101240.4	(Theologis et al., 2000)
CYP703	CYP703A2	NM_100010.3	(Theologis et al., 2000)
CYP75	EoCYP75	HQ268505.1	unpublished
	EgCYP75	U72654.2	unpublished
CYP76	CYP76AH1	JX422213.1	(Guo et al., 2013)
	CYP76M6	Q6Z517	(International Rice Genome Sequencing, 2005)
	CYP76AH3	KR140168.1	(Guo et al., 2016a)
	CYP76AK1	KR140169.1	(Guo et al., 2016a)
	CYP76A26	KF591593.1	(Salim et al., 2014)
	CYP76C3	NM_130120.4	(Lin et al., 1999)
	CYP76A47	MH124060.1	(Wang et al., 2019)
CYP77	CYP77A4	NM_120548.3	(Tabata et al., 2000)
	CYP77B1	NM_101033.4	(Theologis et al., 2000)
CYP92	CYP92	KC841857.1	unpublished
CYP71	CYP71D353	KF460438.1	(Krokida et al., 2013)
	CYP71A16	NM_123623.5	(Tabata et al., 2000)
	CYP71AV9	KF752453.1	(Eljounaidi et al., 2014)
	CYP71E1	AF029858.1	(Kahn et al., 1997)
	CYP71D313	JN604541.1	(Han et al., 2011)
	CYP71D1V1	JN613015.1	(Huang et al., 2012)
	CYP71BE52	KT157042.1	(Trikka et al., 2015)
	CYP71Z18	NM_001147894.2	(Mao et al., 2016)
	CYP71B31	NM_115190.1	(Salanoubat et al., 2000)
CYP706	CYP706A1	NM_118395.3	(Mayer et al., 1999)
	CYP706A2	NM_118397.4	(Mayer et al., 1999)
CYP84	CYP84A4	NM_120515.3	(Weng et al., 2012)
CYP79	CYP79A118	KX931079.1	unpublished
CYP81	CYP81G1	NM_126131.4	(Tabata et al., 2000)
	CYP81D1	NM_123013.4	(Tabata et al., 2000)
	CYP81F2	NM_125104.3	(Tabata et al., 2000)
	CYP81E11	DQ340238.1	unpublished
	CYP81H1	NM_119895.4	(Mayer et al., 1999)
	CYP81K1	NM_121099.4	(Tabata et al., 2000)
	CYP81K2	NM_121098.3	(Tabata et al., 2000)
CYP89	CYP89A2	U61231.1	(Courtney et al., 1996)
	CYP89A3	NM_125525.1	(Tabata et al., 2000)

CYP93	CYP93E1	AF135485.1	(Steele et al., 1999)
	CYP93E2	DQ335790.1	(Li et al., 2007)
	CYP93E3	AB437320.1	(Seki et al., 2008)
	CYP93E4	KF906535.1	unpublished
CYP80	CYP80	U09610.1	(KRAUS and KUTCHAN, 1995)
CYP82	CYP82G1	NM_113423.4	(Mayer et al., 1999)
	CYP82A2	NM_001253148.1	(Schopfer and Ebel, 1998)
	CYP82C2	NM_119348.2	(Mayer et al., 1999)
	CYP82D47	JN604545.1	(Han et al., 2011)
CYP99	CYP99A3	Q0JF01.1	(Feng et al., 2002)
CYP712	CYP712A1	NM_129787.2	(Lin et al., 1999)
	CYP712A2	NM_147845.2	(Tabata et al., 2000)
CYP83	CYP83A1	KP693684.1	(Guo et al., 2016b)
	CYP83B1	KU559565.1	unpublished
	CYP83E8	DQ340234.1	unpublished
CYP705	CYP705A1	NM_117621.5	(Mayer et al., 1999)
	CYP705A5	NM_124173.3	(Tabata et al., 2000)

Supplementary Table 26. The UGT genes used to construct phylogenetic tree in this research.

Subfamily	Gene name	GenBank number	Reference
UGT79	GmUGT79B30	NM_001359019.1	(Shaokang Di, 2015)
	GmUGT79A6	NM_001288595.2	(Rojas Rodas et al., 2014)
	HvUGT13248	GU170355.1	(Wolfgang Schweiger, 2012)
	PhUGT79B31	LC387490.1	(Knoch et al., 2017)
UGT91	GmSGT2	NM_001317455.2	(Shibuya et al., 2010)
	GmSGT3	NM_001253928.2	(Shibuya et al., 2010)
	CaUGT91A1-like	XP_027076440.1	Bioproject: PRJNA506972
	ItUGT91A1-like	XP_031110113.1	Bioproject: PRJNA574454
	LsUGT91D1-like	XP_023735445.1	Bioproject: PRJNA432228
	HaUGT91D1-like	XP_022007978.1	Bioproject: PRJNA396063
	CcUGT91C1	XP_024970787.1	Bioproject: PRJNA453787
	ItUGT91C1	XP_031118563.1	Bioproject: PRJNA574454
UGT94	VpUGT94F1	AB514127.1	(Ono et al., 2010)
	SiUGT94-related	LC484019.1	unpublished
	SiUGT94-related-2	LC484018.1	unpublished
	PgUGT94Q2	JX898530.1	(Jung et al., 2014)
UGT89	NaUGT89A2-like	XM_019370589.1	(Chen and Li, 2017)
	AtUGT89B1	NM_106048.4	(Theologis et al., 2000)
	AtUGT89C1	Q9LNE6	(Yonekura-Sakakibara et al., 2007)
UGT90	HpUGT90A7	EU561019.1	(Witte et al., 2009)
	CtUGT90A14	MH013340.1	(Zhang et al., 2019)
UGT73	AtUGT73B1	NM_119576.4	(Lim et al., 2006)
	AtUGT73B2	AY339370.1	(Lim et al., 2006)
	AtUGT73B3	NM_119574.3	(Lim et al., 2006)
	AtUGT73B4	NM_001202600.1	(Mazel and Levine, 2002)
	AtUGT73B5	NM_127108.4	(Mazel and Levine, 2002)
	AtUGT73C1	NM_129230.3	(Hou et al., 2004)
	AtUGT73C2	NM_129231.3	(Lin et al., 1999)
	AtUGT73C5	NM_129235.4	(Hou et al., 2004)
UGT71	CtUGT71AE1	MH013341.1	(Zhang et al., 2019)
	AtUGT71D1	NM_128527.4	(Lin et al., 1999)
	AtUGT71C1	NM_128529.3	(Hansen et al., 2009; Lim

			et al., 2003)
	AtUGT71C2	NM_128528.4	(Hansen et al., 2009)
	AtUGT71C3	NM_100600.4	(Xie et al., 2012)
	AtUGT71B1	NM_113070.3	(Salanoubat et al., 2000)
	PgUGT71A27	AIZ00429.1	unpublished
	AtUGT71B6	NM_113073.3	(Priest et al., 2006)
	ITUGT71B2	MK704396.1	unpublished
	AtUGT71B5	NM_117616.2	(Mayer et al., 1999)
	RsUGT72B14	KX262844.1	(Yu et al., 2011)
	AtUGT72B1	NM_116337.3	(Brazier-Hicks and Edwards, 2005)
	AtUGT72B3	NM_001331274.1	(Lin et al., 2016)
UGT72	AlUGT72C1	NW_003302552.1	(Hu et al., 2011)
	AtUGT72E1	NM_114934.2	(Lim et al., 2005)
	AtUGT72E2	NM_126067.3	(Lanot et al., 2006; Lim et al., 2005)
	AtUGT72E3	NM_122532.3	(Lanot et al., 2008; Lim et al., 2005)
	AtUGT78D1	NM_102790.4	(Jones et al., 2003)
	AtUGT78D2	NM_121711.5	(Kim et al., 2012)
UGT78	AtUGT78D3	NM_121709.2	(Yonekura-Sakakibara et al., 2008)
	GmUGT78K1	GU434274.1	(Kovinich et al., 2010)
	AtUGT85A2	NM_102086.3	(Theologis et al., 2000)
	TcUGT85A2	EOX92065.1	(Motamayor et al., 2013)
UGT85	AtUGT85A3	NM_102088.3	(Theologis et al., 2000)
	AtUGT85A5	NM_202156.2	(Theologis et al., 2000)
	AtUGT85A7	NM_102085.3	(Theologis et al., 2000)
	AtUGT76C1	NM_120669.4	(Hou et al., 2004)
	TaUGT76C1	KY784575.1	unpublished
	SlUGT76E1	NM_001361347.1	(Sun et al., 2017)
	AtUGT76E1	NM_125350.3	(Tabata et al., 2000)
UGT76	AtUGT76E2	NM_125351.3	(Tabata et al., 2000)
	AtUGT76E11	NM_114534.3	(Li et al., 2018a; Salanoubat et al., 2000)
	AtUGT76E12	NM_114533.2	(Salanoubat et al., 2000)
	AtUGT76D1	NM_128205.3	(Lin et al., 1999)
	VrUGT87A2-like	XP_034687331.1	Bioproject: PRJNA636344
	VvUGT87A2	RVX23022.2	(Roach et al., 2018)
UGT87	CsUGT87A1-like	XP_028057899.1	Bioproject: PRJNA524157
	CaUGT87K1	AUR26629.1	unpublished
	CaUGT87K2	AUR26632.1	unpublished

UGT86	NtUGT86A1-like	XP_009610181.1	Bioproject: PRJNA257218
	CsUGT86A1-like	XP_028052942.1	Bioproject: PRJNA524157
UGT74	AtUGT74B1	NM_102256.3	(Theologis et al., 2000)
	AtUGT74F1	NM_129946.3	(Lin et al., 1999)
	AtUGT74F2	NM_129944.3	(Lin et al., 1999)
	PgUGTP74AE2	JX898529.1	(Jung et al., 2014)
	AtUGT74D1	NM_128733.5	(Tanaka et al., 2014)
	AtUGT74E2	NM_100448.4	(Theologis et al., 2000)
UGT75	SlUGT75C1	NM_001361345.1	(Aoki et al., 2010)
	AtUGT75B1	NM_100435.3	(Theologis et al., 2000)
	AtUGT75B2	NM_100432.2	(Theologis et al., 2000)
	AkUGT75W2	AWU66066.1	(Sun et al., 2018)
UGT84	AtUGT84A1	NM_117638.3	(Mayer et al., 1999)
	AtUGT84A2	NM_113051.3	(Salanoubat et al., 2000)
	AtUGT84B1	NM_127890.3	(Lin et al., 1999)
UGT709	CrUGT709C2	KF302068.1	(Miettinen et al., 2014)
	CaUGT709L1	AUR26631.1	unpublished
UGT95	PgUGT95B2	MH507175.1	(Wilson et al., 2019)
UGT708	CsUGT708C1-like	XP_028096648.1	Bioproject: PRJNA524157
	PpUGT708C1	XP_007216617.1	Bioproject: PRJNA241430
UGT80	LhUGT80A2	XM_031153760.1	unpublished
	AtUGT80B1	NM_001084205.2	(Theologis et al., 2000)
	AtUGT80	KJ396595.1	unpublished
	AtUGT80A2	NM_001337686.1	(Salanoubat et al., 2000)
P. ginseng	PgUGT1	KF377585.1	(Yan et al., 2014)
	PgUGT3	AIE12480.1	(Yan et al., 2014)
	PgUGT4	AIE12477.1	(Yan et al., 2014)
	PgUGT7	AIE12476.1	(Yan et al., 2014)
	PgUGT16	AIE12486.1	(Yan et al., 2014)
	PgUGT17	AKA44597.1	(Wang et al., 2015)
	PgUGT25	AKA44595.1	(Wang et al., 2015)
	PgUGT33	AKA44590.1	(Wang et al., 2015)
	PgUGT39	AKA44591.1	(Wang et al., 2015)
	PgUGT100	AKQ76388.1	(Wei et al., 2015)
	PgUGT101	KP795114.1	(Wei et al., 2015)
	PgUGT102	KP795115.1	(Wei et al., 2015)
	PgUGT103	KP795116.1	(Wei et al., 2015)
	PgUGTPg29	KM401911.1	(Wang et al., 2015)
	PgUGTPg45	KM401918.1	(Wang et al., 2015)
	PgUGT11	AIE12482.1	(Yan et al., 2014)

	PgUGT12	AIE12481.1	(Yan et al., 2014)
	PgUGTPg36	AKA44596.1	(Wang et al., 2015)
	PgUGTPg37	AKA44583.1	(Wang et al., 2015)
UGT83	VrUGT83A1-like	XP_034705962.1	Bioproject: PRJNA636344
	VvUGT83A1	RVW82717.1	(Roach et al., 2018)
	PtUGT83A1	XP_002306038.2	Bioproject: PRJNA17973
UGT82	VrUGT82A1	XP_034676882.1	Bioproject: PRJNA636344
	MrUGT82A1	KAB1210460.1	(Jia et al., 2019)
Others	Zj arabinosyltransferase RRA3-like	XP_015866879.1	Bioproject: PRJNA315994
	Cs arabinosyltransferase RRA3-like	XP_028093686.1	Bioproject: PRJNA524157
	To Nucleotide-diphospho-sugar transferase	PON84914.1	unpublished
	Ac Beta-1,4-xylosyltransferase	PSS36057.1	unpublished
	Cs Beta-1,4-xylosyltransferase IRX14	XP_028090508.1	Bioproject: PRJNA524157
	Cs UDP-glucosyltransferase	AYQ58374.1	unpublished
	Ac Zeatin O-glucosyltransferase	PSS15686.1	unpublished
	Ac Zeatin O-glucosyltransferase-2	PSS01783.1	unpublished
	Mc Glycosyl transferase	OVA05033.1	(Liu et al., 2017)
	Vr galacturonosyltransferase-like 3	XP_034701103.1	Bioproject: PRJNA636344
	Ac Beta-1,4-xylosyltransferase IRX9H	PSS01196.1	unpublished
	Jr galacturonosyltransferase 8-like	XP_018807381.1	Bioproject: PRJNA350852
	Cs galacturonosyltrans	XP_028115305.1	(Li et al., 2017)

	ferase 8-like		
	Ls beta-1,4-xylosyltransferase IRX10L	XP_023744067.1	Bioproject: PRJNA432228
	Cc xyloglucan 6-xylosyltransferase 2-like	XP_024996410.1	Bioproject: PRJNA453787
	Ls xyloglucan 6-xylosyltransferase 2-like	XP_023760243.1	Bioproject: PRJNA432228
	Vr UDP-rhamnose:rhamnosyltransferase 1	XP_034708582.1	Bioproject: PRJNA636344
	Vr UDP-rhamnose:rhamnosyltransferase 2	XP_034708677.1	Bioproject: PRJNA636344

1272

1273

1274 **Supplementary Table 27.** Primers for cloning UGT genes in *P. notoginseng* genome.
1275

Genes	Primers	Sequence (5' to 3')
PnUGT1	Seq1091.2-F	ATGAAGTCAGAATTGATATTCTTGC
	Seq1091.2-R	TTACATAATTTCCTCAAATAGCTTC
PnUGT2	Seq1751.3-F	ATGGATAACCAAAAAGGTAGAATCA
	Seq1751.3-R	CTATTGTGCATCTTTCTTCTTCTTA
PnUGT3	Seq2311.15-F	ATGAAGTCAGAATTGATATTCGTGC
	Seq2311.15-R	TCACATAATTTCCTCAAATAGTTTC
PnUGT4	Seq3403.8-F	ATGGATATCGAAAAAGGTAGAATCA
	Seq3403.8-R	TTAATATTGTGCGTCTTTCTTCATC
PnUGT5	Seq574.8-F	ATGTTGAGCAAAACTCACATTATGT
	Seq574.8-R	TCAGGAGGACACAAGCTTTGAAATG
PnUGT6	Seq1424.9-F	ATGGTTTCTATTCGGAGAACATTGT
	Seq1424.9-R	TCAAAAAATTGTGGTATGAGGAACA
PnUGT7	Seq1517.4-F	ATGCTGGAGCAGTGTTTGGGACAAC
	Seq1517.4-R	TTATACCTTGACGGCTTTAAATGCA
PnUGT8	Seq1543.17-F	ATGGCAGGTCGTAGTAGAGACGGTC
	Seq1543.17-R	TTACTGTTCTGAACCATCAGGGAAG
PnUGT9	Seq1607.6-F	ATGGACTCACAAGTCTCATCACGTC
	Seq1607.6-R	TTACTGATCTGATCGTTCCTCTCTC
PnUGT10	Seq1625.69-F	ATGAGGAACTGGAGTTGGGGTTTTG
	Seq1625.69-R	CTACCATGGTTTGAGGTCTCCCATG
PnUGT11	Seq1743.56-F	ATGGATACGACAAGGCGGAAGGCGG
	Seq1743.56-R	TCAAAAACAATACTGAATTAACTTT
PnUGT12	Seq1790.11-F	ATGGATGGCAAGAGCCTTCACATAG
	Seq1790.11-R	CTAGGAGGCTACGAGAAGGTCTTGC
PnUGT13	Seq1935.39-F	ATGAAGAAGCTGAAGAGCTTTTACA
	Seq1935.39-R	CTATTTGCACTGCATTGGTCGGAAC
PnUGT14	Seq1965.8-F	ATGGAGTCTCCGAATAGACCTCATG
	Seq1965.8-R	TTAAGGTTTGCTAATATTTTTTCCA
PnUGT15	Seq1975.16-F	ATGATCCCCCTCTCCGAAATCGCCC
	Seq1975.16-R	TTATGCTTTCTCCCTTTTCTCTCTG
PnUGT16	Seq2001.6-F	ATGGGTCAGCTTAATGTGTTCTTTT
	Seq2001.6-R	TCAAGAATGATTAGAACTCAATTCT
PnUGT17	Seq2096.16-F	ATGGCTATTCTCCAAACCCAAGACC
	Seq2096.16-R	TCATTTCAATTTCAGTTGTTCCACG
PnUGT18	Seq2308.20-F	ATGAAGCTCTCTGCGCTGCAGCAGA
	Seq2308.20-R	CTACTTCTTACTGGTATGGCTTGCA
PnUGT19	Seq312.4-F	ATGGCGAACACGACGACGTTTCGAA
	Seq312.4-R	TTAGAGACCAAAATTGCAGGCCTGG
PnUGT20	Seq3221.1-F	ATGCCACCAAAACTCCACCTCCCAA
	Seq3221.1-R	TCAGCTGTCAGAATACAAATATTCA
PnUGT21	Seq3651.4-F	ATGAAGAACTCAGAATTGGTATTCG

	Seq3651.4-R	TCACATGATCTCCTCAATTAGTTTC
PnUGT22	Seq3959.10-F	ATGAAGCTCTCTGCGCTACAGCAGA
	Seq3959.10-R	CTACTTCTCACTGGTATGGCTTGCA
PnUGT23	Seq4354.2-F	ATGGAAAATAACCACGTTCTTCATG
	Seq4354.2-R	TTAACTCATCAATTGGGATTTCCTC
PnUGT24	Seq4407.1-F	ATGGCTCAACAAACAATCCCACCTC
	Seq4407.1-R	CTAGGGTGTGATGCCACCCAAAGTC
PnUGT25	Seq4424.12-F	ATGCCAACACAGAAATACTCAACCC
	Seq4424.12-R	TTATTGTTTAGATTTCACACCCATT
PnUGT26	Seq4424.13-F	ATGGAGAAAAAGGACTCAACTCGAC
	Seq4424.13-R	TCATCTCTCCACACCCATCAATTTA
PnUGT27	Seq4424.15-F	ATGGCTGAACAAACAATCCCACCTC
	Seq4424.15-R	CTAGGCTCTGATGCCACCCACAGTC
PnUGT28	Seq4424.16-F	ATGGATCAACCAGCAGCCGAACCTC
	Seq4424.16-R	TTAGCTACGCAAAACTACAGCCATC
PnUGT29	Seq4481.73-F	ATGGCAACTGAAGACCCTAAACTCC
	Seq4481.73-R	TTATCCATTTTTTGATTTCTCAAAA
PnUGT30	Seq4702.22-F	ATGGAGATTAACCGGCATAGGAAGC
	Seq4702.22-R	TTATTTTGTATGATTTTCAAGATAC
PnUGT31	Seq5047.49-F	ATGGGCTCCCTTCCTAAAGTAACTA
	Seq5047.49-R	CTACTTTGCTAACAACACCTGATCC
PnUGT32	Seq5124.21-F	ATGGTGGGTCGTAAAGAGAAGAGCA
	Seq5124.21-R	TTATTGCGTATTTGTTTGCCAGTCA

1276