Effective prediction of biosynthetic pathway genes involved in bioactive polyphyllins in Paris polyphylla

Xin Hua[1,&], Wei Song[2 , &], Kangzong Wang[1,&], Xue Yin[1,] , Changqi Hao[1] , Baozhong Duan[6], Zhichao Xu[3]*, Tongbing Su[4,5]*, Zheyong Xue[1]*

[1] Key Laboratory of Saline-alkali Vegetation Ecology Restoration (Northeast Forestry University) , Ministry of Education, Harbin, China

[2] College of Pharmacy, Zhejiang Chinese Medical University, Hangzhou, China

[3] Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[4] Beijing Vegetable Research Center (BVRC), Beijing Academy of Agriculture and Forestry Science (BAAFS), Beijing, China

[5] National Engineering Research Center for Vegetables, Beijing 100097, China

[6] College of Pharmaceutical Science, Dali University, Dali, China

[&] X.H., W.S. and K.W. contributed equally.

* corresponding author: Zheyong Xue (zyxue@nefu.edu.cn), Tongbing Su (sutongbing@nercv.org) or Zhichao Xu (zcxu@implad.ac.cn).

ABSTRACT

The genes in polyphyllins pathway mixed with other steroid biosynthetic genes form an extremely complex biosynthetic network in *Paris polyphylla* with a giant genome. The lack of genomic data and tissue specificity causes the study of the biosynthetic pathway notably difficult. Here, we report an effective method for the prediction of key genes of polyphyllin biosynthesis. Full-length transcriptome from eight different organs via hybrid sequencing of next generation sequencingand third generation sequencing platforms annotated two 2,3-oxidosqualene cyclases (OSCs), 216 cytochrome P450s (CYPs), and 199 UDP glycosyltransferases (UGTs). Combining metabolic differences, gene-weighted co-expression network analysis, and phylogenetic trees, the candidate ranges of *OSC*, *CYP*, and *UGT* genes were further narrowed down to 2, 15, and 24, respectively. Beside the three previously characterized CYPs, we identified the OSC involved in the synthesis of cycloartenol and the UGT (PpUGT73CR1) at the C-3 position of diosgenin and pennogenin in *P. polyphylla*. This study provides a idea for the investigation of gene cluster deficiency biosynthesis pathways in medicinal plants.

Keywords: 2,3-oxidosqualene cyclase; Metabolic biosynthetic pathways; *Paris polyphylla;* Steroid saponins; UGT glucosyltransferase.

INTRODUCTION

P. polyphylla var. *yunnanensis* is a member of Liliaceae family and one of the most famous medicinal plants in China. The rhizome of this plant is an important component of the traditional Chinese medicines “Yunnan Baiyao” and “Gongxue Ning,”[1] which have pharmacological activities, such as hemostasis, analgesic, sedation, anti-inflammatory, and anti-tumor effects[2-4]. The main active components of this plant are steroidal saponins, also known as polyphyllins, accounting for about 80% of the total number of active compounds[5]. Polyphyllins have aroused great interest for their rich pharmacological activities, including anti-inflammatory, vascular protection, hypoglycemic, immunomodulatory, antiparasitic, hypocholesterolemic, antifungal, anti-parasitic, and anti-tumor effects[4,6,7]. However, the species is at risk of extinction due to its slow growth and excessive exploitation[8]. Given their complex molecular structures, polyphyllins are unlikely to be chemically synthesized for commercial usages. Therefore, metabolic engineering may be an effective method to provide a stable source of polyphyllins. The metabolic engineering strategy largely relies on the biosynthetic pathway of polyphyllins, which still has not been fully elucidated.

Polyphyllins are a group of products with different sugar chains connected at the C-3 or C-26 position of diosgenin or pennogenin. Diosgenin is also an important precursor to the synthesis of over 200 steroidal drugs (e.g., contraceptives, testosterone, progesterone, and glucocorticoids) [8]. However, the sources of diosgenin mainly depend on the extraction from several specific plants, such as yam (*Dioscorea* genus) and fenugreek (*Trigonella foenum-graecum*). The biosynthetic pathway of polyphyllins begins with the condensation of two molecules of isopentenyl diphosphate and one molecule of dimethylallyl diphosphate, which is then catalyzed by farnesyl diphosphate synthase (FPS) to form farnesyl diphosphate (FPP, C15) [9]. Two FPP molecules are catalyzed by squalene synthase (SQS) to produce a linear C30 molecule, squalene, which is further cycled by squalene epoxidase to 2,3-oxidosqualene [9]. Then, 2,3-oxidosqualene is cyclized by a cycloartenol synthase (CAS) to form cycloartenol, which is then modified through a series of oxidation and reduction to form cholesterol[10,11]. Enzymes in the CYP90G family can catalyze the hydroxylation of cholesterol C-16 and C-22 with the closure of E ring. Then, *16S*,*22S*-dihydroxycholesterol is further hydroxylated at C-26 and forms an F ring to produce diosgenin under the action of cytochrome P450s (CYPs), such as PpCYP94D108[12]. However, how steroidal skeleton α-hydroxylates at C-17 form pennogenin is still unknown. Subsequently, diosgenin and pennogenin are glycosylated by UDP glycosyltransferases (UGTs) to form various polyphyllins (Figure 1). To date, the UGTs related to polyphyllin biosynthesis have still not been selected and functionally identified.

Although genomic and transcriptional information of numerous medicine plants have been generated and made available to public, the progress of candidate gene mining and whole pathway dissection of specialized plant metabolites remains slow due to the following factors[6,12-15]. The biosynthesis of several specified metabolites, such as the ginsenosides in Araliaceae family, lacks tissue specificity. A variety of ginsenosides are widely found in multiple tissue parts of the plant, and the complex distribution pattern of ginsenoside components hinders the prediction of the exact genes involved in biosynthetic pathway application by simple differentially expression analysis[16]. The divergent evolution of CYP and UGT families generated numerous individual members that are phylogenetically close and can decorate diverse type of natural products. For this reason, predicting the related pathway gene solely based on phylogenetic analysis is impossible [17]. A previous study reported the extremely huge genome size of Parideae species (about 50 pg) (Pellicer et al., 2014). In addition, the biosynthetic genes of specific plant metabolites are scattered in different regions of the genome, further increasing the difficulty in identifying candidates precisely by physical distance of metabolic pathway-related genes[18,19], despite the generation and good assembly of whole-genome sequences[20,21]. Therefore, an efficient strategy needs to be developed and improved urgently to accurately predict the key genes in the complex none-clustered biosynthetic pathway of specialized plant metabolites.

In this study, full-length transcriptome analysis using hybrid sequencing strategy based on single-molecule sequencing and paired-end mRNA sequencing was performed on eight different tissues from *P. polyphylla* var. *yunnanensis*. The weighted gene co-expression network analysis (WGCNA) combining the distributions of specific metabolites, different gene expressions, and phylogenetic analysis was further used to predict the key genes involved in the biosynthesis of polyphyllins. Then, candidate genes containing several assumed 2,3-oxidosqualene cyclase (OSC) genes and UGTs were functionally verified. This study may provide a strong basis for characterizing the steps of biosynthesis of polyphyllins of *P. polyphylla* var. *yunnanensis,* thus promoting the production of such important chemicals via synthetic biology.

METHODS

Plant material and RNA preparation

Samples of four-year dwarf *P. polyphylla* var. *yunnanensis* were collected from Dali, Yunnan, China. The rhizomes, fibrous roots, stems, leaves, ripe fruits, stigma, petals, and pistil were harvested in 5-year-old healthy plants (Figure S1). The rhizomes, fibrous roots, stems, leaves, and ripe fruits were harvested in October 2018, whereas the stigma, petals, and pistil were harvested in April 2019. All tissues were frozen in liquid nitrogen immediately and stored at −80 °C after collection. Every sample had three biological replicates that were sequenced independently.

RNA isolation, transcriptome sequencing, and gene function annotation

The sequencing samples of rhizomes, fibrous roots, stems, leaves, ripe fruits, stigma, petals, and pistil were from multiple plants, and the full-length transcriptome sequencing samples were from a mixture of different tissues from multiple plants. Total RNA was isolated using an RNA Plus kit (Takara, Qingdao, China), in accordance with the manufacturer's protocol. Three biological replicates of rhizome, fibrous root, stem, leaf, and ripe fruit were determined, and two biological replicates of stigma, anther, and petal were tested due to difficulty in sample collection and insufficient RNA quality. RNA quality was examined using Agilent 2100 (Agilent Technologies, Santa Clara, USA). The cDNA library was constructed and sequenced by Biomarker Technologies Corporation (Beijing, China). A single tissue was sequenced using an Illumina Hiseq 2000 platform, and full-length transcriptome sequencing for the mixture of different tissues was performed using PacBio Sequel platform. The PacBio long reads were filtered, and redundant sequences were removed using CD-HIT-EST program. In full-length transcriptome data, sequences with polymerase read less than 50 bp and sequence accuracy less than 0.9 were filtered out. Then, the clean reads from Illumina sequencing were mapped into the non-redundant long-reads to calculate the Fragments Per Kilobase per Million (FPKM) values using DESeq2. Specifically, FPKM = cDNA fragments / mapped fragments (millions) × transcript length (kb), where cDNA fragments represent the number of fragments aligned to a transcript, mapped fragments (millions) is the total number of fragments aligned to the transcript; transcript length (kb) denotes the transcript length. The filtered full-length transcripts were functionally annotated using non-redundant (nr), SWISSPROT, Gene Ontology, Clusters of Orthologous Genes, EuKaryotic Orthologous Groups (KOG), PFAM, and Kyoto Encyclopedia of Genes and Genomes databases, respectively.

Phylogenetic Analysis

Transcripts belonging to OSCs, CYPs, and UGTs were identified using BLAST software. The transcripts with a length under 1000 bp were removed. Table S1 lists the OSC sequences from different plants used to construct the phylogenetic tree. The phylogenetic trees of OSCs, P450s, and UGTs were constructed using the maximum likelihood method and Jones–Taylor–Thornton model using MEGA 7.0 [22]. A bootstrap resampling analysis with 1000 replicates was performed to evaluate the topology of phylogeny.

Metabolite content determination

The tissues frozen at -80 °C were lyophilized in a freeze dryer. Then, 20 mg dry materials from different tissues were weighted and placed in a 2 mL centrifuge tube. A total of 1 mL 80% methanol containing 20 μg internal standard was added to the sample (digitoxin, ≥95%, Sigma). The samples were further extracted at 1400 rpm for 2 h and centrifuged at 10,000 g for 5 min. The supernatant was transferred to a new centrifuge tube and added with 300 μL n-hexane for extraction. After extraction, the n-hexane layer was sucked and removed, and the process above was repeated. SpinVac was applied for solvent removal in the samples. The samples were redissolved with 500 ml distilled water. Then, extraction was performed twice with 500 mL n-butanol. Nitrogen was used to dry the organic phase of the sample and redissolved sample with the mobile phase during measurement.

The analysis was performed on a Waters ACQUITY ultra-performance liquid chromatography (LC) system coupled with an AB Sciex 5500 Qtrap mass spectrometer (AB Sciex, Milford, MA, USA). Chromatographic separation was achieved on a ACQUITY BEH C18 column (100 × 2.1 mm^2, 1.7 μm) at 40 °C. The 0.1% formic acid water was used as mobile phase A, and 0.1% formic acid in acetonitrile was used as mobile phase B. The gradient was 0–1 min, 5%–52% B; 1–6 min, 52%–56% B; 6–7 min, 56%–95% B; 7–7.5 min, 95%–95% B; 7.5–9 min, 95%–5% B; 9–10 min, 5%–5% B. The flow rate was 0.25 mL min^{-1}, and the injection volume was 5 μL. The electrospray ionization source interface operated in negative ionization mode was used in this study. The ion spray voltage was set at −4500 V. Table S2 shows the optimized multiple reaction monitoring parameters for the analytes and internal standard. Cholesterol detection was completed on a Thermo ISQ-LT gas chromatography–mass spectrometry (GC-MS) system using Thermo TG-5HT column (30 m × 0.25 mm × 0.10 μm). The mass detector was set to SCAN mode, the scanning range was 60–800 m/z, and solvent delay was 10 min. The GC conditions were as follows. The sample (1 μL) was injected in split mode (10:1) at 250 °C under a He flow rate of 1.2 mL $min^{−1}$, and the temperature cycle involved the initial injection temperature of 170 □ for 2 min, 170 □ to 290 □ for 6 □ per minute, holding for 4 min after the temperature reached 290 □, and raising the temperature from 290 □ to 340 □ at 25 □ per minute.

Construction of gene co-expression networks

Gene co-expression networks were constructed using the WGCNA approach with R packages (version 3.2.2)[23]. Here, we used the normalized quantile function in the R software package to normalize the gene expression data. We selected the expression matrix of 31,937 genes with the sum FPKM value in all tissues greater than 1.0 from all genes as the input file for WGCNA to identify gene modules with strong co-expression. Before the construction of the network module, outlier samples should be removed to ensure the accuracy of the results because the analysis results of network module are easily affected by outlier samples. By calculating the correlation coefficient of each sample’s expression level and clustering, the samples with low correlation or those that cannot be clustered on the tree graph are removed. Next, WGCNA network construction and module detection were conducted using an unsigned type of topological overlap matrix (TOM). Based on the TOM, we used the average-linkage hierarchical clustering method to cluster genes, following the standard of hybrid dynamic shearing tree and set the minimum number of genes for each gene network module to 30. The power β was selected based on the scale-free topology criterion. The modules were detected as branches of the dendrogram using the dynamic tree-cut, and a cut-off height of 0.25 was used to merge the branches to final modules.

Finally, the gene visual network was described by using heatmap. The heatmap depicts the TOM among all genes in the analysis. Light color represents a low overlap, and progressively darker red color represents higher overlap. Blocks of darker colors along the diagonal are the modules, and a very strong association existed between the genes that are contained within these red modules. These red modules were the focus of our genetic prediction.

Functional verification of OSC genes

The function of OSC gene was verified by yeast strain and *Nicotiana benthamiana*. Table S3 shows all the selected strains and plasmids used in the yeast experiment. *PpOSC1* and *PpOSC2* were cloned from *P. polyphylla* var. *yunnanensis* and transferred in to pδHis plasmid. The plasmid was transformed into yeast strain BY-SQ1 using the standard lithium acetate approach. The yeast strains SQ-PpOSC1 and SQ-PpOSC2 were precultured in 5 ml synthetic defined medium with glucose as carbon source and uracil and histidine omitted (SD-URA-HIS) at 30 °C and 220 rpm for 24 h. Precultures were inoculated at an initial optical density $(OD)_{600}$ of 0.05 in 50 ml SD-URA-HIS in 250 ml flasks and grown under the same condition for 72 h. The cells were harvested, resuspended in 2 ml 10% KOH (w/v) and 90% ethanol (v/v), heated for 2 h at 75 °C, and cooled and extracted once with 0.5 ml ethyl acetate. After centrifugation, the ethyl acetate phase was collected and dried by centrifugal vacuum evaporator. Derivatization of the dried products was conducted with 1-(trimethylsilyl)imidazole-pyridine mixture at 70 °C for 30 min to prepare the sample for analysis.

In the transient expression system of *Nicotiana benthamiana,* the coding regions of candidate OSC genes were cloned from *P. polyphylla* var. *yunnanensis* into the pEAQ-HT-DEST1 vector. After sequence verification, pEAQ-HT-DEST1 vectors carrying OSC genes were separately transferred into *Agrobacterium tumefaciens* strain GV3101 and cultured overnight at 28 °C and 220 rpm. Then, 1 mL culture was used to inoculate 10 mL Luria–Bertani (LB) medium containing 50 µg/mL kanamycin, 25 µg/mL rifampicin, and 25 µg/mL gentamicin for overnight growth. The following day, the cultures were centrifuged (5000 g, 5 min), and cells were resuspended in infiltration buffer (10 mM MES ($C_6H_{13}NO_4S$), pH 5.6, 10 mM $MgCl_2$, and 100 µM acetosyringone) to a final OD_{600} of 0.4. The leaves of 6-week-old *N. benthamiana* were infiltrated with *A. tumefaciens* solution as follows. A 5 mL needle-free syringe was used to gently push the bacterial mixture into the abaxial surface until the entire leaf was filled with agrobacterium. The infiltrated leaves were cultured at 22 °C, exposed to light for 10 h a day, and harvested at 6th day after infiltration. For metabolite extraction, leaf discs in diameter 1cm were prepared from *Agrobacterium*-infiltrated *N. benthamiana* and dried with a vacuum freeze-dryer. Then, the leaves were ground into powder, and 10 mg powder was weighed and placed a 2 mL tube for use. Then, 2 mL lysate was added to each sample and heated in the water bath (75 °C for 1 h). After the samples were completely dried, 300 µL ethyl acetate and 500 µL water were added and mixed with vortex shock and centrifuged for 10 min to facilitate separation. Next, 100 µL was removed from the upper layer (ethyl acetate layer) and transferred to a special glass tube. The liquid was blow dried with nitrogen and added with 50 µL 1-(trimethylsilyl) imidazole-pyridine mixture. After vortex-mixing–heating at 70 °C for 30 min, the mixture was analyzed with GC-MS same as cholesterol analysis above.

Cloning and prokaryotic expression of *UGT* genes from *P. polyphylla* var. *yunnanensis*

The total RNA from the *P. polyphylla* var. *yunnanensis* was extracted and reverse transcribed to obtain cDNA. The candidate polymerase chain reaction (PCR) primers for UGTs were designed based on the transcriptome sequence. The PCR procedure was as follows: 95 °C for 3 min; 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 90 s in 33 cycles; 72 °C for 5 min. The primers of UGT genes are shown in Table S4. The prokaryotic expression vector pGEX-6p-1 was linearized with restriction endonucleases EcoR I and Sal I (Thermo), recombined with the PCR product through the ClonExpress II One Step Cloning Kit (Vazyme), and transformed into *E. coli* DH5α.

The plasmid with correct sequencing was transformed into Rosetta-gami B (DE3) pLysS and inoculated into LB liquid medium containing ampicillin (100 mg/L) and then cultured at 37 °C at 180 rpm until the OD_{600} of = 0.6. A total of 0.2 mM isopropyl β-d-1-thiogalactopyranoside was added to the culture medium, induced at 16 °C for 16 h, and centrifuged at 4 °C at 5,000 rpm to collect the bacteria. The bacterial cells were suspended in 10 mM phosphate buffer (pH 7.4), and the cells were disrupted by ultrasound in an ice bath. Then, the cells were centrifuged at 12,000 g at 4 °C for 20 min. The bacterial supernatant was purified using glutathione beads (Smart-Life Sciences, Changzhou, China) and concentrated using Millipore ultrafiltration tubes (Meck, Darmstadt, Germany). Pierce BCA Protein Assay Kit (Thermo, Waltham, USA) was used to quantify the target protein.

Enzyme activity analysis

The enzymatic reaction system consisted of 50 mM Tris (pH 8.0), 1 mM $MgCl_2$, 5 mM glucose donor (UDP-glucose), 1 mM glucose receptor (diosgenin/pennogenin), and purified enzyme of PpUGT73CR1 in a final volume of 100 μL. After overnight incubation at 37 °C, an equal volume of ice methanol was added to stop the reaction. The product was concentrated and dried, dissolved in 100 μL chromatographic methanol, and centrifuged at 12,000 rpm for 10 min, and the supernatant was obtained for testing. The reaction products were identified by high-performance LC (HPLC) and LC time-of-flight mass spectrometry (LC-TOF-MS), and the Thermo Hypersil GOLD C18 column (250 mm × 4.6 mm, 5 μm) was used for HPLC detection. The mobile phases were water (A) and acetonitrile (B). The elution gradient was as follows: 0–6 min, 20%–30% B; 6–15 min, 30%–60% B; 15–21 min, 60%–100% B; 21–30 min, 100% B, 30–35 min, 100%–20% B. The flow rate was 1 mL/min, the column temperature was 30 °C, the injection volume was 10 μL, and the detection wavelength was 210 nm. LC-TOF-MS was performed using the AB Sciex Tripletof 6600 (AB Sciex, Milford, MA, USA) in a positive ionization mode.

For the kinetic analysis of UGT73CR1, the reaction mixture contained 50 mM Tris−HCl (pH 8.0), 5 mM UDP-glucose, acceptor substrate (20–400 μM diosgenin and pennogenin), and 1 μg purified UGT73CR1 in a final volume of 100 μL. The reaction was incubated at 37 °C for 30 min. HPLC analysis was used to quantify the target product in each reaction. The Michaelis–Menten parameters were calculated by kinetic model using Prism 7 (GraphPad, San Diego, CA, USA). All data are presented as means ± standard deviation of three independent experiments.

RESULTS

Transcriptome sequencing, assembly, and functional annotation

A total of 292.69 Gb clean data for 21 sequencing libraries, including three biological replicates of rhizome, fibrous root, stem, leaf, and ripe fruit and two biological replicates of stigma, anther, and petals, were obtained by Illumina sequencing (Table S5). A total of 81.81 Gb clean data containing 1,121,119 CCS reads were obtained from PacBio sequencing platform. Among them, 969,450 long reads belong to the full length non-chimeric sequence. The mean read length of CCS was 2,263 bp. The full-length non-chimeric sequences were clustered into 69,009 consensus sequences, and the consensus sequences were polished using Quiver to obtain 68,266 high-quality consensus sequences. The low-quality consensus sequences were further corrected using the Illumina short reads. After removing redundant sequences for the high-quality consensus sequences and corrected low-quality consensus sequences, 39,875 transcript sequences were finally obtained. Using BUSCO [24] to evaluate the integrity of the transcriptome, the results showed that complete and single-copy duplicated transcript sequences accounted for 69.92%, the fragmented ones accounted for 5.97%, and those missing accounted for 26.11%.

First, we compared the known data in the nr database, annotating 38,177 (95.74%) unigenes from 39,875 transcripts. Afterward, we performed comparisons in the Swissprot, eggNOG, KOG, and Pfam databases and annotated 29,475 (73.92%), 37,673 (94.48%), 24,581 (61.65%), and 33,607 (84.28%) unigenes.

Phylogenetic analysis of OSC, CYP, and UGT gene families

In this study, we identified 2 intact OSCs, 216 CYPs, and 199 UGTs using PFAM annotation and BLAST algorithm. The phylogenetic tree of two OSCs from *P. polyphylla* var. *yunnanensis* and other 51 species showed that different branches distinguished the classification of OSC genes. The different OSC subfamilies are distributed in terms of various activities, especially for the skeletons of catalytic products, including cycloartenol, β-amyrin, lanosterol, lupeol, α-amyrin, friedelin, dammarenediol II, and mixed products (Figure 2a). Two identified OSC transcripts from *P. polyphylla* var. *yunnanensis* were classified into the CAS clade.

Phylogenetic analysis of CYPs using *Arabidopsis* as a reference indicated that all the full-length P450s from *P. polyphylla* var. *yunnanensis* can be assigned to CYP51, CYP71, CYP710, CYP711, CYP72, CYP74, CYP85, CYP86, and CYP97 family. Among them, the CYP71 (85) and CYP86 (60) families have the largest number of P450 genes, followed by the CYP85 (37) and CYP72 (21) families, whereas CYP710 (4), CYP97(4), CYP74 (2), CYP711 (2) and CYP51 (1) has the smallest number of P450 genes (Figures 3a and S2).

The phylogenetic analysis of *P. polyphylla* var. *yunnanensis* UGTs was carried out together with the UGTs from *A. thaliana* and *Z. mays*, and the predicted UGT protein sequences of *P. polyphylla* var. *yunnanensis* were clustered into 13 of the 21 known UGT subfamilies[25]. The H, K, M, or N subfamilies of UGTs are lost in *P. polyphylla* var. *yunnanensis*. Among all the subfamilies, D is the largest phylogenetic group in *P. polyphylla* var. *yunnanensis*, containing 57 genes accounting for 28.64% of all UGTs (Figures 3a and S3).

Metabolite and gene co-expression analysis to predict functional genes.

According to the results in Figure S4, a soft threshold β = 9 was selected to build a co-expression network. Then, the function hclust was used to perform hierarchical clustering on dissimilar matrices, whereas Dynamic Tree Cut was utilized to cut the generated cluster tree (Figure S5). In this process, unigenes with similar expression patterns can be combined on the same branch, and each branch represented a co-expression module, with different colors representing various modules. Differential unigenes were correlated and clustered based on their FPKM values. Unigenes with a high correlation were assigned to the same module (Figure S6). In the end, 31,937 unigenes were divided into 26 modules, and the number of unigenes in the modules was 53–7667.

Using LC-MS-MS, we determined the production profiles of eight metabolites (cholesterol, diosgenin, trillin, prosapogenin A, polyphyllin I, polyphyllin II, polyphyllin VI, and polyphyllin VII) in different tissues of *P. polyphylla* var. *yunnanensis* (Figure 3c). Diosgenin and most of its related metabolites (prosapogenin A, polyphyllin I, and polyphyllin II) had relatively similar distribution patterns and were highly accumulated in rhizomes, leaf, ovary, and petal. The distribution of trillin was exceptionally higher in leaf than in other tissues. The accumulation of pennogenin-derived saponins in different tissues of *P. polyphylla* also showed similar patterns. Polyphyllin VI was highly accumulated in rhizomes, fibril, fruit, and ovary, whereas polyphyllin VII was highly accumulated in fibril, fruit, ovary, and petal. The distribution patterns of diverse polyphyllins and transcriptome data were integrated to construct a co-expression network of metaboloic pathway genes and metabolites (Figure S7). The network can combine gene modules and metabolites with similar patterns in various tissues of *P. polyphylla* var. *yunnanensis* and calculate the correlation coefficient between metabolites and gene modules (Figure 3b). When the correlation coefficient approaches 1, the metabolites and genes in the module show similar expression or distribution patterns in different tissues of *P. polyphylla* var. *yunnanensis*.

Based on the correlation coefficient, we observed that trillin clustered well with the lavenderblush3 (0.82), lightblue3 (0.72), and white (0.73) modules in the co-expression network. Polyphyllin I had the highest correlation with Coral (0.72) and orangered4 (0.7) modules, and polyphyllin VII showed a high correlation with antiquewhite4 (0.81) and antiquewhite1 (0.76). Diosgenin and cholesterol were correlated with the antiquewhite4 module, and the correlation coefficients were 0.66 and 0.65, respectively. Polyphyllin II was correlated with lavenderblush3 (0.64) module, but polyphyllin VI and prosapogenin A showed no specific correlation with a certain module. The clustering results indicated that the genes in the orangered4, lavenderblush3, lightblue3, coral, antiquewhite1, white, and antiquewhite4 modules are likely to be involved in the biosynthesis of polyphyllins. Furthermore, we analyzed the upstream genes involved in the metabolic pathways of cholesterol or other phytosterols in the above seven modules. The coral, antiquewhite4, and lavenderblush3 modules contained relatively rich sterol synthesis upstream genes (14, 21, and 11, respectively) (Table S6). However, the orange, lightblue3, white, and antiquewhite1 modules contained less sterol synthesis upstream genes (0, 0, 4, and 6, respectively). Based on the results, the genes contained in the coral, antiquewhite4, and lavenderblush3 modules are likely to play a key role in the biosynthesis of polyphyllins, and subsequent gene prediction will be developed around these three modules.

From the three predicted modules, we obtained 42 CYPs and 48 UGTs. Heat maps were plotted to show the expression patterns of these genes in different tissues (Figure 4a). Genes involved in specialized metabolites biosynthesis often show high expression levels in certain tissues. Therefore, among the genes that have been predicted in our correlation analysis, the *CYP* and *UGT* genes with high expression levels in polyphyllins that accumulated tissues are likely to participate in the biosynthesis of polyphyllins.

In addition, we predicted the candidate *CYP* and *UGT* genes using phylogenetic analysis. P450s that can ~~be~~ hydroxylate ~~in~~ triterpene or steroidal skeletons often belong to the CYP72, CYP90, and CYP94 families. Based on our phylogenetic tree, we detected 21, 11, and 28 genes from the CYP72, CYP90, and CYP94 families, respectively. The UGTs that can add a glycosyl group at the C-3 position of triterpenoid and steroidal aglycone belong to the UGT73 family. A total of 57 UGT73s were annotated from our transcriptome data. Through combinational analysis of phylogenetic tree and WGCNA, we narrowed the *CYPs* and *UGTs* to 15 and 24 candidate genes, respectively (Figure 4b). Among them, three CYP genes, namely, *PpCYP90G4* (F01_transcript/40556), *PpCYP90B27* (F01_transcript/40246), and *PpCYP72A616* (F01_transcript/40246), have been reported to be involved in the biosynthesis of polyphyllins. These three genes ranked the 1st, 3rd, and 6th place in the list of candidate genes, respectively. These candidate genes are very likely to be involved in the biosynthesis of polyphyllins. The genes that appeared in the candidate clades of phylogenetic tree but were not included in the candidate modules of WGCNA should not be ignored. They may also ~~include some genes~~ be involved in the biosynthesis of polyphyllins. All candidate *OSC*, *CYP*, and *UGT* genes can be found in Supplementary File 2.

PpOSC1 but not PpOSC2 catalyzes the conversion of 2,3-oxidesqualene to cyclic triterpenes

OSC is one of the key enzymes in steroid biosynthesis [12,13], and the OSC-catalyzed conversion of 2,3-oxidesqualene to cyclic triterpenes marks the first scaffold diversification reaction in triterpenoid and steroid pathways. In contrast to single-copy *OSC* gene in lower plants, higher plants always have multiple *OSC* gens in their genomes [86]. A total of 13 OSC-related transcriptome variants were found in the transcriptome, which may represent two non-redundant *OSCs* (*PpOSC1* and *PpOSC2*). To investigate the activity of the two candidates, we cloned two genes transferred them into optimized yeast strains (SQSQ-pPOSC1 and SQSQ-pPOSC2). Cycloartenol production was evidently observed in the SQ-PPOSC1 strains, whereas it was absent in SQ-PPOSC2 strains, suggesting that *PpOSC1* gene encodes CAS in *P. polyphylla* var. *yunnanensis* (Figure 2b). Furthermore, *PpOSC1* and *PpOSC2* genes were transfected into *A. tumefaciens* for infected *N. benthamiana* leaf infiltration. Similarly, after infiltration, we collected the leaves and verified the function of *OSC* genes by measuring the content of triterpenes. The results showed that the instantaneous expression of *PpOSC1* can facilitate the production with 3.12-fold more cycloartenol, whereas *PpOSC2* could not increase the content of cycloartenol but can increase that of an uncharacterized triterpene product instead (Figure S8). Therefore, the *PpOSC1* gene plays an important function in the biosynthesis of polyphyllin.

PpUGT73CR1 functions as a glucotransferase at the C-3 position of diosgenin and pennogenin

In the candidate UGTs, transcript/33044 showed the activity of C-3 glucosyltransferase named *PpUGT73CR1*. In the phylogenetic tree, *PpUGT73CR1* is in the same branch with *BvUGT73CR10* and *BvUGT73CR10*, which can glycosylate the C-3 position of β-amyrin[26]. The coding sequence length of the gene is 1473 bp, encoding a UGT with 490 amino acid residues. The molecular weight of the protein is 54.48 kDa, and the molecular weight of the fusion protein with GST tag is 81.68 kDa (Figure 4c). When diosgenin was used as a sugar acceptor, the enzyme encoded by *PpUGT73CR1* can catalyze diosgenin to produce a polar product after the addition of UDP-glucose. The HPLC retention time of the product was 22.1 min, which was consistent with that of trillin. When pennogenin was used as the sugar acceptor, PpUGT73CR1 converted pennogenin into a new product with the retention time at 18.9 min. The LC-TOF-MS analysis showed that the molecular weight of the new product was 593.37 $[M+H]^+$, which was consistent with the molecular weight of floribundasaponin A (Figure 4d). Supplementary File 3 shows the hydrogen and carbon spectrum results of the substrate (pennogenin) and reaction product (floribundasaponin A).

To study the promiscuity of UGT to diverse substrates, we evaluated the catalytic capability of PpUGT73CR1 on diosgenin and pennogenin. Subsequently, the enzymatic kinetics of PpUGT73CR1 catalyzing different substrates were studied (Table 1). The maximum reaction velocity (*Vmax*) of diosgenin and pennogenin were 177.13 ± 8.91 and 87.7 ± 3.27 nmol/min/mg, respectively. Michaelis constant (K_m) reflected the affinity between the enzyme and substrate to a certain extent. Compared with pennogenin (K_m = 73.43 ± 8.16 μM), PpUGT73CR1 had a higher affinity for diosgenin (K_m = 53.69 ± 9.37 μM). The enzymatic catalytic constant (K_{cat}) of PpUGT73CR1 for diosgenin and pennogenin were 0.24 and 0.12 s^{-1}, respectively, and the calculated conversion efficiencies (K_{cat}/K_m) were 4.47 (diosgenin) and 1.62 (pennogenin) $mM^{-1}{\cdot}s^{-1}$. Thus, the catalytic efficiency of PpUGT73CR1 for diosgenin was higher than that of pennogenin.

DISCUSSION

Medicinal plants are rich in a variety of metabolites with pharmacological properties. However, studies on the biosynthetic pathways of these metabolites are still limited due to the complexity of plant genomes and the lack of genomic resources.

Unlike microorganisms, functional gene clusters are rare in plants, and gene redundancy and strict genetic regulation in plants cause difficulty in parsing metabolic pathways[27]. At present, botanists mainly use multi-omics methods to analyze the biological pathways of metabolites. However, in the absence of genomic data, this prediction method often yields a large number of false positive results. For the majority of medicinal plants, an accurate method is needed to predict metabolite biosynthesis pathways without genomic data or metabolic biosynthesis clusters. The research on the biosynthesis of polyphyllin has been a hot topic because of its various pharmacological activities. Given the huge genome, no data report is currently available on the genome of *P. polyphylla* var. *yunnanensis*. At this stage, transcriptome sequencing is the most suitable method to study the biosynthesis pathway of polyphyllin.

Several studies reported the transcriptome data of *P. polyphylla*[6,14,15,28]. However, these transcriptome measurements were all based on Illumina platform and cannot reflect well the complete transcriptome information of *P. polyphylla* var. *yunnanensis*. Published transcriptome information was mainly derived from roots, stems, and leaves of the plant, with little transcriptome information for other tissues, which is insufficient to predict polyphyllin biosynthesis pathway genes using the association analysis of gene expression and metabolites. In this study, we collected samples from eight tissues of *P. polyphylla* var. *yunnanensis* to complete the splicing and full-length transcriptome sequencing, and more than 370 G clean data were obtained. The transcriptome data in our study obtained the most diversity in tissues and deepest sequencing depth among all transcriptome experiments in *P. polyphylla* var. *yunnanensis*. Compared with previous reports of sequencing, it can avoid redundant data and splicing errors and provide data support for the accurate prediction of the biosynthetic pathway of polyphyllin.

We constructed a co-expression network of metabolites with gene expression levels in different tissue of *P. polyphylla* var. *yunnanensis*. We measured the contents of key metabolites related to polyphyllin synthesis in various tissues of *P. polyphylla* var. *yunnanensis*, including cholesterol, diosgenin, trillin, prosapogenin A, polyphyllins I, II, VI, and VII. We predicted three modules that are highly relevant to the above metabolites. We further conducted conditional screening through phylogenetic trees and gene expression levels and obtained reliable candidate genes, including three reported key *CYP* genes involved in the biosynthesis of polyphyllins. We also identified an *OSC* gene responsible for cycloartenol biosynthesis and a *UGT* gene with C-3 glucosyltransferase function from *P. polyphylla* var. *yunnanensis*. Among the predicted modules, the coral module showed a strong correlation only with polyphyllin I, whereas the lavenderblush3 and antiquewhite4 modules exhibited more correlations with cholesterol, diosgenin, and trillin. This finding suggests that the genes contained in the coral module may be more involved in the formation of hydroxylation and glycosylation of polyphyllins. The predicted results also proved our speculation that the three predicted CYP genes with clear function and the C-3 glucosyltransferase gene of polyphyllins (*PpUGT73CR1*) all come from the coral module. These results showed the accuracy and reliability of this prediction method.

Several studies focused on glycosylation modification of steroidal sapogenins. A C-3 glycosyltransferase SAGT4a in *Solanum aculeatissimum* shows the glycosylation activity of diosgenin, nuatigenin, tigogenin, and other glycosyltransferases[29]. In this study, c-3 glycosyltransferases of diosgenin and pennogenin were found in in *P. polyphylla* var.

yunnanensis. According to the study of substrate promiscuity and enzyme kinetics, PpUGT73CR1 had a better affinity and catalytic capability with diosgenin compared with pennogenin.

At present, the analysis of steroidal sapogenin biosynthetic pathways is progressing slowly and remains in the exploratory stage. Research has focused on the cloning and regulation of functional genes upstream of terpenoid biosynthetic pathways, such as HMGR, FPS, SS, CAS, etc.[30,31], and several P450s. Other research reported the glycosylation modification of diosgenin. Therefore, the key genes involved in the biosynthesis pathway of polyphyllin can be possibly predicted by using our prediction method combining the evolutionary tree, co-expression network, and gene expression quantity. This method is also generally applicable to the prediction of key genes in plants lacking genome data.

CONCLUSION

Polyphyllin has a variety of pharmacological activities, but the analysis of the biosynthetic pathway of polyphyllin is incomplete. We performed splicing and full-length transcriptome sequencing of rhizomes, fibrous roots, stems, leaves, ripe fruits, stigma, petals, and pistil tissues of *Paris polyphylla* var. *yunnanensis*, and the gene expression and WGCNA method were used to predict the OSCs, CYPs, and UGTs involved in the biosynthesis of sapogenin. Among the predicted candidate genes, we identified an *OSC* gene (*PpOSC1*) and a diosgenin/pennogenin C-3 UGT gene (*PpUGT73CR1*) for the first time. This study improves our understanding of the biosynthetic pathways of polyphyllins, providing a basis for further elucidation of the pharmacologically active triterpene/sapogenin biosynthesis and an efficient strategy to study the complex pathway of other specialized plant metabolites.

ACKNOWLEDGEMENTS

This work was supported by the program for Natural Science Foundation of Heilongjiang Province of China (YQ2020C020), Opening Project of Zhejiang Provincial Preponderant and Characteristic Subject of Key University (Traditional Chinese Pharmacology), the Zhejiang Chinese Medical University (No. ZYAOX2018012), the National Natural Science Foundation of China (NSFC Grant No.31770332, 31970314), National Key R&D Program of China (2019YFC1711103), and the Fundamental Research Funds for the Central Universities (2572020BD01) and Key project at central government level: The ability establishment of sustainable use for valuable Chinese medicine resources (2060302-2101-17).

The authors also acknowledge the technical support of Dr. Shengnan Tan from Analysis and Test Center, Northeast Forestry University.

AUTHOR CONTRIBUTIONS

Z, X and X.H designed the experiments and coordinated the project. K.W, X.Y and C.H performed the samples collection, phylogenetic tree, OSC function, transcriptomic and metabolomic analyses. X.H wroted and edited most of the manuscript. B,D edited the language. All authors have read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

Raw reads have been deposited as a BioProject under accession PRJCA004404 (https://bigd.big.ac.cn/bioproject/browse/ PRJCA004404).

COMPETING INTERESTS

The authors declare that they have no conflict of interest. Supplementary Information accompanies this paper at website.

REFERENCES

1 Tang, M. J., Zhao, J., Li, X. H. & Yu, S. S. [Advances in studies on chemical constituents and pharmacological activities from plants of Symplocaceae]. *China journal of Chinese materia medica* **29**, 390-394 (2004).

2 Wang, Y., Zhang, Y. J., Gao, W. Y. & Yan, L. L. [Anti-tumor constituents from Paris polyphylla var. yunnanensis]. *China journal of Chinese materia medica* **32**, 1425-1428 (2007).

3 Guo, L. *et al.* Active pharmaceutical ingredients and mechanisms underlying phasic myometrial contractions stimulated with the saponin extract from Paris polyphylla Sm. var. yunnanensis used for abnormal uterine bleeding. *Human reproduction* **23**, 964-971, doi:10.1093/humrep/den001 (2008).

4 Qin, X. J. *et al.* Steroidal saponins with antimicrobial activity from stems and leaves of Paris polyphylla var. yunnanensis. *Steroids* **77**, 1242-1248, doi:10.1016/j.steroids.2012.07.007 (2012).

5 Negi, J. S. *et al.* Paris polyphylla: chemical and biological prospectives. *Anti-cancer agents in medicinal chemistry* **14**, 833-839, doi:10.2174/1871520614666140611101040 (2014).

6 Yin, Y., Gao, L., Zhang, X. & Gao, W. A cytochrome P450 monooxygenase responsible for the C-22 hydroxylation step in the Paris polyphylla steroidal saponin biosynthesis pathway. *Phytochemistry* **156**, 116-123, doi:10.1016/j.phytochem.2018.09.005 (2018).

7 Shuli, M. *et al.* Paridis saponins inhibiting carcinoma growth and metastasis in

vitro and in vivo. *Archives of pharmacal research* **34**, 43-50, doi:10.1007/s12272-011-0105-4 (2011).

8 Patel, K., Gadewar, M., Tahilyani, V. & Patel, D. K. A review on pharmacological and analytical aspects of diosmetin: a concise report. *Chinese journal of integrative medicine* **19**, 792-800, doi:10.1007/s11655-013-1595-3 (2013).

9 Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P. & Osbourn, A. Triterpene biosynthesis in plants. *Annual review of plant biology* **65**, 225-257, doi:10.1146/annurev-arplant-050312-120229 (2014).

10 Cardenas, P. D. *et al.* The bitter side of the nightshades: Genomics drives discovery in Solanaceae steroidal alkaloid metabolism. *Phytochemistry* **113**, 24-32, doi:10.1016/j.phytochem.2014.12.010 (2015).

11 Lu, Y. *et al.* Regulation of the cholesterol biosynthetic pathway and its integration with fatty acid biosynthesis in the oleaginous microalga Nannochloropsis oceanica. *Biotechnology for biofuels* **7**, 81, doi:10.1186/1754-6834-7-81 (2014).

12 Christ, B. *et al.* Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. *Nature communications* **10**, 3206, doi:10.1038/s41467-019-11286-7 (2019).

13 Morozova, O., Hirst, M. & Marra, M. A. Applications of new sequencing technologies for transcriptome analysis. *Annual review of genomics and human genetics* **10**, 135-151, doi:10.1146/annurev-genom-082908-145957 (2009).

14 Liu, T., Li, X., Xie, S., Wang, L. & Yang, S. RNA-seq analysis of Paris polyphylla var. yunnanensis roots identified candidate genes for saponin synthesis. *Plant diversity* **38**, 163-170, doi:10.1016/j.pld.2016.05.002 (2016).

15 Yang, Z. *et al.* Transcriptome analyses of Paris polyphylla var. chinensis, Ypsilandra thibetica, and Polygonatum kingianum characterize their steroidal saponin biosynthesis pathway. *Fitoterapia* **135**, 52-63, doi:10.1016/j.fitote.2019.04.008 (2019).

16 Han, J. Y., In, J. G., Kwon, Y. S. & Choi, Y. E. Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in Panax ginseng. *Phytochemistry* **71**, 36-46, doi:10.1016/j.phytochem.2009.09.031 (2010).

17 Rai, A., Saito, K. & Yamazaki, M. Integrated omics analysis of specialized metabolism in medicinal plants. *The Plant journal : for cell and molecular biology* **90**, 764-787, doi:10.1111/tpj.13485 (2017).

18 Mylona, P. *et al.* Sad3 and sad4 are required for saponin biosynthesis and root development in oat. *Plant Cell* **20**, 201-212, doi:10.1105/tpc.107.056531 (2008).

19 Shang, Y. *et al.* Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084-1088, doi:10.1126/science.1259215 (2014).

20 Xu, J. *et al.* Panax ginseng genome examination for ginsenoside biosynthesis. *Gigascience* **6**, 1-15, doi:10.1093/gigascience/gix093 (2017).

21 Jiang, Z. *et al.* The chromosome-level reference genome assembly for Panax notoginseng and insights into ginsenoside biosynthesis. *Plant Commun* **2**, 100113, doi:10.1016/j.xplc.2020.100113 (2021).

22 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-1874,

doi:10.1093/molbev/msw054 (2016).

23 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

24 Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology* **1962**, 227-245, doi:10.1007/978-1-4939-9173-0_14 (2019).

25 Wilson, A. E. & Tian, L. Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism. *The Plant journal : for cell and molecular biology* **100**, 1273-1288, doi:10.1111/tpj.14514 (2019).

26 Augustin, J. M. *et al.* UDP-Glycosyltransferases from the UGT73C Subfamily in Barbarea vulgaris Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance. *Plant physiology* **160**, 1881-1895, doi:10.1104/pp.112.202747 (2012).

27 Karaiskos, I., Souli, M., Galani, I. & Giamarellou, H. Colistin: still a lifesaver for the 21st century? *Expert opinion on drug metabolism & toxicology* **13**, 59-71, doi:10.1080/17425255.2017.1230200 (2017).

28 Qi, J. *et al.* Mining genes involved in the stratification of Paris polyphylla seeds using high-throughput embryo transcriptome sequencing. *BMC genomics* **14**, 358, doi:10.1186/1471-2164-14-358 (2013).

29 Kohara, A. *et al.* A novel glucosyltransferase involved in steroid saponin biosynthesis in Solanum aculeatissimum. *Plant molecular biology* **57**, 225-239,

doi:10.1007/s11103-004-7204-2 (2005).

30 Babiychuk, E. *et al.* Albinism and cell viability in cycloartenol synthase deficient Arabidopsis. *Plant signaling & behavior* **3**, 978-980, doi:10.4161/psb.6173 (2008).

31 Kumar, S. *et al.* RNA-Seq mediated root transcriptome analysis of Chlorophytum borivilianum for identification of genes involved in saponin biosynthesis. *Functional & integrative genomics* **16**, 37-55, doi:10.1007/s10142-015-0465-9 (2016).

Figure legends

Figure 1. Possible biosynthetic pathways of polyphyllin in *P. polyphylla* var. *yunnanensis*. The established metabolic pathways are represented by solid line arrows, while the speculated metabolic pathways are represented by dotted line arrows. Genes predicted by transcriptome, metabolite profile and WGCAN analysis are shown in yellow background, and those identified by this method are shown in blue background.

Figure 2. Candidate OSC genes and gene function verification. (a) Phylogenetic tree of OSCs. Predicted amino acid sequences of OSCs in *P. polyphylla* var. *yunnanensis* were aligned with selected OSCs from other plant species using MUSCLE. The evolutionary history was inferred using the maximum likelihood method. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed. **(b)** Functional verification of *PpOSC* gene. Two OSC genes were identified in *P. polyphylla* var. *yunnanensis*, and the results of GC-MS showed that *PpOSC1* gene increased the yield of cycloartenol after being transferred to *N.benthamiana*.

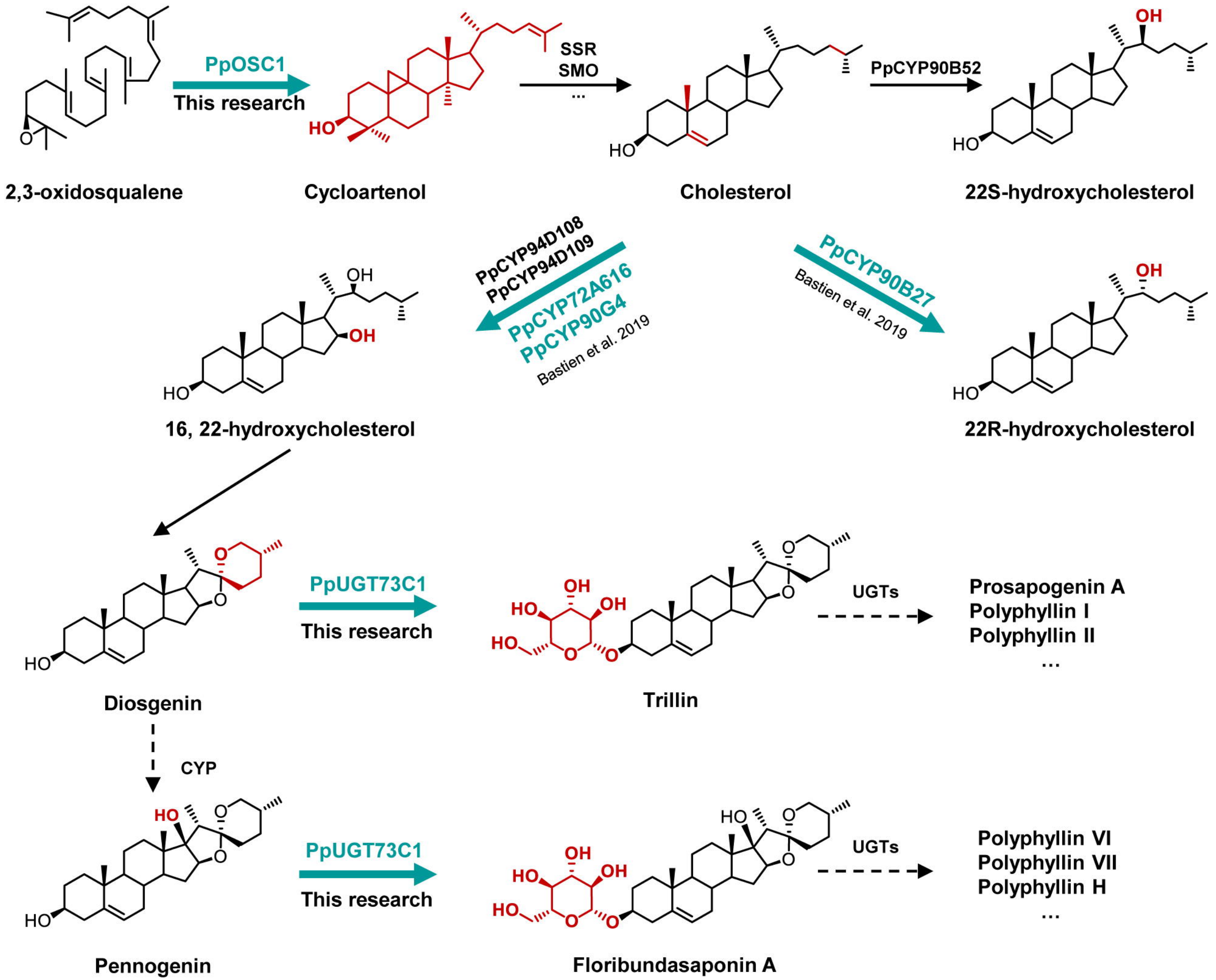
Figure 3. Using phylogenetic tree, metabolic profile and WGCAN analysis to predict that the key genes involved in the biosynthesis of polyphyllin in *P. polyphylla* var. *yunnanensis*. (a) Phylogenetic tree of CYPs and UGTs. **(b)** Module–trait associations. Each row corresponds to a module of characteristic genes, and each column corresponds to a metabolite. Each cell contains the

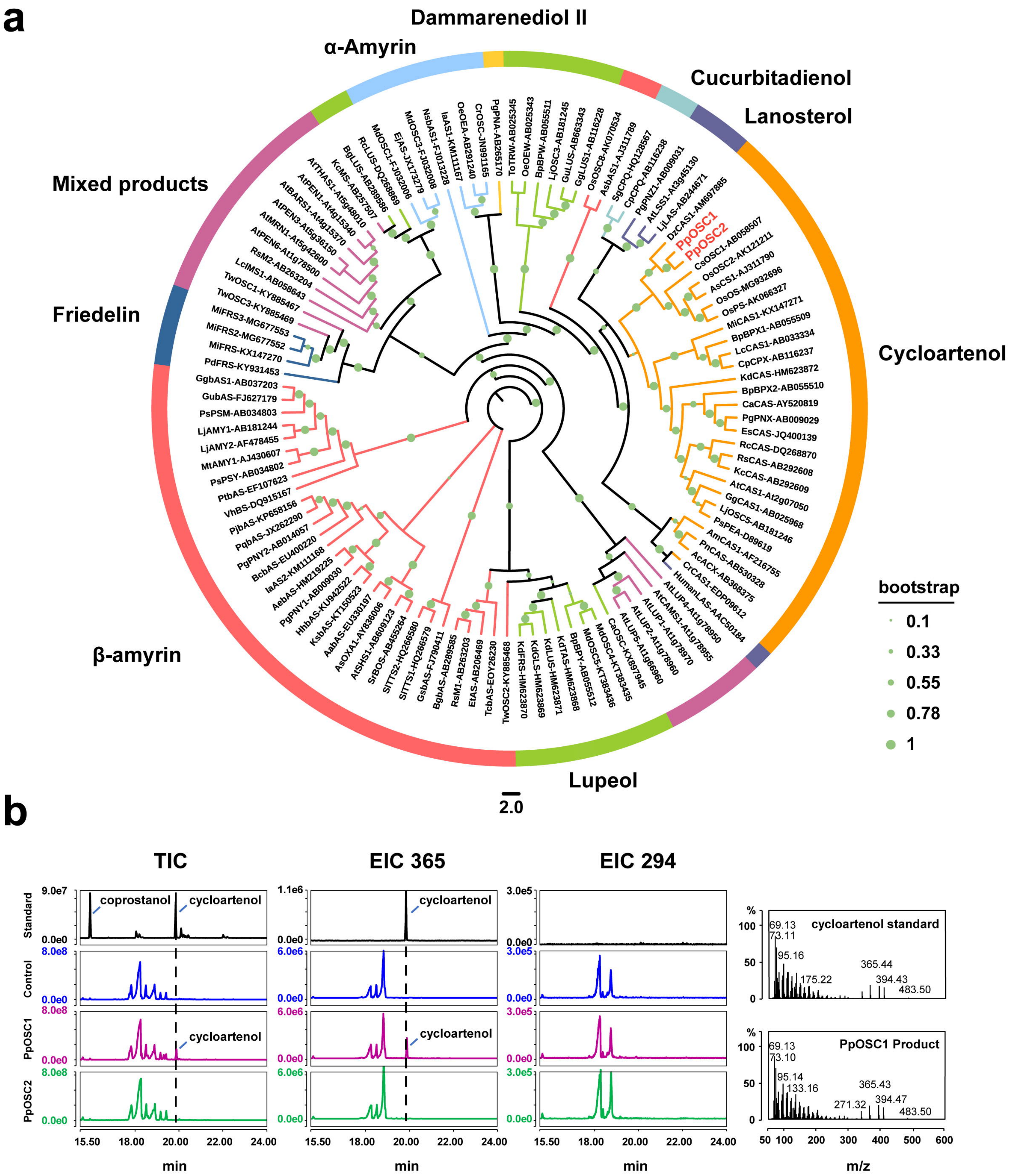
correlation and p value of the genes in the module with the corresponding metabolite. **(c)** The production profiles of key metabolites involved in the biosynthetic pathway of polyphyllin in different tissues. The quantification of Polyphyllins contents was carried out in three separate experiments, in which each sample came from different mixtures of 4 plants.

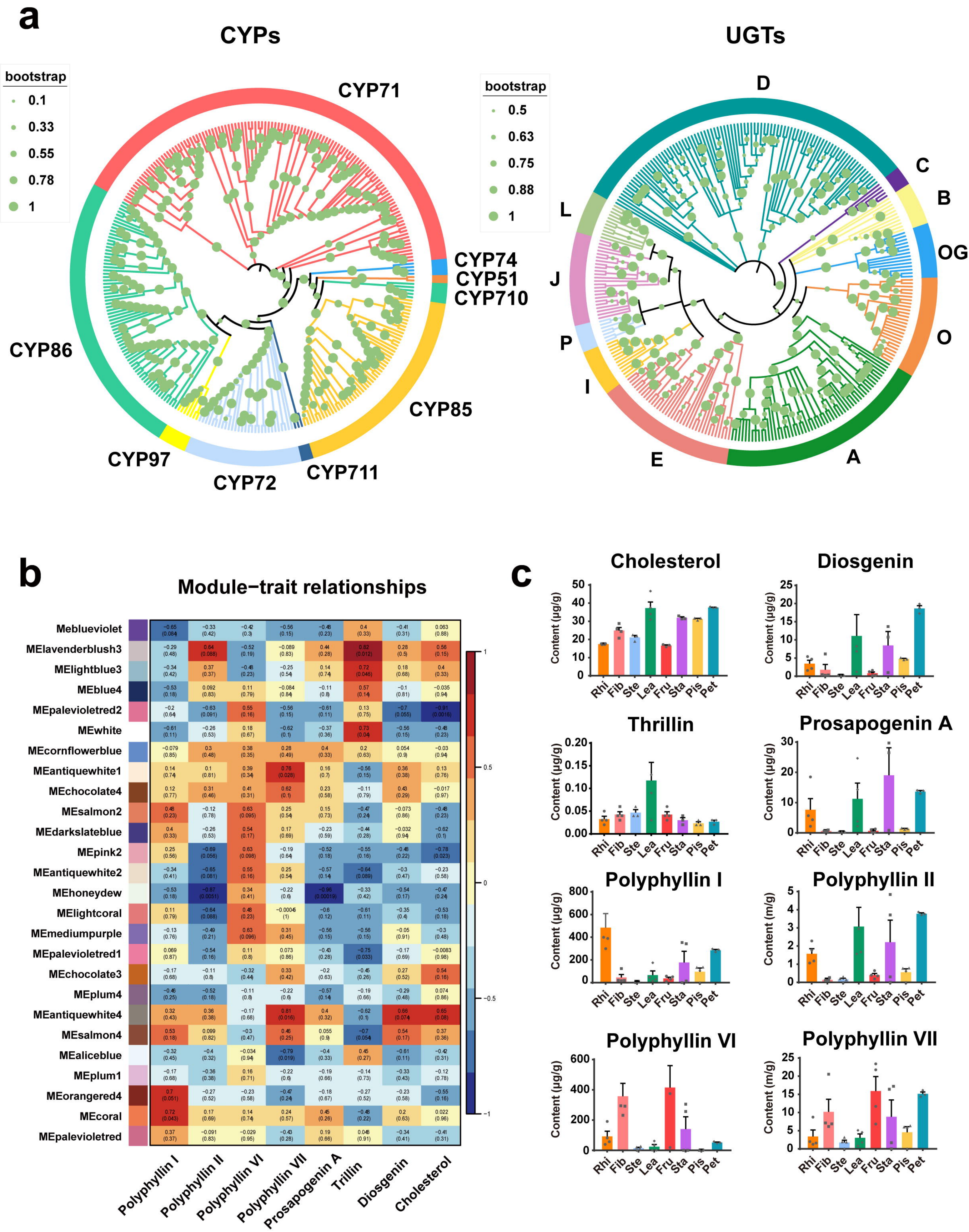
Figure 4. Candidate CYP and UGT genes and gene function verification. (a) Heatmaps of the expression levels of candidate CYPs and UGTs in different tissues of *P. polyphylla* var. *yunnanensis*. All genes are arranged from top to bottom according to the total expression level. The asterisks represent key genes predicted by the evolutionary tree, WGCAN and gene expression. **(b)** Venn diagrams of candidate genes. Phylogenetic tree and WGCAN methods were used to predict candidate CYPs and UGTs, among which 15 and 20 CYPs and UGTs could be predicted by the two methods, respectively. **(c)** SDS-PAGE analysis of expressed PpUGT73CR1 protein. *Lanes*: M, protein molecular weight marker (Thermo fisher); 1, pGEX-6p-1 vector transformed in *E. coli* Rosetta (DE3) cells with IPTG induction; 2, pGEX-*UGT73CR1* vector transformed in *E. coli* Rosetta (DE3) cells without IPTG induction; 3, pGEX-*UGT73CR1* vector transformed in *E. coli* Rosetta (DE3) cells with IPTG induction; 4, the purified recombinant protein of PpUGT73CR1. **(d)** Functional verification of *PpUGT73CR1* gene. The functional verification of candidate UGTs was performed by HPLC and LC-TOF-MS, and the enzyme encoded by *PpUGT73CR1* gene could introduce glucose group into C-3 of diosgenin and pannogenin.

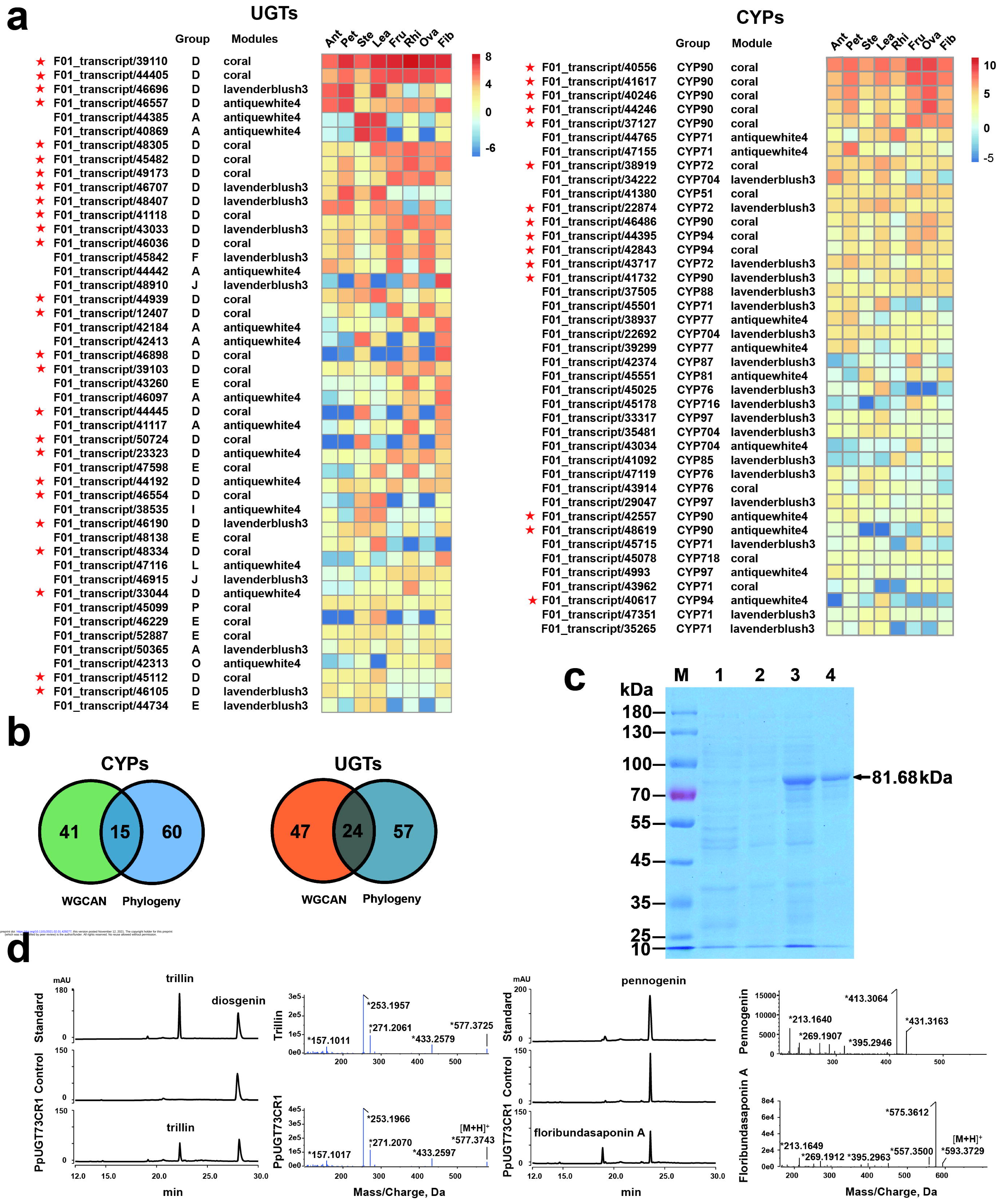
Table 1. The enzymatic kinetics of PpUGT73C1 catalyzed diosgenin and pennogenin.

enzyme	**substrate**	**V_{max}(μM/min)**	**K_m(μM)**	**K_{cat}(s^{-1})**	**K_{cat} / K_m(s^{-1}·mM^{-1})**
PpUGT73C1	diosgenin	1.771 ± 0.089	53.69 ± 9.37	0.24	4.47
	pennogenin	0.877 ± 0.033	73.43 ± 8.16	0.12	1.62


2,3-oxidosqualene
PpOSC1
This research
Cycloartenol
SSR
SMO
...
Cholesterol
PpCYP90B52
22S-hydroxycholesterol
PpCYP94D108
PpCYP94D109
PpCYP72A616
PpCYP90G4
Bastien et al. 2019
16, 22-hydroxycholesterol
PpCYP90B27
Bastien et al. 2019
22R-hydroxycholesterol
Diosgenin
PpUGT73C1
This research
Trillin
UGTs
Prosapogenin A
Polyphyllin I
Polyphyllin II
...
CYP
Pennogenin
PpUGT73C1
This research
Floribundasaponin A
UGTs
Polyphyllin VI
Polyphyllin VII
Polyphyllin H
...


a
Dammarenediol II
α-Amyrin
Cucurbitadienol
Lanosterol
Mixed products
Friedelin
Cycloartenol
β-amyrin
Lupeol
bootstrap
0.1
0.33
0.55
0.78
1
2.0
MiFRS-KX147270
PdFRS-KY931453
MiFRS2-MG677552
MiFRS3-MG677553
TwOSC3-KY885469
TwOSC1-KY885467
LcIMS1-AB058643
RsM2-AB263204
AtPEN6-At1g78500
AtMRN1-At5g42600
AtPEN3-At5g36150
AtBARS1-At4g15370
AtPEN1-At4g15340
AtTHAS1-At5g48010
KcMS-AB257507
BgLUS-AB289586
RcLUS-DQ268869
MdOSC1-FJ032006
EjAS-JX173279
MdOSC3-FJ032008
NsbAS1-FJ013228
IaAS1-KM111167
OeOEA-AB291240
CrOSC-JN991165
PgPNA-AB265170
ToTRW-AB025345
OeOEW-AB025343
BpBPW-AB055511
LjOSC3-AB181245
GuLUS-AB663343
GgLUS1-AB116228
OsOSC8-AK070534
AsbAS1-AJ311789
SgCPQ-HQ128567
CpCPQ-AB116238
PgPNZ1-AB009031
AtLSS1-At3g45130
LjLAS-AB244671
DzCAS1-AM697885
PpOSC1
PpOSC2
CsOSC1-AB058507
OsOSC2-AK121211
AsCS1-AJ311790
OsOS-MG932696
OsPS-AK066327
MiCAS1-KX147271
BpBPX1-AB055509
LcCAS1-AB033334
CpCPX-AB116237
KdCAS-HM623872
BpBPX2-AB055510
CaCAS-AY520819
PgPNX-AB009029
EsCAS-JQ400139
RcCAS-DQ268870
RsCAS-AB292608
KcCAS-AB292609
AtCAS1-At2g07050
GgCAS1-AB025968
LjOSC5-AB181246
PsPEA-D89619
AmCAS1-AF216755
PnCAS-AB530328
AcACX-AB368375
CrCAS1-EDP09612
HumanLAS-AAC50184
AtLUP4-At1g78950
AtCAMS1-At1g78955
AtLUP1-At1g78970
AtLUP2-At1g78960
AtLUP5-At1g66960
CaOSC-KX897945
MdOSC4-KT383435
MdOSC5-KT383436
BpBPY-AB055512
KdTAS-HM623868
KdLUS-HM623871
KdGLS-HM623869
KdFRS-HM623870
TwOSC2-KY885468
TcbAS-EOY26230
EtAS-AB206469
RsM1-AB263203
BgbAS-AB289585
GsbAS-FJ790411
SlTTS1-HQ266579
SlTTS2-HQ266580
SrBOS-AB455264
AtSHS1-AB609123
AsOXA1-AY836006
AabAS-EU330197
KsbAS-KT150523
HhbAS-KU942522
PgPNY1-AB009030
AebAS-HM219225
IaAS2-KM111168
BcbAS-EU400220
PgPNY2-AB014057
PqbAS-JX262290
PjbAS-KP658156
VhBS-DQ915167
PtbAS-EF107623
PsPSY-AB034802
MtAMY1-AJ430607
LjAMY2-AF478455
LjAMY1-AB181244
PsPSM-AB034803
GubAS-FJ627179
GgbAS1-AB037203
b
TIC
EIC 365
EIC 294
Standard
Control
PpOSC1
PpOSC2
9.0e7
0.0e0
8.0e8
0.0e0
8.0e8
0.0e0
8.0e8
0.0e0
coprostanol
cycloartenol
cycloartenol
1.1e6
0.0e0
6.0e6
0.0e0
6.0e6
0.0e0
6.0e6
0.0e0
cycloartenol
cycloartenol
3.0e5
0.0e0
3.0e5
0.0e0
3.0e5
0.0e0
3.0e5
0.0e0
15.50
18.00
20.00
22.00
24.00
min
%
100
50
0
69.13
73.11
95.16
175.22
365.44
394.43
483.50
cycloartenol standard
%
100
50
0
69.13
73.10
95.14
133.16
271.32
365.43
394.47
483.50
PpOSC1 Product
50
100
200
300
400
500
600
m/z


a
CYPs
bootstrap
0.1
0.33
0.55
0.78
1
CYP71
CYP74
CYP51
CYP710
CYP85
CYP711
CYP72
CYP97
CYP86
UGTs
bootstrap
0.5
0.63
0.75
0.88
1
D
C
B
OG
O
A
E
I
P
J
L
b
Module−trait relationships
Meblueviolet −0.65 (0.084) −0.33 (0.42) −0.42 (0.3) −0.56 (0.15) −0.48 (0.23) 0.4 (0.33) −0.41 (0.31) 0.063 (0.88)
MElavenderblush3 −0.29 (0.48) 0.64 (0.088) −0.52 (0.19) −0.089 (0.83) 0.44 (0.28) 0.82 (0.012) 0.28 (0.5) 0.56 (0.15)
MElightblue3 −0.34 (0.42) 0.37 (0.37) −0.48 (0.23) −0.25 (0.54) 0.14 (0.74) 0.72 (0.046) 0.18 (0.68) 0.4 (0.33)
MEblue4 −0.53 (0.18) 0.092 (0.83) 0.11 (0.79) −0.084 (0.84) −0.11 (0.8) 0.57 (0.14) −0.1 (0.81) −0.035 (0.94)
MEpalevioletred2 −0.2 (0.64) −0.63 (0.091) 0.55 (0.16) −0.56 (0.15) −0.61 (0.11) 0.13 (0.75) −0.7 (0.055) −0.91 (0.0016)
MEwhite −0.61 (0.11) −0.26 (0.53) 0.18 (0.67) −0.62 (0.1) −0.37 (0.36) 0.73 (0.04) −0.56 (0.15) −0.48 (0.23)
MEcornflowerblue −0.079 (0.85) 0.3 (0.48) 0.38 (0.35) 0.28 (0.49) 0.4 (0.33) 0.2 (0.63) 0.054 (0.9) −0.03 (0.94)
MEantiquewhite1 0.14 (0.74) 0.1 (0.81) 0.39 (0.34) 0.76 (0.028) 0.16 (0.7) −0.56 (0.15) 0.36 (0.38) 0.13 (0.76)
MEchocolate4 0.12 (0.77) 0.31 (0.46) 0.41 (0.31) 0.62 (0.1) 0.23 (0.58) −0.11 (0.79) 0.43 (0.29) −0.017 (0.97)
MEsalmon2 0.48 (0.23) −0.12 (0.78) 0.63 (0.095) 0.25 (0.54) 0.15 (0.73) −0.47 (0.24) −0.073 (0.86) −0.48 (0.23)
MEdarkslateblue 0.4 (0.33) −0.26 (0.53) 0.54 (0.17) 0.17 (0.69) −0.23 (0.59) −0.44 (0.28) −0.032 (0.94) −0.62 (0.1)
MEpink2 0.25 (0.56) −0.69 (0.056) 0.63 (0.098) −0.19 (0.64) −0.52 (0.18) −0.55 (0.16) −0.48 (0.22) −0.78 (0.023)
MEantiquewhite2 −0.34 (0.41) −0.65 (0.081) 0.55 (0.16) 0.25 (0.54) −0.57 (0.14) −0.64 (0.089) −0.3 (0.47) −0.23 (0.58)
MEhoneydew −0.53 (0.18) −0.87 (0.0051) 0.34 (0.41) −0.22 (0.6) −0.96 (0.00019) −0.33 (0.42) −0.54 (0.17) −0.47 (0.24)
MElightcoral 0.11 (0.79) −0.64 (0.088) 0.48 (0.23) −0.00045 (1) −0.6 (0.12) −0.61 (0.11) −0.35 (0.4) −0.53 (0.18)
MEmediumpurple −0.13 (0.76) −0.49 (0.21) 0.63 (0.096) 0.31 (0.45) −0.56 (0.15) −0.56 (0.15) −0.05 (0.91) −0.3 (0.48)
MEpalevioletred1 0.069 (0.87) −0.54 (0.16) 0.11 (0.8) 0.073 (0.86) −0.43 (0.28) −0.75 (0.033) −0.17 (0.69) −0.0083 (0.98)
MEchocolate3 −0.17 (0.68) −0.11 (0.8) −0.32 (0.44) 0.33 (0.42) −0.2 (0.63) −0.45 (0.26) 0.27 (0.52) 0.54 (0.16)
MEplum4 −0.46 (0.25) −0.52 (0.18) −0.11 (0.8) −0.22 (0.6) −0.57 (0.14) −0.19 (0.66) −0.29 (0.48) 0.074 (0.86)
MEantiquewhite4 0.32 (0.43) 0.36 (0.38) −0.17 (0.68) 0.81 (0.016) 0.4 (0.32) −0.62 (0.1) 0.66 (0.074) 0.65 (0.08)
MEsalmon4 0.53 (0.18) 0.099 (0.82) −0.3 (0.47) 0.46 (0.25) 0.055 (0.9) −0.7 (0.054) 0.54 (0.17) 0.37 (0.36)
MEaliceblue −0.32 (0.45) −0.4 (0.32) −0.034 (0.94) −0.79 (0.019) −0.4 (0.33) 0.45 (0.27) −0.61 (0.11) −0.42 (0.31)
MEplum1 −0.17 (0.68) −0.36 (0.38) 0.16 (0.71) −0.22 (0.6) −0.19 (0.66) −0.14 (0.73) −0.33 (0.43) −0.12 (0.78)
MEorangered4 0.7 (0.051) −0.27 (0.52) −0.23 (0.58) −0.47 (0.24) −0.18 (0.67) −0.27 (0.52) −0.23 (0.58) −0.55 (0.16)
MEcoral 0.72 (0.043) 0.17 (0.69) 0.14 (0.74) 0.24 (0.57) 0.45 (0.26) −0.48 (0.22) 0.2 (0.63) 0.022 (0.96)
MEpalevioletred 0.37 (0.37) −0.091 (0.83) −0.029 (0.95) −0.43 (0.28) 0.19 (0.66) 0.048 (0.91) −0.34 (0.41) −0.41 (0.31)
Polyphyllin I
Polyphyllin II
Polyphyllin VI
Polyphyllin VII
Prosapogenin A
Trillin
Diosgenin
Cholesterol
1
0.5
0
−0.5
−1
c
Cholesterol
Content (μg/g)
0 10 20 30 40 50
Diosgenin
Content (μg/g)
0 5 10 15 20 25
Thrillin
Content (μg/g)
0.00 0.05 0.10 0.15 0.20
Prosapogenin A
Content (μg/g)
0 10 20 30
Polyphyllin I
Content (μg/g)
0 200 400 600 800
Polyphyllin II
Content (m/g)
0 1 2 3 4 5
Polyphyllin VI
Content (μg/g)
0 200 400 600
Polyphyllin VII
Content (m/g)
0 5 10 15 20 25
Rhi Fib Ste Lea Fru Sta Pis Pet


a
UGTs
Group
Modules
Ant
Pet
Ste
Lea
Fru
Rhi
Ova
Fib
8
4
0
-6
★ F01_transcript/39110 D coral
★ F01_transcript/44405 D coral
★ F01_transcript/46696 D lavenderblush3
★ F01_transcript/46557 D antiquewhite4
F01_transcript/44385 A antiquewhite4
F01_transcript/40869 A antiquewhite4
★ F01_transcript/48305 D coral
★ F01_transcript/45482 D coral
★ F01_transcript/49173 D coral
★ F01_transcript/46707 D lavenderblush3
★ F01_transcript/48407 D lavenderblush3
★ F01_transcript/41118 D coral
★ F01_transcript/43033 D lavenderblush3
★ F01_transcript/46036 D coral
F01_transcript/45842 F lavenderblush3
F01_transcript/44442 A antiquewhite4
F01_transcript/48910 J lavenderblush3
★ F01_transcript/44939 D coral
★ F01_transcript/12407 D coral
F01_transcript/42184 A antiquewhite4
F01_transcript/42413 A antiquewhite4
★ F01_transcript/46898 D coral
★ F01_transcript/39103 D coral
F01_transcript/43260 E coral
F01_transcript/46097 A antiquewhite4
★ F01_transcript/44445 D coral
F01_transcript/41117 A antiquewhite4
★ F01_transcript/50724 D coral
★ F01_transcript/23323 D antiquewhite4
F01_transcript/47598 E coral
★ F01_transcript/44192 D antiquewhite4
★ F01_transcript/46554 D coral
F01_transcript/38535 I antiquewhite4
★ F01_transcript/46190 D lavenderblush3
F01_transcript/48138 E coral
★ F01_transcript/48334 D coral
F01_transcript/47116 L antiquewhite4
F01_transcript/46915 J lavenderblush3
★ F01_transcript/33044 D antiquewhite4
F01_transcript/45099 P coral
F01_transcript/46229 E coral
F01_transcript/52887 E coral
F01_transcript/50365 A lavenderblush3
F01_transcript/42313 O antiquewhite4
★ F01_transcript/45112 D coral
★ F01_transcript/46105 D lavenderblush3
F01_transcript/44734 E lavenderblush3
CYPs
Group
Module
Ant
Pet
Ste
Lea
Rhi
Fru
Ova
Fib
10
5
0
-5
★ F01_transcript/40556 CYP90 coral
★ F01_transcript/41617 CYP90 coral
★ F01_transcript/40246 CYP90 coral
★ F01_transcript/44246 CYP90 coral
★ F01_transcript/37127 CYP90 coral
F01_transcript/44765 CYP71 antiquewhite4
F01_transcript/47155 CYP71 antiquewhite4
★ F01_transcript/38919 CYP72 coral
F01_transcript/34222 CYP704 lavenderblush3
F01_transcript/41380 CYP51 coral
★ F01_transcript/22874 CYP72 lavenderblush3
★ F01_transcript/46486 CYP90 coral
★ F01_transcript/44395 CYP94 coral
★ F01_transcript/42843 CYP94 coral
★ F01_transcript/43717 CYP72 lavenderblush3
★ F01_transcript/41732 CYP90 lavenderblush3
F01_transcript/37505 CYP88 lavenderblush3
F01_transcript/45501 CYP71 lavenderblush3
F01_transcript/38937 CYP77 antiquewhite4
F01_transcript/22692 CYP704 lavenderblush3
F01_transcript/39299 CYP77 antiquewhite4
F01_transcript/42374 CYP87 lavenderblush3
F01_transcript/45551 CYP81 antiquewhite4
F01_transcript/45025 CYP76 lavenderblush3
F01_transcript/45178 CYP716 lavenderblush3
F01_transcript/33317 CYP97 lavenderblush3
F01_transcript/35481 CYP704 lavenderblush3
F01_transcript/43034 CYP704 antiquewhite4
F01_transcript/41092 CYP85 lavenderblush3
F01_transcript/47119 CYP76 lavenderblush3
F01_transcript/43914 CYP76 coral
F01_transcript/29047 CYP97 lavenderblush3
★ F01_transcript/42557 CYP90 antiquewhite4
★ F01_transcript/48619 CYP90 antiquewhite4
F01_transcript/45715 CYP71 lavenderblush3
F01_transcript/45078 CYP718 coral
F01_transcript/4993 CYP97 antiquewhite4
F01_transcript/43962 CYP71 coral
★ F01_transcript/40617 CYP94 antiquewhite4
F01_transcript/47351 CYP71 lavenderblush3
F01_transcript/35265 CYP71 lavenderblush3
b
CYPs
41
15
60
WGCAN
Phylogeny
UGTs
47
24
57
WGCAN
Phylogeny
c
M
1
2
3
4
kDa
180
130
100
70
55
45
35
25
10
81.68 kDa
d
mAU
180
0
150
0
150
0
Standard
PpUGT73CR1 Control
trillin
diosgenin
trillin
12.0
15.0
20.0
25.0
30.0
min
Trillin
3e5
2e5
1e5
0e0
*253.1957
*271.2061
*577.3725
*157.1011
*433.2579
PpUGT73CR1
4e5
3e5
2e5
1e5
0e0
*253.1966
*271.2070
[M+H]⁺
*577.3743
*157.1017
*433.2597
200
300
400
500
Mass/Charge, Da
mAU
200
0
150
0
150
0
Standard
PpUGT73CR1 Control
pennogenin
floribundasaponin A
12.0
15.0
20.0
25.0
30.0
min
Pennogenin
15000
10000
5000
0
*413.3064
*213.1640
*431.3163
*269.1907
*395.2946
Floribundasaponin A
8e4
6e4
4e4
2e4
0e0
*575.3612
[M+H]⁺
*213.1649
*269.1912
*395.2963
*557.3500
*593.3729
200
300
400
500
600
Mass/Charge, Da