# Coupling Comprehensive Transcriptome-Metabolome Association and Phylogenetic Analysis Speed Dissecion of Polyphyllins Biosynthetic Pathway

Xin Hua[1,&], Wei Song[2,&], Kangzong Wang[1,&], Xue Yin[1], Changqi Hao[1], 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 150040, 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 100193, China

[4] Beijing Vegetable Research Center (BVRC), Beijing Academy of Agriculture and Forestry Science (BAAFS), Beijing 100097, China

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

*Paris polyphylla* var. *yunnanensis* is an endangered herbaceous plant accumulating polyphyllins, which are widely used in clinical treatment in China. The genes involved in the biosynthesis of polyphyllins are often mixed with other steroids biosynthetic genes, forming a very complex biosynthetic network. The lack of genomic data and tissue specificity of genes makes it extremely difficult to study the biosynthetic pathway of polyphyllins. Here we report an effective method for predicting key genes of polyphyllins biosynthesis. Eight different organs were selected for metabolic analysis and transcriptome sequencing using both of PacBio and Illumina platform generating a total of 370 G of pure data, and two OSCs, 216 CYPs, and 199 UGTs were annotated. By constructing phylogenetic trees, we screened out 60 and 57 candidate genes in the CYP72, CYP90, CYP94 families of CYPs and group D of UGTs, respectively. Metabolic analysis indicates cholesterol, diosgenin, and polyphyllins et al. key metabolites varied in different selected oragns. Three modules were identified by metabolite and gene weighted co-expression network analysis, from which 41 candidate CYPs and 47 candidate UGTs were identified. Combined above information, we narrowed the candidate range of OSC, CYP and UGT genes to 2, 15 and 20. Beside three previous characterized CYPs, we also identified the OSC involved in the synthesis of diosgenin and the glycosyltransferase at the C-3 position of diosgenin for the first time in *P. polyphylla*. This study provides a new idea for the study of gene cluster deficiency biosynthesis pathways in medicinal plants.

## 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", etc. (Tang et al., 2004),

which has pharmacological activities such as hemostasis, analgesia, sedation, anti-inflammatory and anti-tumor effect (Wang et al., 2007, Guo et al., 2008, Qin et al., 2012). The main active component of this plant is the steroids saponin, also known as polyphyllin, accounting for about 80% of the total number of active compounds (Negi et al., 2014). Polyphyllin have aroused great interest for their rich pharmacological activities, including anti-inflammatory, vascular protection, hypoglycaemic, immunomodulatory, antiparasitic, hypocholesterolemic, antifungal, anti-parasitic and anti-tumor effect (Yin et al., 2018, Shuli et al., 2011, Qin et al., 2012). However, the species is at risk of extinction due to slow growth and excessive exploitation (Patel et al., 2013). Due to their complex molecular structure, chemical synthesis of polyphyllin is unlikely to be commercially viable. Therefore, metabolic engineering might be an effective method to provide a stable source of polyphyllins. However, the biosynthetic pathway of polyphyllins still has not been fully elucidated.

Polyphyllins are a group of products with different sugar chains connected at the C3 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) (Patel et al., 2013). However, the sources of diosgenin mainly depend on the extraction from some specified plants, such as yam (Dioscorea genus) and fenugreek (Trigonella foenum-graecum). 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) (Thimmappa et al., 2014). Two FPP molecules were catalyzed by squalene synthase (SQS) to produce linear C30 squalene molecules, squalene, which was further cycled by squalene epoxidase (SQE) to 2, 3-oxidosqualene (Thimmappa et al., 2014).

Then, 2, 3-oxidosqualene is cycled by cycloartenol synthase (CAS) to form cycloartenol, which is then modified by a series of oxidation and reduction to form cholesterol (Cardenas et al., 2015, Lu et al., 2014). In superfamily can catalyze the hydroxylation of cholesterol C-16 and C-22 with the closure of E ring. Then, 16,22 (S) -dihydroxycholesterol further hydroxylates C-26 and forms F ring to produce diosgenin under the action of cytochrome P450s (CYPs) such as PpCYP94D108 (Christ et al., 2019). However, it is still unknown how steroidal skeleton α-hydroxylates at C-17 to form pennogenin. Subsequently, diosgenin and pennogenin were glycosylated by UDP glucuronosyltransferases (UGT) to form various polyphyllins (Figure 1). Until now, the UGTs related to polyphyllins biosynthesis are still not be selected and functionally identified.

Although genomic and transcriptional information of many medicine plants have been generated and available to public, the progress of candidate gene mining and whole pathway desection of plant specialized metabolites are still made slowly due to following factors (Morozova et al., 2009) (Yin et al., 2018, Liu et al., 2016, Yang et al., 2019, Christ et al., 2019) (Xu et al., 2015). The biosynthesis of some specified metabolites lacks tissue specificity, such as the ginsenosides in *Araliaceae* family. A variety of ginsenosides are widely found in multiple tissue parts of the plant, and the complex distribution pattern of ginsenoside components hindering prediction of exact genes involved in biosynthetic pathway application by simple differentially expression analysis (Han et al., 2010). The divergent evolution of CYPs and UGTs families generated many individual members that are phylogenetic closely can decorate diverse type of natural products, respectively. For this reason, it is impossible to predict pathway related gene solely based on phylogenetic analysis (Rai et al., 2017). In addition, the biosynthetic genes of plant specified metabolites which are scattered in different

regions of genome furthermore increased the difficult in identifying candidates precisely by physical distance of metabolic pathway related genes (Mylona et al., 2008, Shang et al., 2014), even if whole genome sequences are generated and well assembled (Xu et al., 2017, Jiang et al., 2021). Therefore, an efficient strategy that can accurately predict the key genes in the complex none-clustered biosynthetic pathway of plant specialized metabolites need to be developed and improved urgently.

In this study, single-molecule based full-length transcriptome sequencing and paired-end based RNA Seq were performed on 8 different tissues from *P. polyphylla var. yunnanensis*. The Weighted Gene Co-Expression Network Analysis (WGCNA) combining the distributions of specified metabolites and the phylogenetic analysis were further predicted the key genes 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 polyphyllin of *P. polyphylla* var. *yunnanensis*, thus promoting the production of those important chemicals by strategy of synthetic biology.

## METHODS

### Plant material and RNA preparation

Samples of four-year dwarf *P. polyphylla* var. *yunnanensis* were collected from Dali, Yunnan, China. The rhizoma, fibrous roots, stems, leaves, ripe fruits, stigma, petals and anthers were harvested in 5 years old healthy plants. The rhizomes, fibrous roots, stems, leaves, and ripe fruits were harvested in October 2018, and the stigma, petals, and anthers were harvested in April 2019. Freeze all tissues in liquid nitrogen immediately and stored at − 80 ℃ after collection. Every sample had three biological replicates that were sequenced independently.

**RNA isolation, transcriptome sequencing and gene function annotation**

The sequencing samples of rhizoma, fibrous roots, stems, leaves, ripe fruits, stigma, petals and anthers 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) according to 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). Single tissue was sequenced using an Illumina Hiseq 2000 platform, and full-length transcriptome sequencing for mixture of different tissues was performed using PacBio platform. The PacBio long reads were filtered and removed redundant sequences 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 FPKM values using DESeq2. The filtered full-length transcripts were functionally annotated using NR, SWISSPROT, GO, COG, KOG, PFAM, and KEGG databases, respectivel.

**Phylogenetic Analysis**

Transcripts belong to OSCs, CYPs and UDP-glycosyltransferases (UGTs) were identified using BLAST software. The transcripts with the length under than 1000 bp were removed. The OSC sequences from different plants used to construct the phylogenetic tree are listed in Table S1. The phylogenetic trees of OSCs, P450s and UGTs were respectively constructed using maximum likelihood method and Jones - Taylor - Thornton (JTT) model using MEGA 7.0 (Kumar et al., 2016b). A bootstrap resampling analysis with 1000 replicates was performed to evaluate the topology of the phylogeny.

**Metabolite content determination**

The tissues frozen at -80 ˚C was lyophilized in a freeze dryer. 20 mg of dry materials was weighted and placed in a 2 mL centrifuge tube. 1 mL of 80 % methanol containing 20 μg of internal standard was added into sample (digitoxin, ≥95%, sigma). The sample was extracted at 1400 rpm for 2 hours, and centrifuged at 10,000 rpm for 5 minutes. Transferred the supernatant to a new centrifuge tube and add 300 μL of n-hexane for extraction. After extraction, the n-hexane layer was sucked and removed, repeated the above. SpinVac was applied for solvent removal of the sample. Redissolved the sample with 500ml of distilled water, then performed extraction twice with 500 mL of n-butanol. Nitrogen was used to dry the organic phase of the sample, then redissolved sample with the mobile phase when measuring.

The analysis was performed on a Waters ACQUITY Ultra Performance Liquid Chromatography (UPLC) system coupled with an AB Sciex 5500 The quadropule ion trap (Qtrap) mass spectrometer (AB Sciex, Milford, MA, USA). Chromatographic separation was achieved on a ACQUITY BEH C18 column (100 × 2.1 mm, 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; flow rate was 0.25 mL $min^{-1}$, Injection volume 5 L. The ESI source interface operated in negative ionization modes were used in this study. The ion spray voltage was set at −4500 V. Table S2 show the optimized multiple reaction monitoring parameters for the analytes and IS. Cholesterol detection is completed on Thermo ISQ-LT Gas Chromatography-Mass Spectrometry (GC-MS), using Thermo TG-5HT column (30m × 0.28mm × 0.10μm). The mass detector is set to SCAN mode, the scanning range is 60-800 m/z, and solvent delay is 10 minutes. The temperature of the injection port is set to 250 □, and the temperature cycle is the initial injection temperature of 170 □ for 2 minutes, 170 to 290 □ for 6 □ per minute, and after reaching 290 □, it stays for 4 minutes, and 290 to 340 □ for 10 □ per minute.

**Construction of gene co-expression networks**

Gene co-expression networks were constructed using the Weighted Gene Co-Expression Network Analysis (WGCNA) approach with R packages (version 3.2.2) (Langfelder and Horvath, 2008). We selected the expression matrix of 31,937 genes with the sum Fragments Per Kilobase per Million (FPKM) value in all tissue greater than 1.0 from all genes as the input file for WGCNA analysis to identify gene modules with strong co-expression. Before the construction of 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). The power β was chosen 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 topological overlap matrix (TOM) among all genes in the analysis. Light color represents low overlap and progressively darker red color represents higher overlap. Blocks of darker colors along the diagonal are the modules, and there is a very strong association between the genes that are contained within these red modules. These red modules will be the focus of our genetic prediction.

**Transient Expression of OSCs in *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, 220 rpm. Then 1 mL culture was used to inoculate 10

mL Luria-Bertani medium containing 50 μg $mL^{-1}$ kanamycin, 25 μg $mL^{-1}$ rifampicin and 25 μg $mL^{-1}$ gentamicin for overnight growth. The following day, the cultures were centrifuged (4000 g, 5 min) and cells were resuspended in infiltration buffer (10 mM MES, 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 following: a 5 mL needle-free syringe was used to gently push the bacterial mixture into the abaxial surface until the entire leaf is filled with agrobacterium. Infiltrated leaves were cultured at 22 □, exposed to light for 10 hs a day, and harvested 6 d after infiltration. For metabolites extraction, leaf discs 1 cm in diameter were prepared from Agrobacterium-infiltrated *N. benthamiana* and dried with a vacumm freeze-dryer. Then the leaves were ground into powder, 20 mg powder was weighed and put into a 2 mL tube for use. Added 2 μL lysate to each sample and heated in the water bath (75 $^{o}$C for 1 hour). After the samples were completely dried, 500 μL ethyl acetate and 500 μL water were added and mixed it with vortex shock and centrifuged for 10 minutes to facilitate separation. 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 30 μL of derivating reagent was added. After vortex mixing, the mixture was heated at 70 $^{o}$C for 30 min, and the mixture was analyzed with GC-MS same as cholesterol analysis above.

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

The total RNA from the *P. polyphylla* var. *yunnanensis* was extracted and reverse transcribed to obtain cDNA. The candidate PCR primers for UGTs were designed according to the transcriptome sequence. The PCR procedure was as follows: 95°C 3 min; 95°C 30 s, 60°C 30 s, 72°C 90 s, 33 cycles and 72°C 5 min. The primers of verified UGT genes are shown in Table S3. 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 OD600 = 0.6. 0.2 mM IPTG was added to the culture medium, induced at 16°C for 16 hours, and centrifuged at 4°C at 5,000 rpm to collect the bacteria. The bacterial cells were suspended in 100 mM phosphate buffer (pH 8.0), and the cells were disrupted by ultrasound in an ice bath. Then the cells were centrifuged at 12,000 rpm at 4°C for 20 min. The bacterial supernatant was purified using Glutathione Beads (Smart-Lifesciences) and concentrated using Millipore ultrafiltration tubes (Meck). Pierce BCA Protein Assay Kit (Thermo) was used to quantify the target protein.

**Enzyme activity analysis**

The enzymatic reaction system consisted of 50 mM Tris (pH 8.0), 1 mM MgCl2, 5 mM glucose donor (UDP-glucose), 1 mM glucose receptor (diosgenin/pennogenin) and purified enzyme of transcript/44939 in a final volume of 100 μL. After overnight incubation at 37°C, equal volume of ice methanol was added to stop the reaction. The product was concentrated and dried, dissolved in 100 μL chromatographic methanol, centrifuged at 12,000 rpm for 10 min, and the supernatant was taken for testing. The reaction products were identified by HPLC and LC-TOF-MS, and the Thermo Hypersil GOLD C18 column (250 mm × 4.6 mm, 5 μm) was used for High Performance Liquid Chromatography (HPLC) detection. The mobile phases were water (A) and acetonitrile (B). Elution gradient: 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 is 1 mL/min, the column temperature is 30°C, the injection volume is 10 μL, and the detection wavelength is 210 nm. Liquid Chromatography Time-Of-Flight Mass Spectrometry LC-TOF-MS was determined using the AB Sciex Tripletof 6600 (AB Sciex, Milford, MA, USA) in a positive ionization mode.

**Results**

***Transcriptome sequencing and assembly***

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 petal was obtained by Illumina sequencing (Table S4). A total of 81.81Gb

clean data containing 1,121,119 CCS reads was obtained from PacBio sequencing platform. Among them, 969,450 long-reads belong to the full length non-chimeric sequence. And 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 (Seppey et al., 2019) to evaluate the integrity of the transcriptome, the results showed complete, single-copy duplicated transcript sequences account for 69.92%, fragmented account for 5.97% and missing account for 26.11%.

Among the final transcripts, 38, 353 (91.68%) transcripts were annotated against the public databases including NR, SWISSPROT, eggNOG, KOG, and PFAM. Furthermore, we enriched the transcripts into GO, COG, and KEGG databases. Among them, 25, 032 (62.78%) annotated transcripts were assigned GO terms, including cellular component, molecular function and biological process. A total of 17, 249 (43.26%) transcripts were functionally predicted and classified using COG database, and 394 transcripts were predicted to be involved in biosynthesis, transport and catabolism of secondary metabolites. KEGG analyses showed that 17, 265 (43.30%) transcripts could map onto 127 pathways (Figure S3). The largest five groups were most associated with carbon metabolism (827), Biosynthesis of amino acids (762), Protein processing in endoplasmic reticulum (686), Ribosome (585), Starch and sucrose metabolism (581).

***Transcriptome functional annotation***

First, we compared the known data in the nr database, annotating 38,177 (95.74%) unigenes from 39,875 transcripts. After that, 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.

GO terms were assigned to 25,032 (62.78%) annotated unigenes, that belonged to three GO categories, cellular component, molecular function and biological process. Among the cellular components, cell, membran, organelle and cell part are the main components; catalyst and binding activity dominate the molecular function, and in the biological process, the metabolic process, cellular process and single-organism process were most prevalent (Figure S1).

Based on the comparison of the COG database, a total of 17,249 (43.26%) unigenes were functionally predicted and classified. The category of transcription (2444) and replication, recombination and repair (2408) were dominant, followed by signal transduction mechanisms (2247) and posttranslational modification, protein turnover, chaperones (1940) (Figure S2). 394 unigenes were related to the biosynthesis, transport and catabolism of secondary metabolites. Among them, 394 genes were predicted to be involved in biosynthesis, transport and catabolism of secondary metabolites.

KEGG analyses showed that 17,265 (43.30%) unigenes mapped onto 127 pathways (Figure S3). The largest five groups were most associated with carbon metabolism (827), Biosynthesis of amino acids (762), Protein processing in endoplasmic reticulum (686), Ribosome (585), Starch and sucrose metabolism (581).

***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 shows that different branches clearly distinguish the classification of OSCs genes. The different OSC subfamilies are distributed in the 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 cycloartenol synthase evolutionary branch (Figure S4).

Phylogenetic analysis of CYPs using Arabidopsis as a reference indicates all the full-length P450s from *P. polyphylla* var. *yunnanensis* could 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 P450s genes, followed by the CYP85 (37) and CYP72 (21) families, CYP710 (4), CYP97(4), CYP74 (2), CYP711 (2) and CYP51 (1) has the smallest number of P450s genes (Figure 2A and S5).

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 UGTs protein sequences of *P. polyphylla* var. *yunnanensis* were clustered into 13 of the 21 known UGT subfamilies (Wilson and Tian, 2019). 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 which account for 28.64 % of all UGTs (Figure 2A and S6).

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

The WGCNA package in R software was used to construct a weighted gene co-expression network. In order to make the network conform to the scale-free network distribution, the function Pick Soft Threshold in the WGCNA package is used to calculate the weight value. According to the results in Figure S7, a soft threshold $\beta = 9$ is selected to build a co-expression network. Then use the function hclust to perform hierarchical clustering on the dissimilar matrices, and use Dynamic Tree Cut to cut the generated cluster tree (Figure S8). In this process, unigenes with similar expression patterns could be combined on the same branch, and each branch represented a co-expression module, and different colors represented different modules. Differential unigenes were correlated and clustered according to their FPKM values. Unigenes with higher correlation were assigned to the same module (Figure S9). In the end, 31,937 unigenes were divided into 26 modules, and the number of unigenes in the modules was from 53-7667.

Using LC-MS-MS, we determined the expression 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 2C). The

metabolites involved in the metabolic pathway with diosgenin as the substrate (diosgenin, prosapogenin A, polyphyllin I and polyphyllin II) have relatively similar expression profiles, and are highly expressed in rhizoma, leaf, ovary and petal. Different from them, trillin has a higher content in leaves. The expression levels of polyphyllin VI and polyphyllin VII synthesized from pennogenin were also similar, polyphyllin VI was highly expressed in rhizoma, fibril, fruit and ovary, while polyphyllin VII was significantly expressed in fibril, fruit, ovary and petal.

The expression profile data of metabolites and WGCAN data are integrated to construct a co-expression network of genes and metabolites (figure S10). The network can combine gene modules and metabolites with similar expression in various tissues of *P. polyphylla* var. *yunnanensis,* and calculate the correlation coefficient between metabolites and gene modules (Figure 2B). When the correlation coefficient approaches 1, it means that the metabolites and the genes in the module show similar expression levels in different tissues of *P. polyphylla* var. *yunnanensis.*

According to correlation coefficient, it can be clearly seen that polyphyllin I has the highest correlation with modules Coral (0.72) and orangered4 (0.7); polyphyllin II has the highest correlation with modules lavenderblush3 (0.64), and polyphyllin VI and prosapogenin A do not show specific correlation with a certain module; Polyphyllin VII showed good correlation with antiquewhite4 (0.81) and antiquewhite1 (0.76); Trillin clustered well with lavenderblush3 (0.82), lightblue3 (0.72) and white (0.73) in the co-expression network; And diosgenin and cholesterol were co-expressed with antiquewhite4, the correlation coefficients are 0.66 and 0.65 respectively . The clustering results indicate that the genes in the orangered4, lavenderblush3, lightblue3, coral, antiquewhite1, white and antiquewhite4 modules are likely to be involved in the biosynthesis of polyphyllin. Furthermore, we analyzed the upstream genes involved in the metabolic pathways of cholesterol and other phytosterols in these 7 modules, and the results showed that coral, antiquewhite4 and lavenderblush3 contained relatively complete cholesterol synthesis genes, with 14, 21 and 11 respectively (Table S5).

However, the orange, lightblue3, white and antiquewhite1 modules did not contain enough cholesterol synthesis genes, and the number of upstream genes was 0, 0, 4 and 6, respectively. Based on the results, the genes contained in the three modules of coral, antiquewhite4, and lavenderblush3 are likely to play a key role in the synthesis of polyphyllins, and subsequent gene prediction will be developed around these three modules.

From the three predicted modules, we obtained a total of 42 CYPs and 48 UGTs genes. Heat maps were plotted using information about the expression of these genes in different tissues, as shown in Figure 3A. In the heat map, the redder the color of the module, the higher the gene expression in the corresponding tissue. Genes that can perform important functions in plants often require sufficient expression levels. Therefore, among the genes that have been predicted, the higher the expression level, the more likely it is to participate in the synthesis of polyphyllin.

At the same time, we predict the possible CYP and UGT genes from the phylogenetic tree. According to previous reports, the P450 genes that can be hydroxylated in triterpene or steroidal skeletons often belong to the CYP72, CYP90 and CYP94 families. In the phylogenetic tree, we have obtained 21, 11 and 28 genes from the CYP72, CYP90 and CYP94 families. As for UGT genes, the genes that can add a glycosyl at position C-3 are always in the UGT73 family. In our transcriptome results, there are a total of 57 genes belonging to this family. The genes obtained by the phylogenetic tree and WGCNA were compared, and the CYPs and UGTs predicted by the two methods were 15 and 20, respectively (Figure 3B). These common genes are very likely to be involved in the biosynthesis of polyphyllin. Those genes that appear in the phylogenetic tree and are not included in the WGCAN should not be ignored. They may also include some genes involved in the biosynthesis of polyphyllin.

***Functional characterization of 2,3-oxidosqualene cyclases***

OSC is one of the key enzymes in steroid biosynthesis [12,13], and OSC catalyzed conversion of 2,3-oxidesqualene to cyclic triterpenes marks the first scaffold diversification reaction in the triterpene pathway. In contrast to single copy OSC in lower plants, multiple

OSCs are encoded in the higher plant genomes [86]. A total of 13 OSC-related transcriptome variants were found in the transcriptome, which may represent 2 non-redundant OSCs (PpOSC1 and PpOSC2). In order to investigate the activity of these OSC candidates in the cholesterol transformation reaction, we cloned 2 genes from cDNA prepared from RNA of mixed samples of plants, and successfully transferred these 2 genes into *A.tumefaciens* and infected *N.benthamiana*. After infection, we collected the leaves and verified the function of OSC gene by measuring the content of cycloartenol. The results showed that the instantaneous expression of PpOSC1 could facilitate the production of 3.12 fold more cycloartenol, while PpOSC2 could not increase the content of cycloartenol, instead of an uncharacterized product (Figure 3C). Therefore, the PpOSC1 gene plays an important function in the biosynthesis of polyphyllin.

***Identification of glycotransferase at c-3 position of diosgenin***

In the candidate UGTs, transcript/44939 showed the activity of C-3 glucosyltransferase, named PpUGT73E5. In the phylogenetic tree, PpUGT73E5 is in the same branch with *Bv*UGT73C10 and *Bv*UGT73C10, which can glycosylate the C-3 position of β-amyrin(Augustin et al., 2012). The CDS sequence length of the gene was 1398bp, encoding 465 amino acids, and the molecular weight of the protein was 51.14kDa. When diosgenin is used as a sugar acceptor, the enzyme encoded by PpUGT73E5 can catalyze diosgenin to produce a more polar product after adding UDP-glucose. The HPLC retention time of product is 22.1 min, which is consistent with trillin. When pennogenin is used as the sugar acceptor and UDP-glucose is the sugar donor, PpUGT73E5 catalyzes pennogenin to produce a new product peak at 18.9 min. The LC-TOF-MS detection result shows that the molecular weight of the new product is 593.37 $[M+H]^+$ , Which is consistent with the molecular weight of pennogenin-3-O-glucoside (Figure 3D).

**DISSCUSSION**

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 not common in plants, and gene redundancy and strict genetic regulation in plants make it more difficult to parse metabolic pathways (Karaiskos et al., 2017). At present, botanists mainly use multi-omics (genome, transcriptome, and metabolome) methods to analyze the biological pathways of metabolites. However, in the absence of genomic data, this prediction method often has a large number of false positive results. For the majority of medicinal plants , a more accurate method is needed to predict metabolite biosynthesis pathways without genomic data.

The research on the biosynthesis of polyphyllin has been a hot topic because of its various pharmacological activities. Due to the huge genome, there is currently no data report on the genome of *P. polyphylla* var. *yunnanensis*. In order to better study the biosynthesis pathway of polyphyllin, transcriptome sequencing is the most suitable method at this stage.

RNA-Seq technology based on RNA fragment splicing often cannot accurately obtain or assemble complete transcripts, and cannot identify isoform, homologous genes, superfamily genes and alleles genes, which makes it difficult to understand the deeper meaning of this life activity (Ardui et al., 2018, Zhao et al., 2019). Full-length transcriptome sequencing based on PacBio SMRT single-molecule real-time sequencing technology does not need to interrupt RNA, and directly reverse-transcribes to obtain full-length cDNA without subsequent assembly, which greatly improves the quality of transcriptome sequencing and obtains more accurate transcriptome information (Zhao et al., 2019, Ardui et al., 2018).

At present, there have been some reports on the transcriptome data of *P. polyphylla* (Yang et al., 2019, Yin et al., 2018, Liu et al., 2016) (Qi et al., 2013). However, these transcriptome measurements were all based on Illumina platform and could not better reflect the complete transcriptome information of *P. polyphylla var. yunnanensis*, published transcriptome information is mainly derived from roots, stems, and leaves of the plant, with little transcriptome information from other tissues.

In this study, we collected samples from 8 tissues of *P. polyphylla var. yunnanensis* to

complete the splicing and full-length transcriptome sequencing, a total of more than 370 G of clean data was obtained. This is the transcriptome data that contains the most tissues and the deepest sequencing in *P. polyphylla var. yunnanensis*. Compared with previous reports of sequencing, it can avoid a lot of redundant data and splicing errors, and provides data support for more accurate prediction of the biosynthetic pathway of polyphyllin.

Weighted gene co-expression network analysis (WGCNA) is an important method for studying gene functions through the network (Yang et al., 2018a, Pei et al., 2017). Genes in the same module and several modules with strong correlation in the gene co-expression network may have similar biological functions. Therefore, using WGCNA method and RNA-Seq data can better excavate modules related to specific biological functions. And through the analysis of hub gene to find genes related to biosynthesis and predict gene function. Using the WGCAN method, Yan Yang et al. predicted key genes and pathways related to male sterility of eggplant (Yang et al., 2018b), and Yuling Tai et al. investigated the regulation of the biosynthetic pathways of catechins, theanine, and caffeine in the tea plant (Tai et al., 2018).

In this article, we tried to construct 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, polyphyllin I, II, VI and VII (Figure 4). The results of WGCAN construction are very satisfactory. We have predicted three modules that are very relevant to metabolites. We further conducted conditional screening through phylogenetic trees and gene expression levels, and obtained some very good alternative genes. In candidate genes, we identified the OSC gene and a UGT gene with C-3 glucosyltransferase function from *P. polyphylla* var. *yunnanensis* for the first time. The predicted CYP genes included those that have been reported to play an important role in the synthesis pathway of polyphyllin. These results show the accuracy and reliability of this prediction method.

At present, the analysis of steroidal saponin biosynthetic pathways is progressing slowly

and is still 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.(Babiychuk et al., 2008, Kumar et al., 2016a), and several P450s. There are also few reports on glycosylation modification of diosgenin. Therefore, it is possible to predict key genes involved in the synthesis pathway of polyphyllin by using our prediction method combining the evolutionary tree, co-expression network and gene expression quantity, and 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 not complete. We performed splicing and full-length transcriptome sequencing of rhizoma, fibrous roots, stems, leaves, ripe fruits, stigma, petals and anthers 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 saponin. Finally, we confirmed that PpOSC1 can catalyze the formation of cycloartenol from 2, 3-oxidosqualene in *N. benthamiana*. To our knowledge, this study measured the full-length transcriptome and characterized the function of OSC and C-3 glycosyltransferase of *Paris polyphylla* var. *yunnanensis* for the first time. This study improves our understanding of the biosynthetic pathways of polyphyllin, making a basis for further elucidating the pharmacologically active triterpene / saponin biosynthesis, and also providing an efficient strategy to study complex pathway of other plant specialized metabolites .

## Acknowledgements

This work was supported by the Opening Project of Zhejiang Provincial Preponderant and Characteristic Subject of Key University (Traditional Chinese Pharmacology), and the Zhejiang Chinese Medical University (No. ZYAOX2018012), and the National Natural Science Foundation of China (NSFC Grant No.31770332, 31970314).

## 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 wrote and edited most of the manuscript. All authors have read and approved the final manuscript.

## Data availability

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

## Conflict of interest

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

## References

ARDUI, S., AMEUR, A., VERMEESCH, J. R. & HESTAND, M. S. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res, 46, 2159-2168.

AUGUSTIN, J. M., DROK, S., SHINODA, T., SANMIYA, K., NIELSEN, J. K., KHAKIMOV, B., OLSEN, C. E., HANSEN, E. H., KUZINA, V., EKSTROM, C. T., HAUSER, T. & BAK, S. 2012. UDP-Glycosyltransferases from the UGT73C Subfamily in Barbarea vulgaris Catalyze Sapogenin 3-O-Glucosylation in Saponin-Mediated Insect Resistance. Plant Physiology, 160, 1881-1895.

BABIYCHUK, E., BOUVIER-NAVE, P., COMPAGNON, V., SUZUKI, M., MURANAKA, T., VAN MONTAGU, M., KUSHNIR, S. & SCHALLER, H. 2008. Albinism and cell viability in cycloartenol synthase deficient Arabidopsis. Plant Signal Behav, 3, 978-80.

CARDENAS, P. D., SONAWANE, P. D., HEINIG, U., BOCOBZA, S. E., BURDMAN, S. & AHARONI, A. 2015. The bitter side of the nightshades: Genomics drives discovery in Solanaceae steroidal alkaloid metabolism. Phytochemistry, 113, 24-32.

CHRIST, B., XU, C., XU, M., LI, F. S., WADA, N., MITCHELL, A. J., HAN, X. L., WEN, M. L., FUJITA, M. & WENG, J. K. 2019. Repeated evolution of cytochrome P450-mediated spiroketal steroid biosynthesis in plants. Nat Commun, 10, 3206.

GUO, L., SU, J., DENG, B. W., YU, Z. Y., KANG, L. P., ZHAO, Z. H., SHAN, Y. J., CHEN, J. P., MA, B. P. & CONG, Y. W. 2008. 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. Hum Reprod, 23, 964-71.

HAN, J. Y., IN, J. G., KWON, Y. S. & CHOI, Y. E. 2010. Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in Panax ginseng. Phytochemistry, 71, 36-46.

JIANG, Z., TU, L., YANG, W., ZHANG, Y., HU, T., MA, B., LU, Y., CUI, X., GAO, J., WU, X., TONG, Y., ZHOU, J., SONG, Y., LIU, Y., LIU, N., HUANG, L. & GAO, W. 2021. The chromosome-level reference genome assembly for Panax notoginseng and insights into ginsenoside biosynthesis. Plant Commun, 2, 100113.

KARAISKOS, I., SOULI, M., GALANI, I. & GIAMARELLOU, H. 2017. Colistin: still a lifesaver for the 21st century? Expert Opin Drug Metab Toxicol, 13, 59-71.

KUMAR, S., KALRA, S., SINGH, B., KUMAR, A., KAUR, J. & SINGH, K. 2016a. RNA-Seq mediated root transcriptome analysis of Chlorophytum borivilianum for identification of genes involved in saponin biosynthesis. Funct Integr Genomics, 16, 37-55.

KUMAR, S., STECHER, G. & TAMURA, K. 2016b. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol, 33, 1870-4.

LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

LIU, T., LI, X., XIE, S., WANG, L. & YANG, S. 2016. RNA-seq analysis of Paris polyphylla var. yunnanensis roots identified candidate genes for saponin synthesis. Plant Divers, 38, 163-170.

LU, Y., ZHOU, W., WEI, L., LI, J., JIA, J., LI, F., SMITH, S. M. & XU, J. 2014. Regulation of the cholesterol biosynthetic pathway and its integration with fatty acid biosynthesis in the oleaginous microalga Nannochloropsis oceanica. Biotechnol Biofuels, 7, 81.

MOROZOVA, O., HIRST, M. & MARRA, M. A. 2009. Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet, 10, 135-51.

MYLONA, P., OWATWORAKIT, A., PAPADOPOULOU, K., JENNER, H., QIN, B., FINDLAY, K., HILL, L., QI, X., BAKHT, S., MELTON, R. & OSBOURN, A. 2008. Sad3 and sad4 are required for saponin biosynthesis and root development in oat. Plant Cell, 20, 201-12.

NEGI, J. S., BISHT, V. K., BHANDARI, A. K., BHATT, V. P., SINGH, P. & SINGH, N. 2014. Paris polyphylla: chemical and biological prospectives. Anticancer Agents Med Chem, 14, 833-9.

PATEL, K., GADEWAR, M., TAHILYANI, V. & PATEL, D. K. 2013. A review on pharmacological and analytical aspects of diosmetin: a concise report. Chin J Integr Med, 19, 792-800.

PEI, G., CHEN, L. & ZHANG, W. 2017. WGCNA Application to Proteomic and Metabolomic Data Analysis. Methods Enzymol, 585, 135-158.

QI, J., ZHENG, N., ZHANG, B., SUN, P., HU, S., XU, W., MA, Q., ZHAO, T., ZHOU, L., QIN, M. & LI, X. 2013. Mining genes involved in the stratification of Paris polyphylla seeds using high-throughput embryo transcriptome sequencing. BMC Genomics, 14, 358.

QIN, X. J., SUN, D. J., NI, W., CHEN, C. X., HUA, Y., HE, L. & LIU, H. Y. 2012. Steroidal saponins with antimicrobial activity from stems and leaves of Paris polyphylla var. yunnanensis. Steroids, 77, 1242-1248.

RAI, A., SAITO, K. & YAMAZAKI, M. 2017. Integrated omics analysis of specialized metabolism in medicinal plants. Plant J, 90, 764-787.

SEPPEY, M., MANNI, M. & ZDOBNOV, E. M. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol, 1962, 227-245.

SHANG, Y., MA, Y., ZHOU, Y., ZHANG, H., DUAN, L., CHEN, H., ZENG, J., ZHOU, Q., WANG, S., GU, W., LIU, M., REN, J., GU, X., ZHANG, S., WANG, Y., YASUKAWA, K., BOUWMEESTER, H. J., QI, X., ZHANG, Z., LUCAS, W. J. & HUANG, S. 2014. Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. Science, 346, 1084-8.

SHULI, M., WENYUAN, G., YANJUN, Z., CHAOYI, M., LIU, Y. & YIWEN, L. 2011. Paridis saponins inhibiting carcinoma growth and metastasis in vitro and in vivo. Arch Pharm Res, 34, 43-50.

TAI, Y., LIU, C., YU, S., YANG, H., SUN, J., GUO, C., HUANG, B., LIU, Z., YUAN, Y., XIA, E., WEI, C. & WAN, X. 2018. Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). BMC Genomics, 19, 616.

TANG, M. J., ZHAO, J., LI, X. H. & YU, S. S. 2004. [Advances in studies on chemical constituents and pharmacological activities from plants of Symplocaceae]. Zhongguo Zhong Yao Za Zhi, 29, 390-4.

THIMMAPPA, R., GEISLER, K., LOUVEAU, T., O'MAILLE, P. & OSBOURN, A. 2014. Triterpene biosynthesis in plants. Annu Rev Plant Biol, 65, 225-57.

WANG, Y., ZHANG, Y. J., GAO, W. Y. & YAN, L. L. 2007. [Anti-tumor constituents from Paris polyphylla var. yunnanensis]. Zhongguo Zhong Yao Za Zhi, 32, 1425-8.

WILSON, A. E. & TIAN, L. 2019. Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism. Plant J, 100, 1273-1288.

XU, J., CHU, Y., LIAO, B., XIAO, S., YIN, Q., BAI, R., SU, H., DONG, L., LI, X., QIAN, J., ZHANG, J., ZHANG, Y., ZHANG, X., WU, M., ZHANG, J., LI, G., ZHANG, L., CHANG, Z., ZHANG, Y., JIA, Z., LIU, Z., AFREH, D., NAHURIRA, R., ZHANG, L., CHENG, R.,

ZHU, Y., ZHU, G., RAO, W., ZHOU, C., QIAO, L., HUANG, Z., CHENG, Y. C. & CHEN, S. 2017. Panax ginseng genome examination for ginsenoside biosynthesis. Gigascience, 6, 1-15.

XU, Z. C., PETERS, R. J., WEIRATHER, J., LUO, H. M., LIAO, B. S., ZHANG, X., ZHU, Y. J., JI, A. J., ZHANG, B., HU, S. N., AU, K. F., SONG, J. Y. & CHEN, S. L. 2015. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. Plant Journal, 82, 951-961.

YANG, T., LI, K., HAO, S., ZHANG, J., SONG, T., TIAN, J. & YAO, Y. 2018a. The Use of RNA Sequencing and Correlation Network Analysis to Study Potential Regulators of Crabapple Leaf Color Transformation. Plant Cell Physiol, 59, 1027-1042.

YANG, Y., BAO, S., ZHOU, X., LIU, J. & ZHUANG, Y. 2018b. The key genes and pathways related to male sterility of eggplant revealed by comparative transcriptome analysis. BMC Plant Biol, 18, 209.

YANG, Z., YANG, L., LIU, C., QIN, X., LIU, H., CHEN, J. & JI, Y. 2019. Transcriptome analyses of Paris polyphylla var. chinensis, Ypsilandra thibetica, and Polygonatum kingianum characterize their steroidal saponin biosynthesis pathway. Fitoterapia, 135, 52-63.

YIN, Y., GAO, L., ZHANG, X. & GAO, W. 2018. A cytochrome P450 monooxygenase responsible for the C-22 hydroxylation step in the Paris polyphylla steroidal saponin biosynthesis pathway. Phytochemistry, 156, 116-123.

ZHAO, L., ZHANG, H., KOHNEN, M. V., PRASAD, K., GU, L. & REDDY, A. S. N. 2019. Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. Front Genet, 10, 253.

## 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 first identified by this method are shown in blue background.

**Figure 2. 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 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 expression profiles of key metabolites involved in the biosynthetic pathway of polyphyllin in different tissues.

**Figrure 3. Candidate 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 asterisk 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) 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*. (D) Functional verification of *Pp*UGT genes. The functional verification of candidate UGTs was performed by HPLC and LC-TOF-MS, and the enzyme encoded by transcript_44939 gene could introduce glucose group into C-3 of diosgenin and pannogenin.

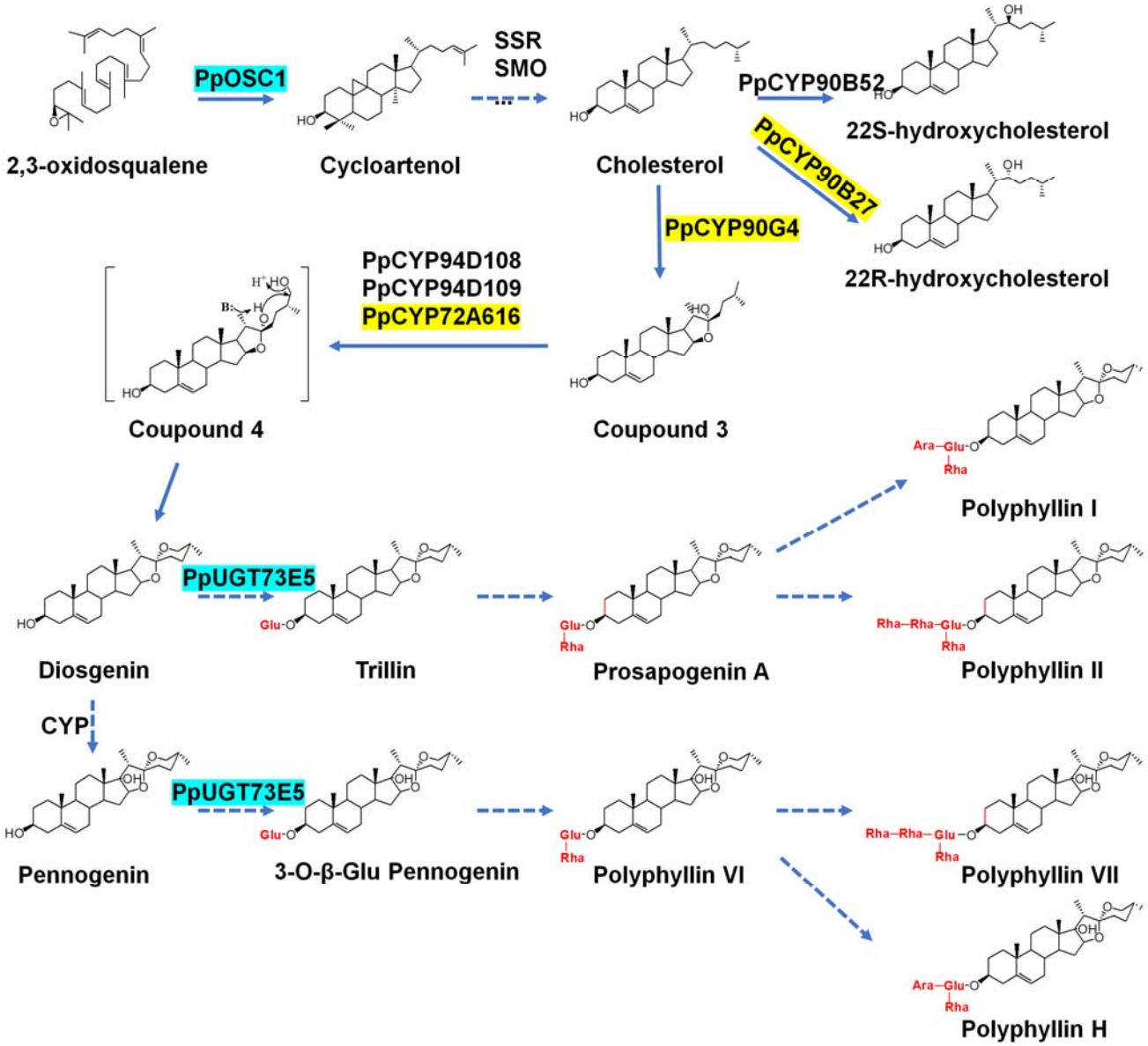


**Figure 1**

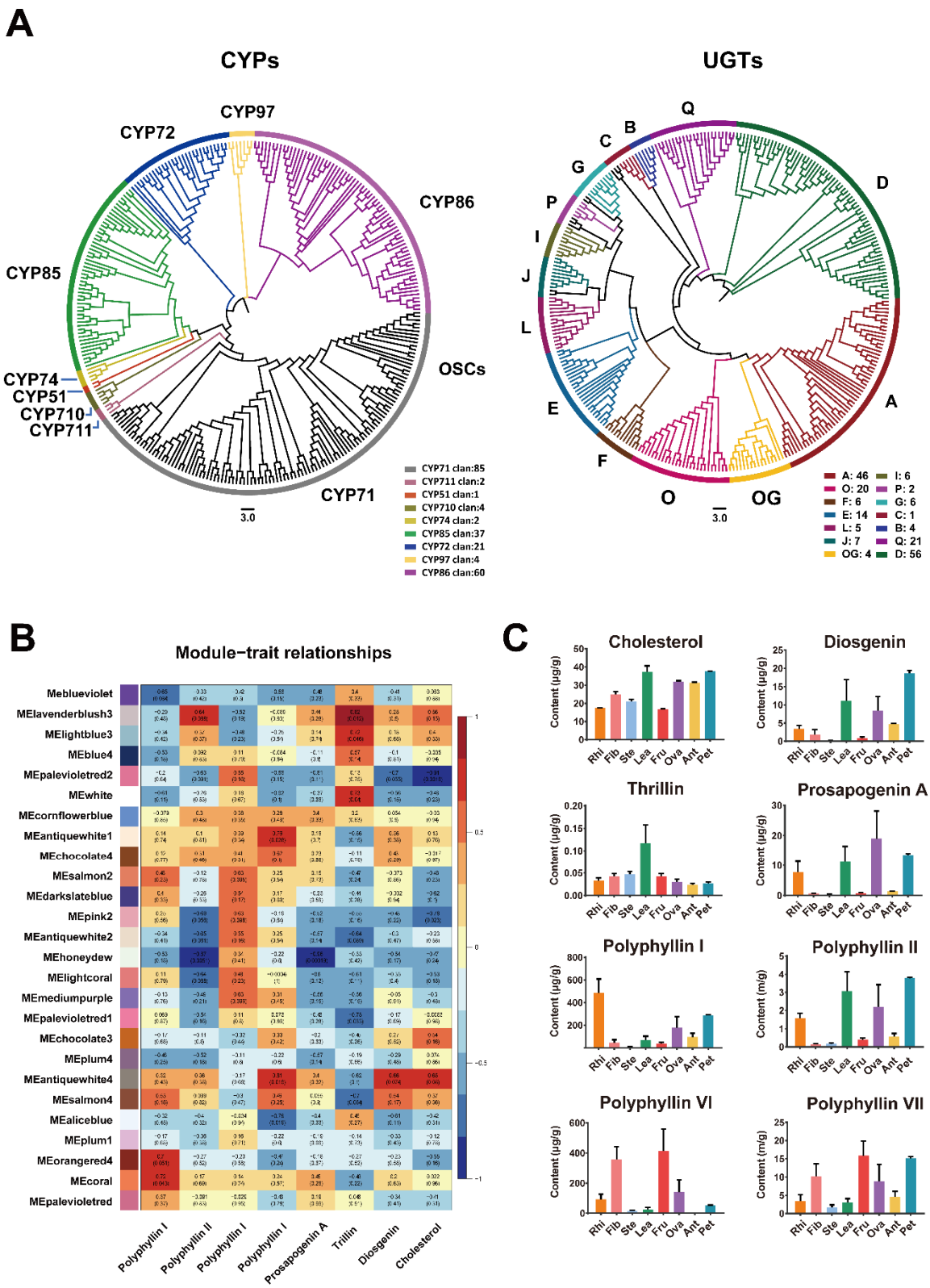


**Figure 2**

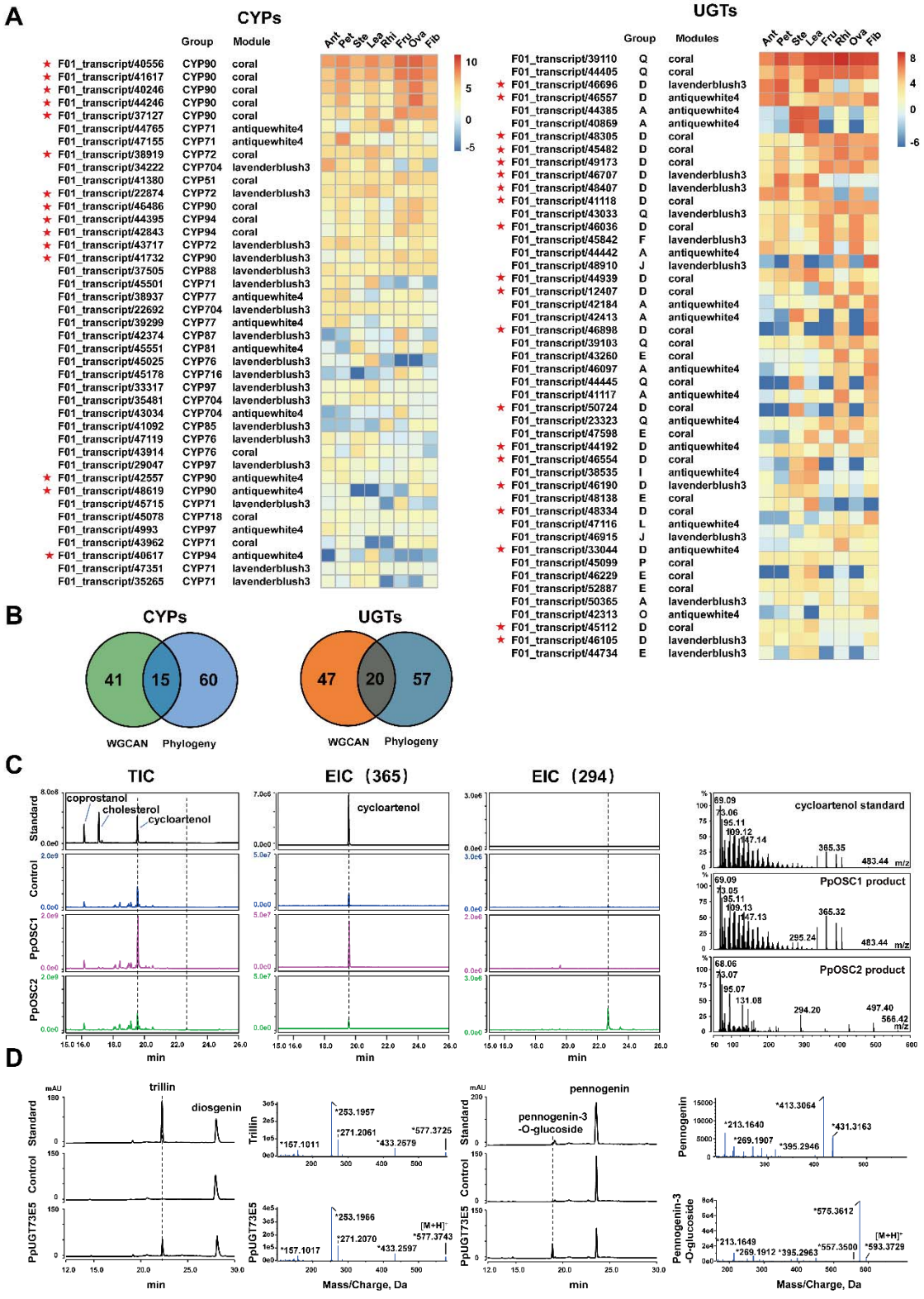


**Figure 3**