

Accumulation mechanism of metabolites markers identified by machine learning between Qingyuan and Xiushui counties in *Polygonatum cyrtoneura* Hua

Qiqi Gong

Zhejiang A&F University

Jianfeng Yu

Zhejiang A&F University

Zhicheng Guo

Ke Fu

Zhejiang A&F University

Yi Xu

Zhejiang A&F University

Hui Zou

Yipuyuan Huangjing Technology Co., Ltd

Cong Li

Zhejiang A&F University

Jinping Si

Zhejiang A&F University

Shengguan Cai

Zhejiang University

Donghong Chen

Zhejiang A&F University

Zhigang Han

hanzg@zafu.edu.cn

Zhejiang A&F University

Research Article

Keywords: *Polygonatum cyrtoneura* Hua, phenylpropanoids, machine learning, metabolomics, biosynthesis

Posted Date: October 12th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3420530/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on March 6th, 2024. See the published version at <https://doi.org/10.1186/s12870-024-04871-6>.

Abstract

Polygonatum cyrtoneura Hua (*P. cyrtoneura*) is well known for its high medicinal value due to a wide variety of secondary metabolites. Nonetheless, the unclearness persists regarding the distribution and buildup processes of these metabolites across various regions. Using UPLC-ESI-MS/MS, a grand total of 482 metabolites were detected and identified in this research. Cluster analysis showed distinct metabolite profiles between Qingyuan County and Xiushui County. The identification of secondary metabolites, such as flavonoids, phenolic acids, and lignans, between the two regions was performed using support vector machine-recursive feature elimination (SVM-RFE) and random forest (RF) machine learning techniques. Subsequently, the process of comparative transcriptomics and weighted gene co-expression analysis (WGCNA) uncovered genes associated with flavonoids such as *CHI*, *UGT1*, *UGT2*, *ERF*, as well as phenylpropane-related genes *UGT3* and *NAC*. In addition, by comparing transcriptomes, four genes (*PcOMT10/11/12/13*) were selected as differentially expressed. Their impact on metabolic fluxes of the phenylpropane pathway was confirmed using a transient expression system in tobacco. The findings enhanced our comprehension of the variation in accumulation of secondary metabolites mediated by phenylpropanoids across various locations, and offered valuable genetic assets for the synthesis of bioactive compounds.

1. Introduction

Polygonatum cyrtoneura Hua (*P. cyrtoneura*) is an ancient and timeless herbal plant. According to recent studies, the rhizome of this plant contains abundant non-starch polysaccharides, oligosaccharides, triterpenoid saponins, steroids, flavonoids, and other beneficial substances. These components have various pharmacological benefits, including the prevention of diabetes in obese individuals, enhancement of insulin secretion, improvement of insulin resistance, antibacterial properties, anti-tumor effects, anti-inflammatory properties, anti-aging benefits, bone protection, and prevention of Alzheimer's disease [1–3]. In China, *P. cyrtoneura* (Fig. 1A) primarily grown in forested areas, this plant is found in the southern part of the region [4]. Nevertheless, the precise composition and number of bioactive substances in *P. cyrtoneura* remain largely undisclosed. Hence, it is significant to detect specific metabolic indicators for differentiating the quality of *P. cyrtoneura* grown in various regions.

Presently, there have been advancements in machine learning methods for the detection of biomarkers or characteristic genes associated with human ailments [5, 6]. The support vector machine-recursive feature elimination (SVM-RFE) is a common multivariate approach in machine learning that is extensively employed for selecting features and classifying high-throughput sequencing data [7]. The random forest (RF) is a classification technique that utilizes decision trees by dividing data, and it is extensively employed in predictive models for human illnesses because of its remarkable precision and comprehensibility [8]. As a result, machine learning has efficiently enabled the extraction of distinctive metabolites from extensive metabolome data.

Plants often contain plant metabolites that originate from the phenylpropanoid pathway, such as coumarins, lignans, flavonoids, phenolic acids, and lignins. Plant growth, development, and responses to environmental stress are influenced by hydroxylated aromatic rings in these secondary metabolites [9]. Flavonoids, which are the predominant polyphenols found in plants, possess various biological activities including antioxidant properties as well as anti-inflammatory, anticancer, anti-allergic, and neuroprotective effects [10]. The dietary health benefits and functional properties of phenolic acids, including antioxidant, immune regulatory, anti-atherosclerotic, anti-thrombotic, cardioprotective, anti-cancer, and anti-diabetic activities, have also been widely recognized and appreciated [11]. According to Jang et al., lignans, which are derived from two molecules of phenylpropanoid derivatives, are natural compounds that can provide protective effects, especially against chronic western diseases, when consumed regularly in foods rich in lignans [12]. Hence, it is imperative to investigate the accumulation process of phenylpropanoid substances in order to enhance the therapeutic worth of *P. cyrtonema*.

The phenylpropanoid pathway typically initiates in plants by synthesizing phenylalanine through the glycolysis and shikimic acid pathways that precede it. Afterwards, PAL, C4H, and 4CL catalyze the reaction in a sequential manner to generate cinnamic acid, hydroxycinnamic acid, and CoA-linked 4-coumarate, respectively. This process was described by Muro-Villanueva [13]. Afterwards, they enter two major downstream pathways, namely the biosynthesis of monophenols and flavonoids. According to D'Auria, lignin production is primarily regulated by hydroxycinnamoyl transferase (HCT), which plays an essential role in controlling the subsequent monophenol metabolism [14]. This is followed by the catalytic activities of cinnamoyl-CoA reductase (CCR) and cinnamyl-alcohol dehydrogenase (CAD). The downstream flavonoid metabolism is believed to be regulated by Chalcone synthase (CHS), which is considered as the pivotal enzyme. After CHS synthesizes the chalcone, it undergoes conversion to flavanone through the action of CHI. Subsequently, a sequence of enzymes including F3H, FLS, and FNS facilitate the generation of dihydroflavonols, quercetin, and apigenin, correspondingly [15]. Despite the extensive knowledge on the phenylpropanoid pathway and certain gene groups in *Arabidopsis*, the comprehension of *P. cyrtonema*'s biosynthesis of phenylpropanoid compounds remains largely elusive.

In recent times, the utilization of weighted gene co-expression network analysis (WGCNA) has become prevalent in uncovering the genetic composition of secondary compounds. This approach combines extensive transcriptome and metabolome analyses, and is considered a robust technique [16]. The WGCNA algorithm has the ability to partition genes into modules by employing a hierarchical clustering technique, which is more intricate compared to the conventional approach of pairwise differential gene expression. This method takes into account the connections between genes by analyzing the pairwise correlations of gene expression profiles [17].

Using a widely targeted metabolome approach, we discovered the metabolites of 27 different wild genotypes of *P. cyrtonema* in two distinct regions during this research. Subsequently, the RF and SVM-RFE machine learning techniques were employed to distinguish the primary characteristic metabolites between Qingyuan and Xiushui counties, which exhibited clear clustering into two distinct groups based on the metabolome data. Afterwards, the biosynthesis mechanism of these distinctive metabolites is

explored through comparative transcriptome analysis and WGCNA analysis. Furthermore, the tobacco transient transformation system was utilized to characterize the biological functions of genes that were expressed differentially. Identifying the biomarkers of metabolites in various areas and uncovering the mechanism of metabolite accumulation will be advantageous for enhancing genetic enhancement and cultivation of superior *Polygonatum* species.

2. Materials and approaches

2.1. Plant materials and sampling

From the southern region of China, a collection of 27 untamed genotypes of *P. cyrtoneura* was acquired. During the year 2017, the rhizomes of *P. cyrtoneura* underwent disinfection using a 1% concentration of carbendazim for a duration of 60 minutes. Subsequently, they were washed with tap water and placed in a pot containing nutrient-rich soil. During the year 2020, three sets of *P. cyrtoneura* rhizome samples were collected from the containers, thoroughly cleansed, and had their fibrous roots eliminated. Collected rhizomes were rapidly frozen in liquid nitrogen and stored at -80°C after drying on filter paper.

2.2. Identification and quantification of metabolites through the application of a widely targeted metabolome approach

Metabolites were extracted, identified, and quantified following the method described by Han et al. (2023). In short, the rhizome samples were freeze-dried and then pulverized using a mixer (Retsch, Haan, Germany). The freeze-dried tissue (0.1 g) was dissolved in extraction buffer and the mixture was incubated in a refrigerator at 4 °C for 12 hours. The centrifugation process was applied to the extraction solution at a speed of 14,000 revolutions per minute for a duration of 8 minutes. Subsequently, the extracted samples underwent analysis utilizing a UPLC-ESI-MS/MS setup. Following UPLC, the effluent was connected in succession to an ESI-triple quadrupole linear ion trap (Q TRAP)-MS.

The triple quadrupole linear ion trap mass spectrometer (QTrap) was operated using Analyst 1.6 software (AB Sciex) to gather scans from the linear ion trap (LIT) and triple quadrupole (QQQ). Polypropylene glycol solutions at concentrations of 10 and 100 mol/L were used for instrument tuning and mass calibration for QQQ and LIT scans. QQQ scans were obtained using the multiple reaction monitoring (MRM) mode, with the collision gas (nitrogen) set at 5 psi. For each MRM transition, the DP and CE were optimized. The eluted metabolites were monitored at each time interval by monitoring a specific MRM transition.

Simca-P software (version 13.0, Umetrics AB, Umea, Sweden) was used to perform orthogonal partial least squares discriminant analysis (OPLS-DA) and unsupervised principal component analysis (PCA) on the processed data. The R software was used to conduct hierarchical clustering analysis of metabolites across samples. To identify metabolites that have accumulated differentially (DAMs), the project utilized

screening criteria such as $\log_2(\text{Fold Change}) \geq 1$, $p\text{value} \leq 0.05$, or variable importance in the projection (VIP) ≥ 1 . Han et al. (2023) provided metabolic data for two wild genotypes (T118_21 and T26_14).

2.3. Analysis of RNA sequencing

The MiniBEST Plant RNA Extraction Kit from the Japanese company TaKaRa was used to extract total RNA. RNA concentration and quality were evaluated with the Agilent 2100 Bioanalyzer. Using the NEB Next Ultra RNA Library Preparation Kit (NEB, E7530) and the NEB Next Multiplex Oligos (NEB, E7500), a total of 45 RNA-seq libraries were created. Following that, paired-end sequencing was carried out using an Illumina HiSeq 2500 platform. The Trinity software was utilized to compute the transcript abundance of genes in transcripts per million (TPM). Genes identified as differentially expressed were those with $|\log_2(\text{fold change})| \geq 1$ and a $p\text{value} (P_{adj}) < 0.05$, as detected by DESeq.

2.4. Identification of signature metabolites accumulated in Qingyuan and Xiushui counties

According to Han et al. (2023) identified feature metabolites by utilizing the RF function in the “caret” package. The selection was based on the significance of these metabolites in classification, considering the variations among the metabolome samples. A total of twenty metabolites were included in the analysis. To prevent sample size bias and overfitting, 100 replicas of the random forest model were trained and tested using a well-balanced ten-fold cross-validation approach. Using the “e1071” library in R, a support vector machine (SVM) model was built, and metabolites with an average rank below 30 were chosen as the featured metabolites. Next, we deduced the metabolites that overlapped between the two classifiable models.

2.5. Analysis of gene co-expression networks using weightings

The WGCNA package (version 1.6.6) in R software (version 3.4.4) was utilized to construct weighted gene co-expression networks and identify associated modules. Here, $r > 0.90$ and $p < 0.001$ were defined as the criteria for identifying significant modules related to flavonoids, whereas $r \geq 0.80$ and $p \leq 0.002$ were established as the criteria for significant modules connected to phenolic acids and lignans. To be considered as hub genes related to the compounds, two criteria had to be met: (1) A single gene-trait correlation over 0.90, and (2) an edge weight larger than 0.1 are also required. Gephi was used to display the network of co-expression genes. The RNA-seq data included in this study were obtained from Han et al. (2023).

2.6. *Transient expression of four genes in tobacco leaves*

PcOMT10, *PcOMT11*, *PcOMT12*, and *PcOMT13* CDS were cloned and subsequently injected into *Agrobacterium tumefaciens* GV3101. All constructions were expressed transiently in tobacco leaves. After three days of injection, positive transgenic leaves were identified using quantitative real-time PCR. Widely

targeted metabolomics were used to identify phenylpropanoids in over-expressed transgenic tissues. In these trials, three biological replicates were employed.

2.7. Extraction of RNA and quantitative real-time PCR (qRT-PCR)

Initially, using the MiniBEST Plant RNA Extraction kit (produced by TaKaRa in Japan), total RNA for qRT-PCR was extracted. Afterwards, the PrimerScript RT Enzyme Mix I kit from TaKaRa, Japan, was used to reverse-transcribe the RNA that had been extracted previously into cDNA. SYBR® Premix Ex Taq II (TaKaRa, Japan) and a CFX96 Touch™ Real-Time PCR System (BIO-RAD, USA) were used for the qRT-PCR analysis. We evaluated the relative expression of DcACTIN using the $2^{-\Delta\Delta CT}$ technique. Each analysis required technical replicates and three independent biological replicates, respectively. Primers design were conducted by Primer Premier 5.

3. Results

3.1. Metabolite identification and statistics among 27 genotypes

To better understand fluctuating variations in bioactive substances among 27 different types from Qingyuan and Xiushui counties, a comprehensive total of 482 metabolites in *P. cyrtonema* rhizomes were successfully identified using UPLC-ESI-MS/MS (Table S1). The PCA analysis revealed a strong correlation between samples from Qingyuan and Xiushui counties within each county, but there was a distinct separation observed between the two counties (Fig. 1B). These metabolites were then categorized into 11 distinct groups, which include lignans and coumarins (8 metabolites), lipids (89 metabolites), nucleotides and derivatives (17 metabolites), organic acids (36 metabolites), steroids (52 metabolites), phenolic acids (50 metabolites), alkaloids (51 metabolites), amino acids and derivatives (50 metabolites), Carbohydrates (23 metabolites), Flavonoids (81 metabolites), and Others (25 metabolites) (Table S2; Fig. 1C and S1). And phenylpropanoid metabolites accounted for 28.83% of the total metabolites. The clustering analysis and heatmap, based on the entirety of these metabolomics data, also demonstrated a noticeable differentiation between the two regions (Fig. 1D), suggesting variations in the accumulation of these metabolites under different cultivation conditions.

3.2. Machine learning-based selection of the best metabolite markers

The SVM-RFE method was used to screen a total of 482 metabolites identified through widely targeted metabolome analysis. Based on the average rank and cv error, 29 feature metabolites were ultimately determined (Table S3). Furthermore, the top RF model, which relied on 44 trees, recognized 20 metabolites as feature indicators (Table S4; Fig. 2A and B). After the classification by machine learning, the metabolites that were considered most significant were quercetin-3-O-(6"-malonyl) galactoside,

padelaoside C, 9-Hydroxy-12-oxo-10(E),15(Z)-octadecadienoic acid, epipinoresinol, and salicin (Fig. 2C). Based on the aforementioned findings, flavonols, lignans, and phenolic acids were identified as the three primary metabolites, suggesting a substantial presence of phenylpropanoid metabolites. Then, we obtain three common compounds using the two classification models (Fig. S2). Furthermore, we successfully grouped all metabolites associated with phenylpropanoids, particularly flavonoids and phenolic acids, from both areas. Interestingly, we observed a distinct regional differentiation for these compounds as well (Fig. 2D). The results indicated that the metabolites related to phenylpropanoids were significant markers that could effectively differentiate the germplasm resources of *P. cyrtoneura* in Qingyuan and Xiushui counties.

3.3. Mining of key candidate genes in flavonoid biosynthesis

To uncover the mechanism of flavonoid accumulation, we chose T118_21 (from Qingyuan County, Zhejiang Province) and T26_14 (from Xiushui County, Jiangxi Province) as the representative genotypes. By comparing the metabolomes of these two genotypes, we found that 14 metabolites had significantly different content. The strict screening criteria for these differences were $|\log_2(\text{Fold Change})| > 2$ and $p\text{value} < 0.05$. This is shown in Fig. 3A. Additionally, the heatmap in Fig. 3B clearly illustrates the variations in their relative contents.

To uncover the molecular process of flavonoid production, a comparative analysis of transcriptomes was conducted on T118_21 and T26_14. T118_21 had 5074 genes that were up-regulated and 4166 genes that were down-regulated, which is different from T26_14. Significantly different genotypes from two regions were analyzed to identify significantly differentially expressed genes related to the biosynthesis of plant phenylpropanoids and their downstream pathways, due to the notable variation in phenylpropanoid compounds. In the end, a total of 69 genes related to flavonoid structure, 71 genes related to phenylpropanoid structure, and 485 regulatory genes were identified when comparing T118_21 and T26_14 using the screening criteria of $|\log_2(\text{Fold Change})| > 1$ & $p\text{value} < 0.05$ (Fig. 3D). Among them, four candidate genes related to the synthesis of phenylpropane metabolites with highly significant differences were identified ($|\log_2(\text{Fold Change})| > 3$ & $p\text{value} < 0.01$) (Fig. S3). For phylogenetic analysis of these four candidate genes, we aligned their protein sequences with characterized genes in *Arabidopsis thaliana*, and then we constructed a neighbor-joining tree (Fig. S4) using the MEGA7 software. As a result, the genes *AT1G59950*, *AT1G59960*, *AT5G54160*, and *AT4G35150* were clustered with them and found to belong to the *O*-methyltransferase family (Table S5) and named *PcOMT10*, *PcOMT11*, *PcOMT12* and *PcOMT13*.

Combining transcriptome and 14 flavonoid metabolites with significantly differential content, WGCNA was applied to identify associated modules with these compounds. The findings showed a strong and significant correlation between quercetin-3-O-(6"-malonyl) galactoside with the darkturquoise module. Additionally, there was a significant correlation observed between vitexin-2"-O-xyloside and chrysoeriol-7-O-gentiobioside and the yellow module, as depicted in Fig. 3C. Furthermore, there was a strong correlation

between apigenin-6-C-xyloside-8-C-arabinoside and apigenin-6-C-glucoside-8-O-xylcoside with the green module. Additionally, schaftoside and vitexin-O-xyloside exhibited a significant association with the black module ($r > 0.95$, $P\text{-value} < 0.001$). Hub genes, which include the flavonoids structural genes *CHI*, *UGT1*, and *UGT2*, as well as an ERF gene that positively regulates the *UGT1* gene, were selected from the four modules based on their weight value (> 0.1) and correlation relationship with the flavonoid content ($r > 0.90$). These hub genes are potential pathway genes involved in the biosynthesis of flavonoids.

3.4. Identification of crucial candidate genes involved in the synthesis of phenolic acids and lignans through mining

There are 11 different contents of metabolites, including three upregulated and eight downregulated metabolites, were identified in T118_21 compared to T26_14 through a comparative analysis of phenolic acids and lignans. The screening criteria used was $|\log_2(\text{Fold Change})| > 1$ and $p\text{value} < 0.05$, as shown in Fig. 4A and 4B.

According to WGCNA analysis, there was a strong positive correlation ($r > 0.9$, $p < 0.001$) between p-Coumaroylmalic acid and the blue module, whereas Salicin exhibited the highest correlation with the lightgreen module (Fig. S5). Furthermore, Epipinoresinol exhibited a strong correlation with the lightpink4 module. By utilizing a weight threshold (> 0.1) and examining the correlation between phenolic acids and lignans contents and module eigengene values ($r > 0.80$), we identified hub genes from four modules. These hub genes include *UGT3* (the phenylacetone structural gene) and *NAC*, which negatively regulate the *UGT3* gene and were ultimately selected.

3.5. Characterization of important genes in responsible for phenylpropanoids

In order to investigate the role of *PcOMT10*, *PcOMT11*, *PcOMT12*, and *PcOMT13* in the accumulation of phenylpropanoids, we conducted transient transformation experiments on tobacco leaves for studying their biological function. The analysis of these four genes was conducted using qRT-PCR with the primers and actin (Table S6). The findings indicated that, in tobacco leaves, the expression levels of these genes were markedly upregulated compared to the control vector (Fig. 3E).

The expression of *PcOMT10*, *PcOMT11*, *PcOMT12* genes significantly changed the content of three phenylpropanoids in tobacco leaves, namely chrysoeriol-6,8-di-C-glucoside (Fig. 3F), syringaresinol-4'-O-glucopyranosid (Fig. 4C), 1-O-Sinapoyl-D-glucose (Fig. 4D). The results showed that their content increased or decreased by as much as a factor of two. The variations in content changes for the remaining compounds identified were comparatively modest, with merely different levels of content increase or reduction. Red genes are utilized in Fig. 3G and 4E to show promotion of compound accumulation, and green genes indicate inhibition of compound accumulation, in order to better comprehend the changes after gene overexpression with related compounds. Collectively, this suggests that these four genes exhibit a strong catalytic specificity within the phenylpropane pathway in tobacco.

3.6. Biosynthetic model of phenylpropanoid compounds in *P. cyrtonema*.

Based on metabolites markers between two regions and comparative transcriptome analysis of two representative genotypes, we constructed a biosynthetic model of important phenylpropanoids in *P. cyrtonema*. The manifestation of four genes *PcOMT10/11/12/13* in tobacco, did align with the pattern of compound accumulation in *P. cyrtonema* (Fig. 5). Surprisingly, *PcOMT12* has a promotional effect on the synthesis of several oxymethylated phenylpropane compounds, as epipinoresinol, 4-O-Glucosyl-sinapate and chrysoeriol-6,8-di-C-glucoside. Moreover, *PcOMT13* somewhat inhibited the accumulation of chrysoeriol-7-O-gentiobioside content. However, *PcOMT10* and *PcOMT11* exhibit different roles for the synthesis of different compounds.

4. Discussion

There are a number of biological properties of *P. cyrtonema*, such as anti-aging, nourishing Yin, anti-inflammation, and immune regulation. For more than two millennia, China has been cultivating it as an alternative source of sustenance [4]. Although it is widely planted in China, there are large differences in germplasm quality among regions (Fig. S1), and the mechanism of related compound synthesis has not been well elucidated.

Applying machine learning techniques to examine biomarkers in liver tissue following bariatric surgery [18]; investigating the genetic biomarkers and molecular mechanisms linked to postmenopausal osteoporosis caused by smoking [19], among other studies. Nonetheless, there have been limited investigations employing machine learning algorithms on botanical specimens. Over the past few years, a small number of individuals have begun utilizing machine learning techniques for fruit flavor identification [5]. Additionally, they have employed metabolomic and genomic data to identify marker genes for flavonoids in *P. cyrtonema*, thus contributing to crop improvement breeding [20]. Using machine learning, we employed a study to detect significant marker metabolites linked to the distinctions between Qingyuan and Xiushui counties. Multiple species were found to contain these primary marker metabolites, including quercetin-3-O-(6"-malonyl) galactoside, salicin, and epipinoresinol, which were identified using machine learning in this research. Quercetin-3-O-(6"-malonyl) galactoside, one of them, can lead to the browning of iceberg lettuce as a result of interactions with enzymes [21]. In *Ribes nigrum* L., its concentration was observed to be greater in the northern regions compared to the southern regions when analyzing its growth site [22]. Salicin is frequently derived from herbaceous plants to alleviate pain in the management of acute rheumatism [23]. Forsythoside synthesis involves the participation of epipinoresinol, a compound commonly present in forsythia [24]. Although we have identified these crucial marker metabolites and established their status as primary phenylpropanoids, the molecular synthesis mechanism behind them remains unclear.

This study utilized WGCNA and transcriptome analysis to discover four *CHI* structural genes, *UGT1/2/3* genes, and two regulatory genes *ERF* and *NAC*, which control *UGT1* and *UGT3*, respectively. Naringenin,

typically produced by chalcone isomerase (CHI), serves as an intermediate compound in the flavonoid pathway, facilitating the production of diverse flavonoids and isoflavonoids [25]. The study found a positive correlation between CHI and vitexin-2"-O-galactoside. Over the past few years, research on plant glycosyltransferases has revealed notable variations in glycosyltransferases involved in plant secondary metabolism, including *Arabidopsis* [26], cereals [27], and rice [28]. Nevertheless, the glycosylation process of *P. cyrtonema*, a significant plant in terms of medicinal properties due to its glycosylated flavonoids, remains largely unexplored, and there is limited information on the regulatory genes involved. The study revealed a decrease in *UGT2* expression, however, there was a notable rise in the levels of compounds associated with its production, such as apigenin-6-C-xyloside-8-C-arabinoside and apigenin-6-C-glucoside-8-O-xylcoside. This suggests that *UGT2* potentially plays a suppressive role in the accumulation of these compounds. According to Wan et al. (2023), discovered that the temporary amplification of *CsERF003* in citrus fruits resulted in a notable enhancement in the levels of flavanones, flavonoids, and flavonols. Additionally, it also led to an increase in the expression of crucial genes associated with these metabolites. These findings indicate that *CsERF003* might play a role in regulating the biosynthesis of flavonoids by influencing the expression of *UGTs* [29]. The findings of this research are comparable, and *ERF* might enhance the expression of *UGT1* to facilitate the buildup of quercetin-3-O-(6"-malonyl) galactoside. According to previous research conducted by Zhao et al. (2022), it has been demonstrated that NAC plays a significant role in plant stress response primarily through the regulation of flavonoid synthesis [30]. In *Arabidopsis*, the regulation of flavonoid biosynthesis in response to high light stress is controlled by the NAC transcription factor *ANAC078* [31]. Similarly, in Norway spruce, a stress-induced transcription factor from the NAC gene family, overexpression of *PaNAC03*, results in decreased flavonol biosynthesis and abnormal embryo development [32]. In the current investigation, it is our conjecture that T118_21 genotype experienced various types of stress. This is supported by the notable increase in *NAC* gene expression observed in T118_21 compared to T26_14, which provides a more plausible explanation for the considerable reduction in Salicin content subsequent to *NAC* overexpression.

The *O*-methyltransferase enzyme exhibits unexpectedly wide compatibility and selectivity towards various substrates, which is always an increasingly popular topic of study for several scientists. Spanning from basic catechols to complex phenylpropanoids and isoquinoline alkaloids, as observed in studies conducted by Joshi and Chiang [33] and Frick and Kutchan [34]. Functional diversity is exhibited by flavonoid *O*-methyltransferases. In comparison to *AtOMT1*, *liOMT3* methylated the 3'-hydroxy moiety of flavonoids like eriodictyol and 3'-hydroxydaizein. Additionally, it methylated the 7-OH positions of flavones and facilitated the conversion of caffeic acid into ferulic acid [35]. In this study, substrate binding and specificity in plant *O*-methyltransferases can be well verified by the results of widely targeted metabolomics of *P. cyrtonema* and overexpressed transgenic tobacco. Meanwhile, the four genes (*PcOMT10/11/12/13*) mined in *P. cyrtonema* showed similar expression patterns to those in tobacco for catalyzing the generation of target products from substrates, demonstrating that these enzymes are relatively functionally conserved for the catalytic mechanism among different species. Among them, the *PcOMT12* gene played a positive role in different target compound accumulation in multiple reactions and significantly increased chrysoeriol-6,8-di-C-glucoside content after overexpression in tobacco, which

is presumed to have high catalytic activity and widely catalytic substrates. It is expected to be further exploited in the future.

In this study, we used machine learning for the first time to identify significant metabolites affecting inter-regional variation in *P. cyrtanema*, then used multi-omics to mine synthetically related candidate genes and transiently expressed genes in tobacco leaves. The results of this research offer fresh insights into the molecular basis of phenylpropanoid accumulation in *P. cyrtanema*.

5. Conclusion

The study focused on the metabolomics of *P. cyrtanema* in Qingyuan County (Zhejiang Province) and Xiushui County (Jiangxi Province). Based on machine learning, discovering three phenylpropanoid compounds (quercetin-3-O-(6"-malonyl) galactoside, epipinoresinol, salicin) as the best markers metabolites between these two regions. Additionally, comparative transcriptome analysis and WGCNA helped identify structural genes (*CHI*, *UGT1/2/3*) and regulatory genes (*ERF*, *NAC*) that likely influence their biosynthesis. Furthermore, temporary over-expression trials conducted in tobacco leaves validated the role in altering metabolic flow for the accumulation of phenylpropanoids in these four genes that were expressed differentially (*PcOMT10/11/12/13*). These results are expected to provide an important basis for subsequent genetic studies and precision breeding of *P. cyrtanema* and produce high-value-added medicinal products.

Declarations

Statement of authorship contribution using CRediT

Qiqi Gong is responsible for conceptualizing, curating data, conducting formal analysis, investigating, visualizing, and writing drafts. Jianfeng Yu, Ke Fu: Writing draft and Formal analysis. Yi Xu: Experiments. Zhicheng Guo, Hui Zou, Cong Li, Shengguan Cai, Jinping Si: Investigation and Conceptualization. Donghong Chen and Zhigang Han contributed to the conceptualization, data curation, funding acquisition, methodology, project administration, supervision, and writing - review & editing of the study.

Disclosure of Conflict of Interest

Author Hui Zou was employed by Zhigang Han. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The National Natural Science Foundation of China (32200214), the Natural Science Foundation of Zhejiang Province (LQ22H280003), and the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2022C04035) provided funding for this study.

Data Availability Statement

Transcriptome data can be found in NCBI (<https://www.ncbi.nlm.nih.gov/>), and the submission numbers is SUB13866596.

Ethical statement

Ethics approval was not required for this research.

References

1. Shi Y, Si D, Chen D, Zhang X, Han Z, Yu Q, et al. Bioactive compounds from Polygonatum genus as anti-diabetic agents with future perspectives. *Food Chemistry*. 2023;408:135183.
2. Si J, Zhu Y. Polygonati rhizoma—a new high-quality crop with great potential and not occupying farmland. *Sci Sin-Vitae*. 2021;51:1477–84.
3. Zhao L, Xu C, Zhou W, Li Y, Xie Y, Hu H, et al. Polygonati Rhizoma with the homology of medicine and food: A review of ethnopharmacology, botany, phytochemistry, pharmacology and applications. *Journal of Ethnopharmacology*. 2023;309:116296.
4. Chen D, Han Z, Si J. Huangjing (Polygonati rhizoma) is an emerging crop with great potential to fight chronic and hidden hunger. *Science China Life Sciences*. 2021;64.
5. Colantonio V, Ferrão LFV, Tieman DM, Bliznyuk N, Sims C, Klee HJ, et al. Metabolomic selection for enhanced fruit flavor. *Proc Natl Acad Sci USA*. 2022;119:e2115865119.
6. Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell*. 2020;182:59-72.e15.
7. Zhang X, Lu X, Shi Q, Xu X, Leung HE, Harris LN, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*. 2006;7:197.
8. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinformatics*. 2016;17:60.
9. Dong N, Lin H. Contribution of phenylpropanoid metabolism to plant development and plant–environment interactions. *J Integr Plant Biol*. 2021;63:180–209.
10. Tao H, Li L, He Y, Zhang X, Zhao Y, Wang Q, et al. Flavonoids in vegetables: improvement of dietary flavonoids by metabolic engineering to promote health. *Critical Reviews in Food Science and Nutrition*. 2022;0:1–15.
11. Rashmi HB, Negi PS. Phenolic acids from vegetables: A review on processing stability and health benefits. *Food Research International*. 2020;136:109298.
12. Jang WY, Kim M-Y, Cho JY. Antioxidant, Anti-Inflammatory, Anti-Menopausal, and Anti-Cancer Effects of Lignans and Their Metabolites. *IJMS*. 2022;23:15482.
13. Muro-Villanueva F, Mao X, Chapple C. Linking phenylpropanoid metabolism, lignin deposition, and plant growth inhibition. *Current Opinion in Biotechnology*. 2019;56:202–8.

14. D'Auria JC. Acyltransferases in plants: a good time to be BAHD. *Current Opinion in Plant Biology*. 2006;9:331–40.
15. Lv Y-Q, Li D, Wu L-Y, Zhu Y-M, Ye Y, Zheng X-Q, et al. Sugar signal mediates flavonoid biosynthesis in tea leaves. *Horticulture Research*. 2022;9:uhac049.
16. Han Z, Ahsan M, Adil MF, Chen X, Nazir MM, Shamsi IH, et al. Identification of the gene network modules highly associated with the synthesis of phenolics compounds in barley by transcriptome and metabolome analysis. *Food Chemistry*. 2020;323:126862.
17. The DREAM Module Identification Challenge Consortium, Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, et al. Assessment of network module identification across complex diseases. *Nat Methods*. 2019;16:843–52.
18. Jiang Y, Wang Y, Chen S, Liu Z, Yang H, Jiao Y, et al. Screening of Biomarkers in Liver Tissue after Bariatric Surgery Based on WGCNA and SVM-RFE Algorithms. *Disease Markers*. 2023;2023:1–23.
19. Li S, Chen B, Chen H, Hua Z, Shao Y, Yin H, et al. Analysis of potential genetic biomarkers and molecular mechanism of smoking-related postmenopausal osteoporosis using weighted gene co-expression network analysis and machine learning. *PLoS ONE*. 2021;16:e0257343.
20. Han, Z., Gong, Q., Huang, S., Meng, X., Xu, Y., Li, L., et al. Machine learning uncovers accumulation mechanism of flavonoid compounds in *Polygonatum cyrtoneura* Hua. *Plant Physiology and Biochemistry*. 2023;201:107839.
21. Mai F, Glomb MA. Isolation of Phenolic Compounds from Iceberg Lettuce and Impact on Enzymatic Browning. *J Agric Food Chem*. 2013;61:2868–74.
22. Yang W, Alanne A-L, Liu P, Kallio H, Yang B. Flavonol Glycosides in Currant Leaves and Variation with Growth Season, Growth Location, and Leaf Position. *J Agric Food Chem*. 2015;63:9269–76.
23. Ferguson GB. REMOVAL OF A MELANOTIC SARCOMA, ORIGINATING IN THE SHEATH OF THE SARTORIUS MUSCLE, FROM A MAN AGED SEVENTY-FIVE. *The Lancet*. 1876;107:384–5.
24. Rahman MMA, Dewick PM, Jackson DE, Lucas JA. Biosynthesis of lignans in *Forsythia intermedia*. *Phytochemistry*. 1990;29:1841–6.
25. Tohge T, de Souza LP, Fernie AR. Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. *Journal of Experimental Botany*. 2017;68:4013–28.
26. Gandia-Herrero F, Lorenz A, Larson T, Graham IA, Bowles DJ, Rylott EL, et al. Detoxification of the explosive 2,4,6-trinitrotoluene in *Arabidopsis*: discovery of bifunctional *O*- and *C*-glucosyltransferases. *The Plant Journal*. 2008;56:963–74.
27. Brazier-Hicks M, Evans KM, Gershtater MC, Puschmann H, Steel PG, Edwards R. The C-Glycosylation of Flavonoids in Cereals. *Journal of Biological Chemistry*. 2009;284:17926–34.
28. Ko JH, Kim BG, Hur H-G, Lim Y, Ahn J-H. Molecular cloning, expression and characterization of a glucosyltransferase from rice. *Plant Cell Rep*. 2006;25:741–6.
29. Wan H, Liu Y, Wang T, Jiang P, Wen W, Nie J. Combined transcriptomic and metabolomic analyses identifies CsERF003, a citrus ERF transcription factor, as flavonoid activator. *Plant Science*.

2023;334:111762.

30. Zhao X, Wu T, Guo S, Hu J, Zhan Y. Ectopic Expression of AeNAC83, a NAC Transcription Factor from *Abelmoschus esculentus*, Inhibits Growth and Confers Tolerance to Salt Stress in *Arabidopsis*. *IJMS*. 2022;23:10182.
31. Morishita T, Kojima Y, Maruta T, Nishizawa-Yokoi A, Yabuta Y, Shigeoka S. *Arabidopsis* NAC Transcription Factor, ANAC078, Regulates Flavonoid Biosynthesis under High-light. *Plant and Cell Physiology*. 2009;50:2210–22.
32. Dalman K, Wind JJ, Nemesio-Gorriz M, Hammerbacher A, Lundén K, Ezcurra I, et al. Overexpression of PaNAC03, a stress induced NAC gene family transcription factor in Norway spruce leads to reduced flavonol biosynthesis and aberrant embryo development. *BMC Plant Biol*. 2017;17:6.
33. Joshi CP, Chiang VL. Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases.
34. Frick S, Kutchan TM. Molecular cloning and functional expression of O-methyltransferases common to isoquinoline alkaloid and phenylpropanoid biosynthesis. *Plant J*. 1999;17:329–39.
35. Tan Y, Yang J, Jiang Y, Sun S, Wei X, Wang R, et al. Identification and characterization of two *Isatis indigotica* O-methyltransferases methylating C-glycosylflavonoids. *Horticulture Research*. 2022;9:uhac140.

Figures

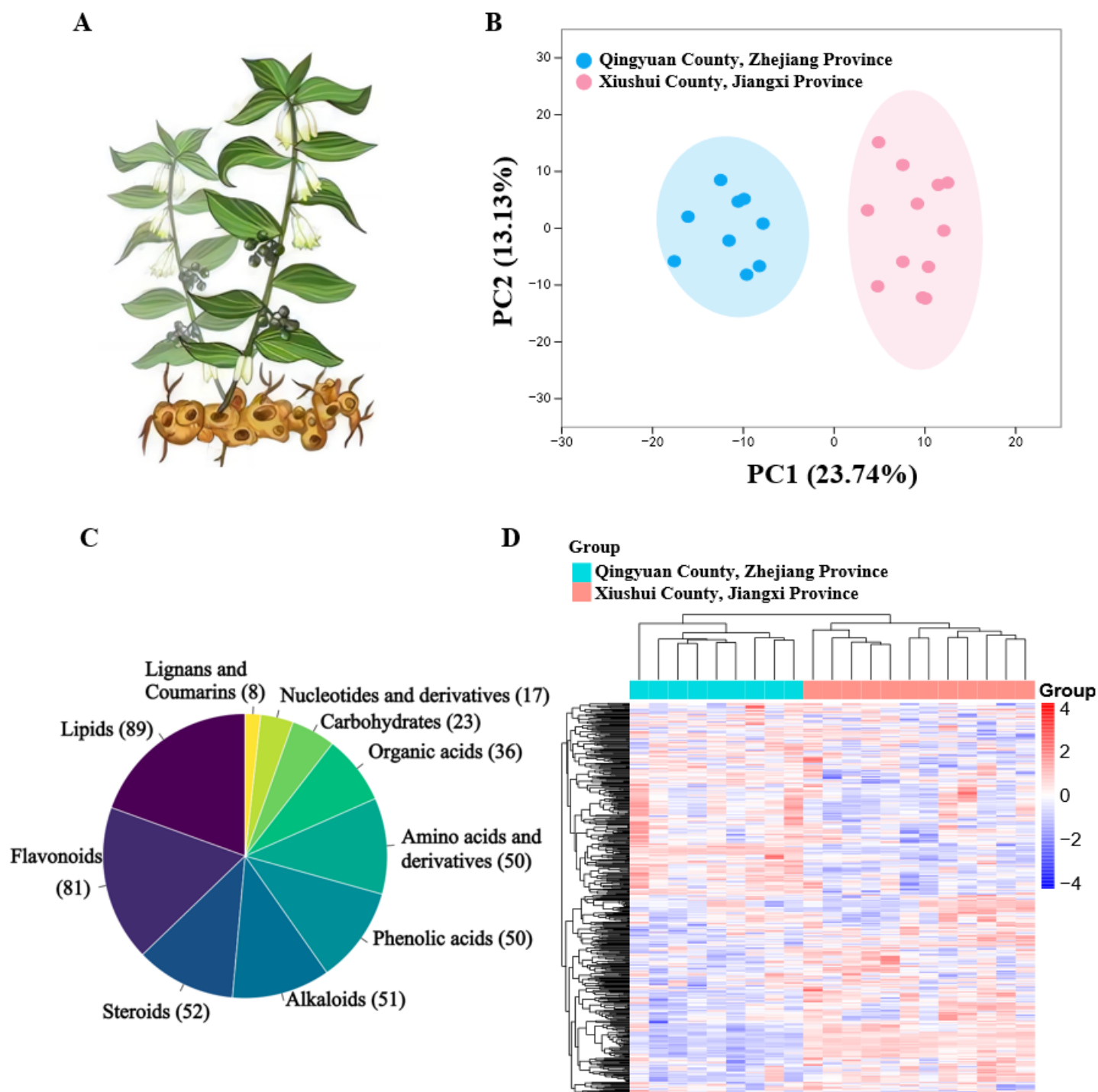


Figure 1

Metabonomic analysis of *P. cyrtonema* from two sites in Qingyuan and Xiushui Counties. (A) Diagram of *P. cyrtonema* plant. (B) PCA analysis of the two sites. (C) Component analysis of the identified metabolites from *P. cyrtonema*. (D) Complexheatmap of all metabolites identified by relative quantification.

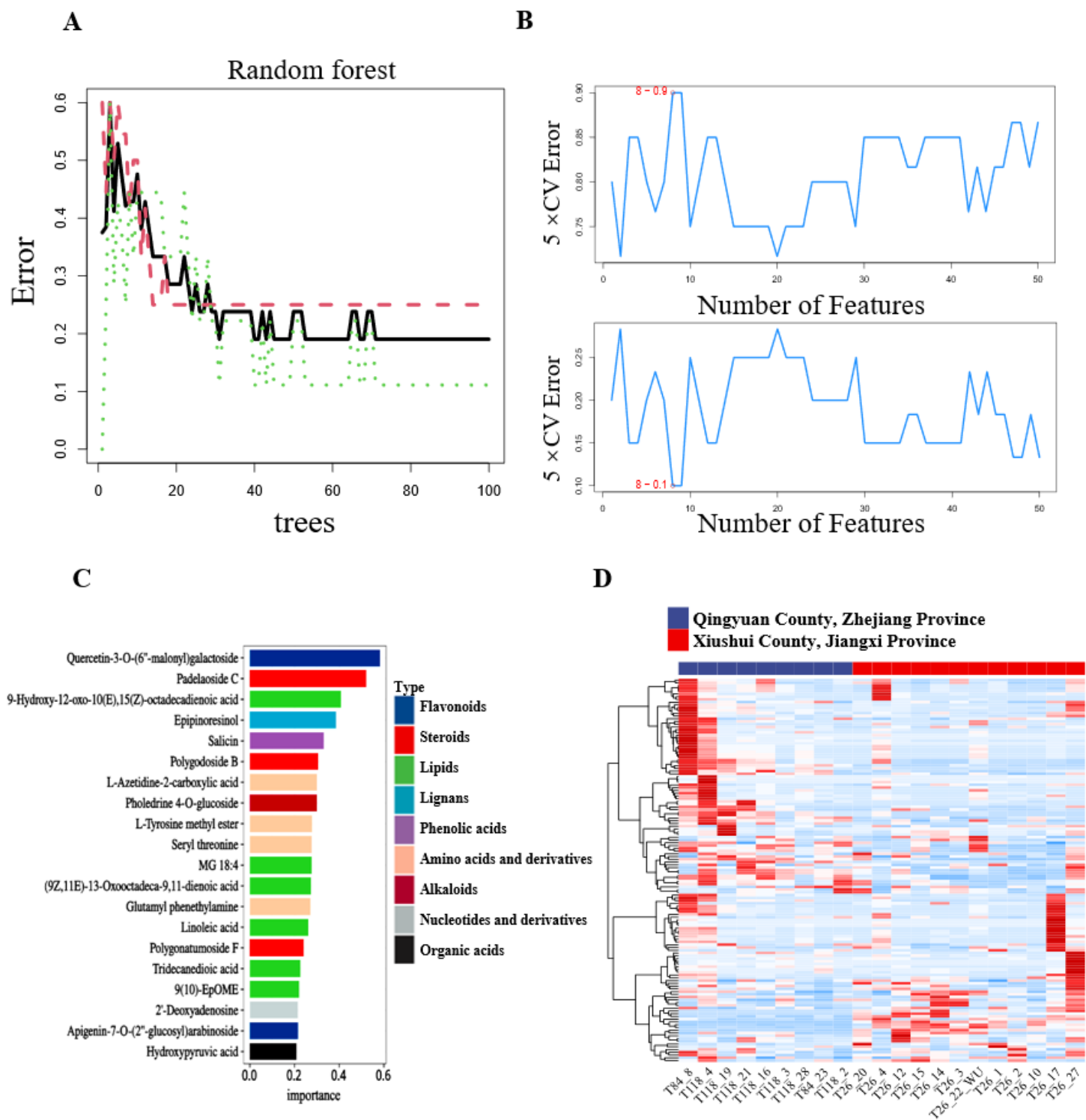


Figure 2

Machine learning methods identify characteristic metabolites. (A) The number of decision trees and the relationship to the error rate. (B) SVM-RFE, a support vector machine - recursive feature elimination approach. (C) Top 20 of feature metabolite results in the random forest method. (D) Complexheatmap of all phenylpropane metabolites in two sites.

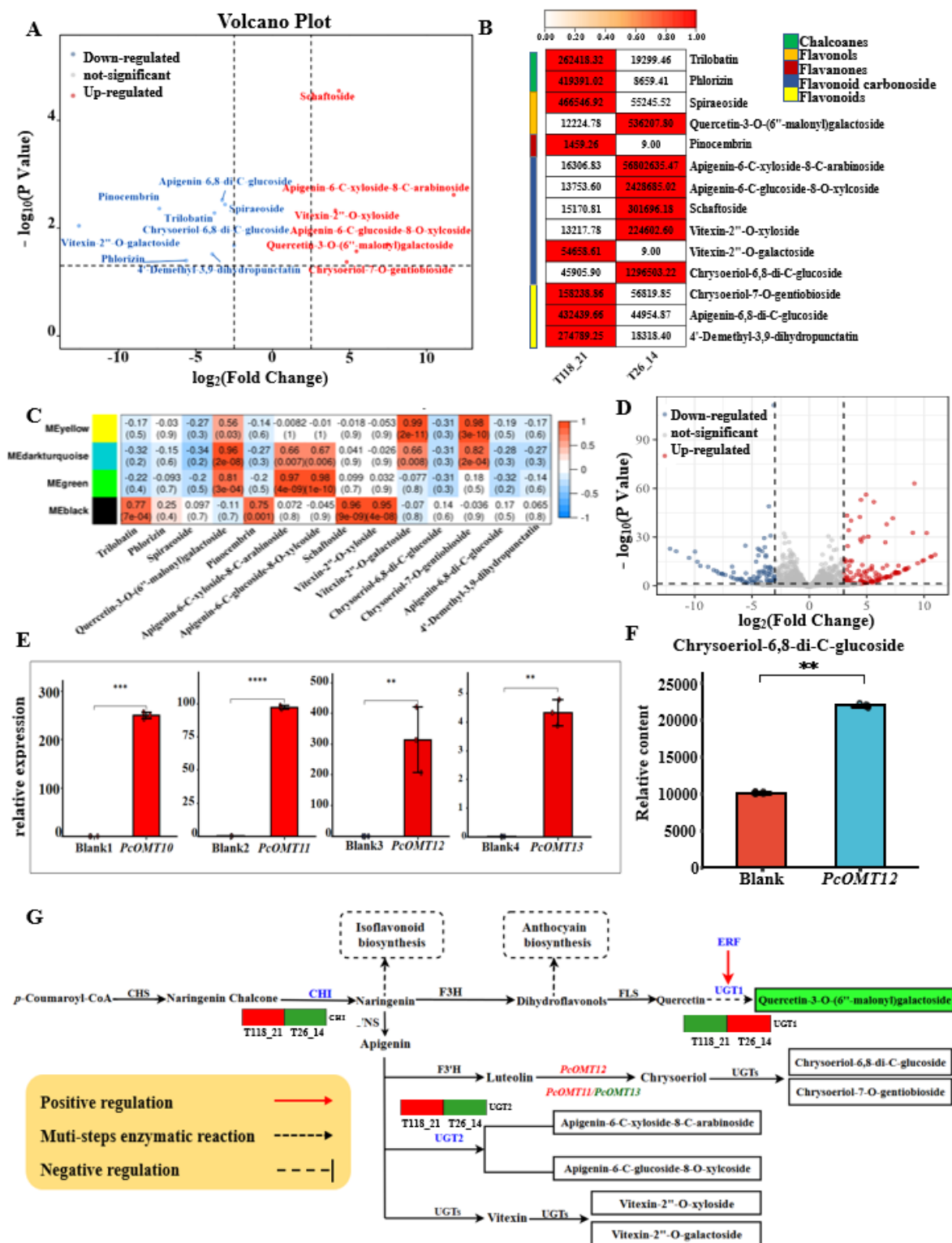


Figure 3

Mining of key candidate genes in flavonoid biosynthesis. (A) Volcano plot of flavonoids metabolites; (B) Relative content of flavonoids compounds. (C) Correlation of flavonoids and modules. (D) Volcano map of differentially expressed genes(pvalue < 0.05). (E) qRT-PCR of *PcOMT10/11/12/13* in tobacco leaves. (F) Changes in chrysoeriol-6,8-di-C-glucoside content after overexpression in tobacco leaves. (G) Flavonoid biosynthesis pathway and expression of key genes in the *P. cyrtoneura*.

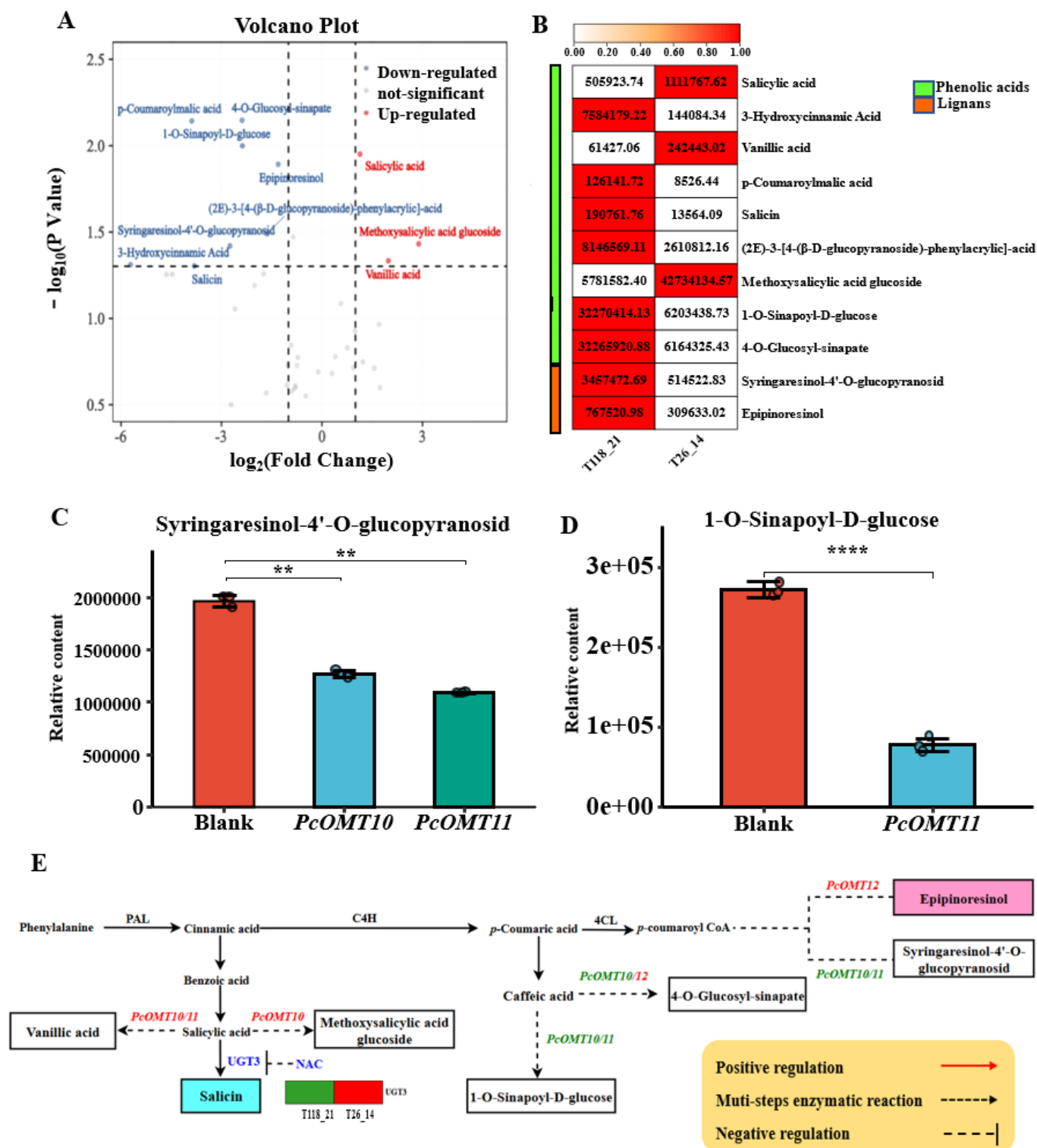


Figure 4

Mining of key candidate genes in the biosynthesis of phenolic acids and lignans. (A) Volcano plot of phenolic acids and lignans metabolites. (B) Heatmap of relative content of phenolic acids and lignans compounds. (C) Changes in syringaresinol-4'-O-glucopyranosid content after gene overexpression in tobacco leaves. (D) Changes in 1-O-Sinapoyl-D-glucose content after gene overexpression in tobacco

leaves. (E) Phenolic acids and lignans biosynthesis pathway and expression of key genes in the *P. cyrtoneura*.

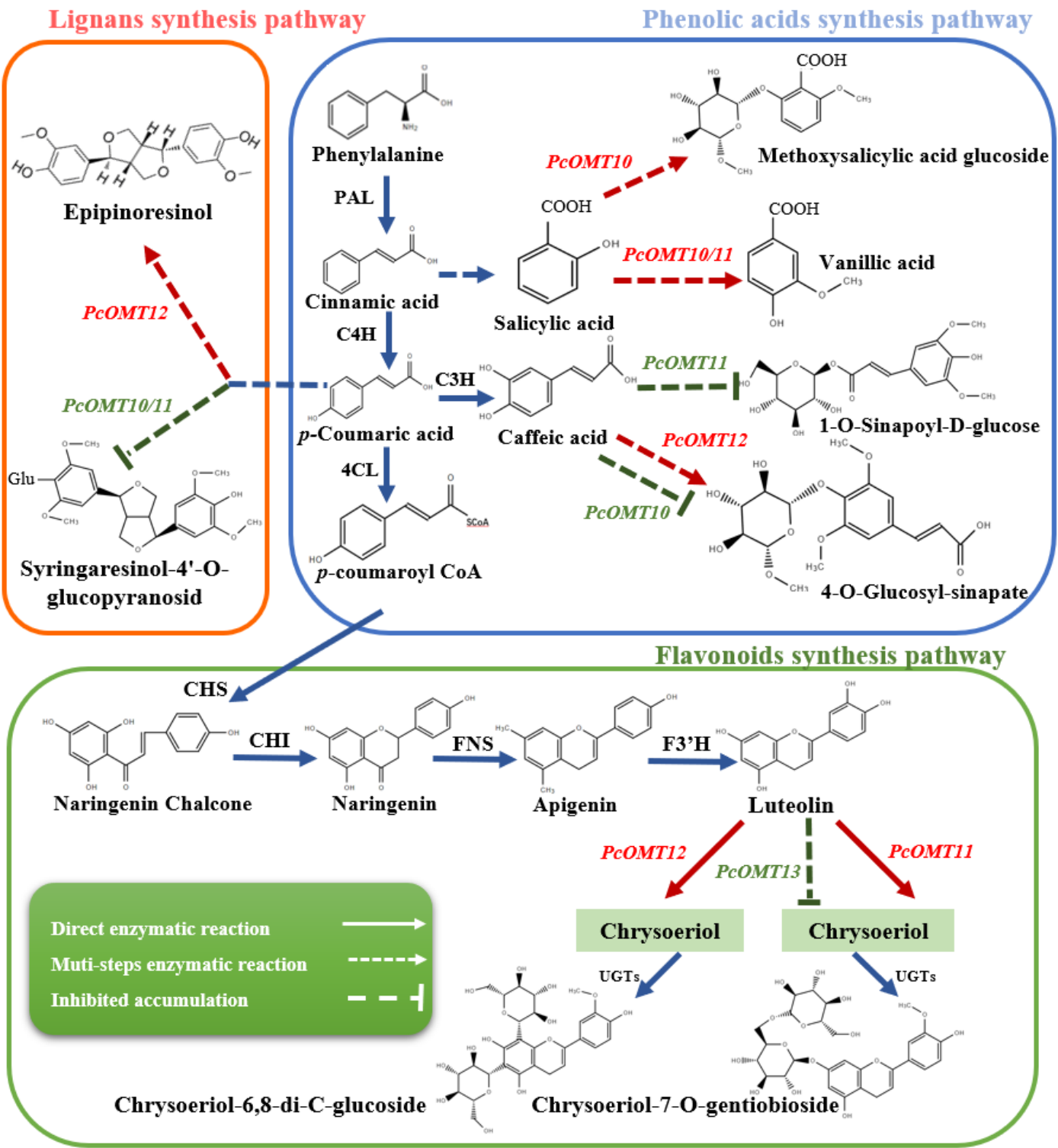


Figure 5

Analysis of the biosynthetic pathway of phenylpropane in *P. cyrtoneura*. In the pathway diagram, red genes indicate promotion of compound accumulation, while green genes indicate inhibition of compound

accumulation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS16.xlsx](#)
- [FigureS15.pptx](#)