Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative Transcriptome and Metabolome Analysis Reveals the Regulatory Networks and Key Genes of Wild and Cultivated Gentiana macrophylla Pall

Juanjuan Liu
Gansu University of Traditional Chinese Medicine
Jialing Zhang
Gansu University of Traditional Chinese Medicine
Yiyang Chen
Gansu University of Traditional Chinese Medicine
Ke Li
Gansu University of Traditional Chinese Medicine
Liangcai Wang
Gansu University of Traditional Chinese Medicine
Xiaohui Ma
Gansu University of Traditional Chinese Medicine
Ling Jin
zyxyjl@163.com

Gansu University of Traditional Chinese Medicine

Article

Keywords: Gentiana macrophylla Pall., wild, cultivated, metabolomics, transcriptomics, terpenoids biosynthesis, flavonoid biosynthesis.

Posted Date: November 15th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5016486/v1

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

Additional Declarations: No competing interests reported.

Abstract

Gentiana macrophylla Pall. is a medicinally important perennial herb used in traditional Chinese and ethnic medicines. Iridoids are the main chemical constituents of *G. macrophylla.* The scarcity of the wild resource has led to increased attention to cultivated *G. macrophylla*. However, little is known about the metabolic differences and the regulatory mechanisms between cultivated and wild *G. macrophylla*. This study utilized untargeted metabolomics and transcriptomics to reveal differences in accumulated metabolites and expressed genes between wild and cultivated. The metabolomics profiling revealed 25587 differentially accumulated metabolites (DAMs) while the transcriptomic profiling identified 6830 differentially expressed genes (DEGs). Analysis revealed that DEGs were predominantly enriched for processes associated with monoterpenoid biosynthesis and flavonoid biosynthesis. In addition, we verified six differential genes involved in monoterpenoid biosynthesis and flavonoid biosynthesis by RT-qPCR. The results showed that the relative expression levels in cultivation were higher than those in the wild. According to KEGG pathway analysis, *10HGO*(8-hydroxygeraniol dehydrogenase) may be a key enzyme encoding secoiridoid biosynthesis. The comprehensive results of transcriptome and metabolomics analysis revealed a strong correlation between metabolite content and gene expression, which is helpful to further study the function of *G. macrophylla* Pall. and the regulation of biosynthesis of active components. It provides a theoretical basis for further development and protection of wild species of *G. macrophylla* in the future.

1. Introduction

Gentiana macrophylla is a traditional medicinal plant, which was first recorded in *the Shennong's Herbal Classic*. It has the effects of dispelling rheumatism, clearing damp heat, relieving arthralgia, and relieving deficiency heat. The *Chinese Pharmacopoeia* stipulates that the medicinal part of *Gentiana macrophylla* Pall., *Gentiana straminea* Maxim., *Gentiana crassicaulis* Duthie ex Burk., and *Gentiana dahurica* Fisch. are dry root, which is mainly used to treat rheumatism and arthralgia, stroke hemiplegia, tendon and vein spasm, joint pain, damp heat jaundice, bone steaming hot flashes, and infantile malnutrition fever [1]. Modern pharmacological studies have shown that *G. macrophylla* has anti-inflammatory, analgesic, antibacterial, antiviral, anti-tumor, immune regulation, liver protection, antioxidant, and other pharmacological activities [2–5]. The main chemical constituents include iridoids, triterpenoids, lignans, flavonoids, alkaloids, steroids, polysaccharides, and other types. Among them, iridoids are the main chemical constituents of *G. macrophylla*, belonging to monoterpenes [6].

The market survey found that due to the mixed origin of *G. macrophylla*, the resources of wild *G. macrophylla* Pall. have been reduced year by year due to long-term over-harvesting, this has resulted in a diminished market presence for this species, which is further complicated by its mixed genetic origins. The producing areas are mostly concentrated in Ningxia, Shaanxi, and Shanxi in China. The cultivated *G. macrophylla* Pall. has a large market share in Shaanxi, and studies have shown that the chemical composition, quality, and genetic diversity of Chinese medicinal materials are affected by the growth environment and production methods [6–9]. That is, there are quality differences between wild and

cultivated Chinese medicinal materials, and the clinical efficacy is also uneven, resulting in large price fluctuations of wild and cultivated Chinese medicinal materials (Jiang et al., 2023). Therefore, artificial cultivation of *G. macrophylla* Pall. has attracted widespread attention. However, it is well known that the role of *G. macrophylla* Pall. depends on its various components (including primary and secondary metabolites) and has different beneficial effects. However, limited studies have explored the differences in the integrated metabolic profiles and molecular mechanisms of manipulating metabolic diversity between wild and cultivated *G. macrophylla* Pall [10].

Metabolomics is a powerful means to the analysis and identification of the overall metabolic profiles in an organism. It is widely used in a variety of plants and Chinese herbal medicines. High-throughput transcriptome sequencing provides solutions for the discovery of functional genes associated with primary and secondary metabolites biosynthesis. Currently, techniques combining high-throughput metabolomics and RNA-sequencing (RNA-seq) technologies have been developed to study the dramatic changes of various primary and secondary metabolites in plants and Chinese herbs, and reveal the DEGs associated with biosynthesis of metabolites [11]. In this study, metabolic differences and associated gene expression between cultivated and wild types were investigated using a strategy integrating non-targeted metabolomics and transcriptomics [10]. The study not only helps to better understand the molecular differences between wild and cultivated *G. macrophylla* Pall but also provides a reliable basis for the quality research of cultivated *G. macrophylla* Pall [12].

2. Materials and Methods

2.1 Chemicals and Reagents

HPLC-grade acetonitrile and Methanol were purchased from CNW Technologies. Ammonium acetate was purchased from Sigma-Aldrich, Inc. (St. Louis, MO, USA). Ethanoic acid was purchased from Fisher Chemical (Shanghai Civi Chemical Technology Co., Ltd, Shanghai, China), and ddH2O is Watsons (Shanghai Civi Chemical Technology Co., Ltd, Shanghai, China). β-mercaptoethanol was purchased from Xilong Science (Xilong Science Co., Ltd, Shantou, China).

2.2 Sample Collection

The samples utilized in this study were sourced from the authentic geographical origins or primary cultivation regions of the Chinese medicinal materials, ensuring their representativeness. The wild and cultivated specimens of G. macrophylla Pall. were specifically procured from the Qinling region, a location renowned for its rich biodiversity and traditional medicinal plant cultivation. A total of 10 wild samples were collected and 10 cultivated. It was identified as G. macrophylla Pall in the Gentian family by Professor Jin Ling of the School of Pharmacy, Gansu University of Traditional Chinese Medicine. Wild and cultivated roots were mixed and divided into 3 groups as 3 biological replicates. The fresh material was cut directly into small pieces, immediately frozen in liquid nitrogen and stored at − 80 ℃ until further transcriptomic and metabolomic analysis.

2.3 UHPLC-QE-MS non-target metabolomics Analysis

2.3.1 Extraction and Separation of Metabolites

The samples were subjected to a freeze-drying process to remove moisture while preserving their structural integrity. Once freeze-dried, they were meticulously ground into a fine powder, ensuring a uniform consistency that is conducive to the subsequent analytical procedures. Weigh each sample (20mg) into an EP tube and add 1000μL of the extraction solution (methanol: acetonitrile: water = 2:2:1, with the isotopically-labeled internal standard mixture) [13]. The sample was then homogenized for 4 minutes using a homogenizer with a frequency of 35 Hz, followed by ultrasound in an ice water bath for 5 minutes. This combined process of homogenization and sonication was performed in three cycles to ensure thorough extraction. Subsequently, the samples were incubated at − 40°C for 1 hour and centrifuged at 4°C for 15 minutes at 12,000 rpm (RCF = 13800 (×g), R = 8.6 cm). The supernatant containing the extracted component was carefully filtered through a 0.22 μm microporous filter membrane and transferred to a glass vial for further analysis [14].

Ultra-high-performance liquid chromatography coupled with quadrupole-Orbitrap mass spectrometry (UHPLC-QE-MS) (Thermo Fisher Scientific, USA) was employed for non-targeted metabolomics analysis. It using a UHPLC system (Vanquish, Thermo Fisher Scientific) with a Phenomenex Kinetex C18 (2.1mm×100mm, 2.6μm) coupled to Orbitrap Exploris 120 mass spectrometer (Orbitrap MS, Thermo). The mobile phase consisted of 0.01% acetic acid aqueous solution (A) and isopropanol: acetonitrile (1:1, v/v) (B). The temperature of the automatic sampler was 4℃, and the injection volume was 2 μL[15]. The orbital well Exploris 120 mass spectrometer was used to acquire MS/MS spectra in information Dependent acquisition (IDA) mode under the control of acquisition software (Excalibur, Thermo Corporation). In this mode, the acquisition software continuously evaluates the full-scan MS spectrum. The ESI source conditions were set as follows: sheath gas flow rate as 50 Arb, Aux gas flow rate as 15 Arb, capillary temperature 320℃, full MS resolution as 60000, MS/MS resolution as 15000 collision energy as 20/30/40 in SNCE mode, spray Voltage as 3.8 kV (positive) or -3.4 kV (negative), respectively [16,17].

2.3.2 Pre-identification and Analysis of Metabolites

ProteoWizard converts raw data into mzXML format for peak detection, extraction, alignment, and integration. [18]. The internal MS2 database (BiotreeDB) was then used for metabolite annotation. The threshold for comments is set to 0.3. Multivariate data analysis, including principal component analysis (PCA) and orthogonal partial least squares discriminant analysis (OPLS-DA), was performed on the processed data and adjusted to the total peak intensity.

2.4 RNAseq-based transcriptome analysis

The experimental process of transcriptome sequencing includes RNA extraction, RNA detection, library construction, and computer sequencing [19]. Total RNA for 3 biological repeats for RNAseq was extracted

from the root, with about 1 μg RNA per sample as input material for RNA sample preparation. The purity, concentration, and integrity of RNA were detected by NanoDrop 2000 spectrophotometer and Agient2100/LabChip GX detector. Sequencing libraries were generated using NEBNext®Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) and index codes were added to attribute sequences to each sample.

Cluster the indexed encoded samples on the cBot Cluster Generation System using the TruSeq PE Cluster Kit V3-Cbott-HS (Illumia). After cluster generation, the library was prepared for sequencing and double-terminal reads were generated. The sequences were further processed using the online platform of Baimai Cloud (www.biocloud.net). Clean reads were obtained by removing reads containing joints, reads containing ploy-N and low-quality reads from the original data. At the same time, Q20, Q30, base composition and sequence repetition level of clean data were calculated. All downstream analysis is based on high-quality, clean data.

The DESeq R package (1.10.1) was used for differential expression analysis between the two groups. DESeq provides statistical routines for determining differential expression in digital gene expression data using a model based on negative binomial distribution. The P-values of the results were adjusted using Benjamini and Hochberg's method to control for error discovery rates. The genes with a p-value of 0.05 after DESeq correction was DEGs. After screening DEGs, DESeq2 was performed according to DEGs (v1.34.0, fold-change ≥ 2, correction p value ≤ 0.05). Clustering heat maps of different samples for functional annotation and enrichment analysis of DEGs, novel gene analysis [20]. Subsequently, upregulated and down-regulated DEGs were submitted to GO and KEGG enrichment analyses [21]. The GO enrichment analysis of DEGs was performed using the Kolmogorov-Smirnov test based on the topGO R package. KOBAS [22] software was used to conduct statistical enrichment tests for differentially expressed genes in KEGG pathway [23].

2.5 Integrated Analysis Between Metabolite Profiling and Transcriptome

The results of differentially expressed metabolite (metabolome) analysis were combined with the results of transcriptome analysis [19]. The genes showing altered transcriptomic, as well as metabolomic profiles, were mapped to the KEGG pathway map, showing pathways of simultaneous enrichment of differential metabolites and differential genes. The correlation between DEGs and metabolites in each group was analyzed and the network diagram was drawn. The correlation between them was then analyzed [19].

2.6 Validation of Quantitative Real-Time PCR

To verify the expression profile of DEGs obtained by RNA-seq, 6 genes were selected for q PCR verification. Using EASYspin Plus Polyphenols/Complex Plant RNA Rapid Extraction Kit (Adele) kit, RT-q PCR was performed for each tissue sample of the collected plant material with 3 biological replicates and at least 3 technical replicates per biological replicate [24]. cDNA was obtained using a reverse transcription kit (full gold). PCR amplification was performed using a 2X M5 UltraSYBR Mixture (Low

ROX). Each reaction volume was 16 μL, including 2 μL cDNA, 8 μL of 2 × M5 HiPer SYBR ® Premix Es Taq TM, 0.32 μL upstream and downstream primers (10umol $\cdot L^{-1}$), and 5.36 μL of ddH_2O. The q PCR procedure was set to 95°C for 10 min. 40 cycles of 95°C for 15 s, and 60°C for 30 s, and 94°C for 1 min, and 60°C for 1 min 30 s, and 94°C for 10 s. The relative expression was calculated by the $2^{-\Delta\Delta CT}$ method. The expression levels of 6 genes were consistent with RNA-seq data. This indicates that the sampling method and RNA-seq used in this study are suitable for studying the transcriptome model of *G. macrophylla* TRs.

2.7 Statistical analysis

Microbioinformatics (online bioinformatics analysis, visualization cloud platform) and chiplot online platform were used to map differential metabolites and differential gene data, such as cluster heat map, volcano map, and matchstick map. SIMCA 14.1 software was used to draw the metabolite PCA and OPLS-DA maps. GraphPad Prism 8 was used to visualize the significance of the combined analysis of differential metabolites and differential genes. SPSS Statistics 26 software was used to analyze whether the data were statistically significant (An alpha value of $p < 0.05$ was considered statistically significant).

3. Results

3.1 Metabolite Profiling of Samples

3.1.1 Multivariate statistical analysis of metabolites

UHPLC-QE-MS non-target metabolites analysis was performed to determine metabolic profiles of *G. macrophylla* Pall. root tissues. Based on the qualitative and quantitative metabolomics results of the UHPLC-QE-MS, univariate statistical analysis and Multivariate statistical analysis were carried out Analysis (MVA) to screen for metabolites with significant differences (VIP>1 and P-value<0.05) [25], among which shikimate, Shikimates and Phenylpropanoids(25.51%), Terpenoids（10.38%）, and Fatty acids（6.42%） are the main metabolites. They included Flavonoids (132), Iridoids monoterpenoids (32), Secoiridoid monoterpenoids (10), Lignans (28), and Coumarins (47).

Multivariate statistical analysis methods, such as PCA and OPLS-DA model, were inserted into the metabolite list after processing for further analysis to reduce the data dimension and improve the interpretability and validity of the data. All the samples in the PCA score plots were in the 95% confidence ellipse, and the proportion of two principal components was 62.3%. The coefficient of variation of the first principal component was 39.9%, and the coefficient of variation of the second principal component was 22.4%, indicating that the metabolites of wild and cultivated *G. macrophylla.* Pall (Figure.1a). PCA algorithms based on visual discrimination generally cannot account for the exact differences between samples. In this study, we used the OPLS-DA model for supervised classification to measure the degree of metabolite conversion in significantly separated samples of different wild and cultivated *G. macrophylla*. Figure 1b and Fig. 1c show a high degree of differentiation between the

sample groups in the OPLS-DA plots, with significant differences between the different groups. Our experiments also reveal excellent model parameters. Through cross-validation and response replacement test (RPT), OPLS-DA model does not appear overfitting. Based on these findings, OPLS-DA was proved to be an effective prediction model for analyzing metabolic differences between wild and cultivated *G. macrophylla* Pall.

3.1.2 Screening and classification of the differentially accumulated metabolites

A volcano plot was utilized to visualize the results of 25,587 differential metabolites (as depicted in Fig. 2a). Among these, 60,935 metabolites were upregulated, marked in red, while 9,596 were downregulated, indicated in blue. Metabolites that did not show significant changes were represented in gray. The identified metabolites were compared against the Human Metabolome Database (HMDB), leading to the successful annotation of 220 out of the 25,587 metabolites. A thorough analysis was performed to identify the top differentially expressed metabolites, both upregulated and downregulated. The top 10 metabolites in each category were narcotine, camelledionol, ferulic acid, anhydroicaritin, panaxadiol, pseudolaric acid b-o-beta-d-glucopyranoside, arjungenin, colchicine, 8-prenylnaringenin, tulipinolide, as shown in the matchstick analysis (Fig. 2b). According to the mean value of relative quantitative metabolites, cultivated *G. macrophylla* Pall. was 0.809, wild *G. macrophylla* Pall. was 0.737, and cultivated was higher than wild. Flavonoids, Phenolic acids, and Lignans were higher in cultivation than in the wild, while Isoflavonoids, Stilbenoids, and Coumarins were higher in the wild than in cultivation. Wild terpenoid metabolites were higher than cultivated ones, and the difference in Iridoids monoterpenoids metabolism was the same as that of cultivated ones, with 36 species up-regulated and 39 species down-regulated, among which Iridoids monoterpenoids were up-regulated and 6 species down-regulated. The number of wild fatty acids was higher than that of cultivated fatty acids, and the number of up-regulated (17) and down-regulated (18) fatty acids was higher than that of cultivated fatty acids.

3.1.3 Metabolic pathway enrichment analysis of differentially accumulated metabolites

Metabolic pathway enrichment analysis can provide insight into the mechanisms of metabolic changes in different experimental samples [26]. The KEGG (www.genome.jp/kegg/) pathways of *G. macrophylla* Pall. can be divided into cellular processes, environmental information processing, genetic information processing, metabolism, and organic systems, which we mapped to 78 KEGG pathways [27]. When comparing wild and cultivated *G. macrophylla* Pall., the enrichment of the top 20 differential metabolic pathways was performed, and the Rich Factor (the ratio of the number of differential metabolites annotated in one pathway to the number of all metabolites in the pathway) was calculated according to the enrichment results of the differential metabolites in the KEGG metabolic pathway. The greater the value, the greater the degree of enrichment (Figure.3a). Metabolic pathways with significant enrichment of metabolites include glyoxylate and dicarboxylate metabolism, monobactam biosynthesis, C5-

branched dibasic acid metabolism, flavone, and flavonol biosynthesis, alanine, aspartate and glutamate metabolism, histidine metabolism, D – amino acid metabolism, galactose metabolism, cyanoamino acid metabolism, plant hormone signal transduction, arginine biosynthesis, 2 – Oxocarboxylic acid metabolism, phenylpropanoid biosynthesis, tyrosine metabolism, beta – alanine metabolism, pantothenate and CoA biosynthesis, biosynthesis of amino acids, aminoacyl – tRNA biosynthesis, ABC transporters, and flavonoid biosynthesis. It is mainly subdivided into 11 categories (Figure.3b).

3.2 Transcriptome Analysis of Samples

3.2.1 Annotation and Expression of the Unigenes

A transcriptomic dataset of the cultivated and wild *G. macrophylla* Pall. was constructed and the results were summarized [24]. The transcriptome sequencing of 6 samples was completed, and a total of 43.82 Gb of Clean Data was obtained, with the Clean Data of each sample reaching 6.56 Gb, and the percentage of Q30 base was 93.17% or above. A total of 64,601 Unigene were obtained after assembly. Among them, there are 19,362 Unigene with a length of more than 1kb. A total of 31,722 Unigene annotation results were obtained, All the unigene assemblies were annotated into eight public databases, including nonredundant protein sequences (NR), Swiss-Prot (www.sib.swiss/swiss-prot/), Pfam (http://pfam.xfam.org/), Clusters of Orthologous Groups (COG, http://www.ncbi.nlm.nih.gov/COG/), Gene Ontology (GO, https://geneontology.org/), and the Kyoto Encyclopedia of Genes and Genomes (KEGG), and euKaryotic Ortholog Groups(KOG), and evolutionary genealogy of genes (eggNOG). NR is the most annotated database, with 30,652 unigenes, accounting for 96.63% of the total number of unigenes. The GO, KEGG, COG, Swiss-Prot, Pfam, KOG, and eggNOG, databases annotated 24,691 (77.83%), 20,036 (63.16%), 7519 (23.70%), 19,176 (60.45%), 20,413 (64.34%), 16,918 (53.33%), and 24.750 (78.02%) unigenes, respectively. All database petal figures (Fig. 4a).

KEGG database was used to annotate unigenes to further elucidate its role in biosynthesis and metabolic pathways. These unigenes were categorized into five main groups, which encompassed 50 sub-clusters. The largest number of KEGG identifiers was in Plant-pathogen interaction (182 annotated genes), followed by Pentose and glucuronate interconversions (82 annotated genes), Plant hormone signal transduction, and Phenylpropanoid biosynthesis (81 annotated genes) (Fig. 4b).

3.2.2 Functional Analysis of DEGs

The genes with significantly different expression levels in different samples are called DEGs, and the annotation and analysis of the metabolic pathway of DEGs will help to further understand the function of the genes. In the screening process, FDR < 0.01 and FC (fold change) ≥ 2 were used as screening criteria. Because gene expression is spatiotemporal specific, we compared the transcriptome data from two different growth modes. A total of 3438 up-regulated genes and 3392 down-regulated genes were detected in the gene expression profiles of wild and cultivated *G. macrophylla* Pall. The expression of 3438 genes in cultivated samples was higher than that in wild samples, with little difference (Fig. 5a) [27].

Genes with similar expression patterns may have the same function. In order to visually display the expression differences of genes in different groups and mine new functional genes, hierarchical clustering analysis was performed on all DEGs screened. The genes with the same or similar expression patterns in different samples were clustered and displayed by heat map (Fig. 5b), and the potential functions of genes with similar expression patterns were discovered according to genes with known functions.

3.2.3 Simple sequence repeat (SSR) analysis

Simple Sequence Repeats (SSR) is one of the effective molecular markers for detecting genetic diversity and constructing genetic map. Through SSR analysis of unigenes larger than 1 KB screened by MISA software, six types of SSRS were identified: (1) mono-nucleoside repeat SSR 5719, (2) di-nucleoside repeat SSR1660, (3) tri-nucleoside repeat SSR2463, (4) tetra-nucleoside repeat SSR148, (5) penta-nucleoside repeat SSR60, and (6) hexa-nucleoside repeat SSR70. In addition, the analysis identified 706 compound repeat simple sequence repeats (SSRs) and 33 overlapping compound type SSRs. A total of 10861 SSRs were characterized, contributing to the genetic diversity and offering potential markers for further study.[27].

3.3 Integrated analysis of the metabolome and transcriptome

Based on metabolomic and transcriptomics data, the KEGG pathway was selected as a vector and integrated analysis was performed. The KEGG analysis showed that the Monoterpenoid biosynthesis (map00902) and Flavonoid biosynthesis (map00941) pathways were significantly different between the two groups, and identical pathways were enriched in the transcriptome and metabolome. According to Fig. 6, there were 37 metabolites and 8 DEGs involved in the metabolic pathways of the two comparison groups. Therefore, these findings are of great significance and provide references for mining key genes in the biosynthesis pathway of secondary metabolites.

Terpenoids (isoprenoids) are a class of important chemicals produced by plants [28]. Iridoids are a special class of monoterpenes. In the synthetic pathway, iridoids are intermediate products in the biosynthetic pathway of indole alkaloids [29]. Iridoids monoterpenoids and Secoiridoid monoterpenoids are one of the main active components of *G. macrophylla*. Therefore, we focused on the analysis and identification of the putative genes in the Monoterpenoid pathway. Loganic acid had the highest correlation with *CYP76F14* ((E)-8-carboxylinalool synthase). Sweroside had the highest correlation with *PGT1* and *HCT* (shikimate O-hydroxycinnamoyltransferase). Gentiopicroside was negatively correlated with *CYP76F14*, *CYP93G1* (flavone synthase II) and *CHS* (chalcone synthase), but positively correlated with *10HGO*, *PGT1* (phlorizin synthase) and *HCT*, but the correlation was weak. Swertiamarin was positively correlated with *PGT1* and *HCT*, but negatively correlated with other differential genes. Quercetin had the highest correlation with *PGT1*. In the monoterpenoid pathway, *10HGO* is particularly important gene. The relative expression of *10HGO* was also higher in the root. Additionally, the relative expression levels of *10HGO* in both wild and cultivated samples were validated using qRT-PCR. The qRT-PCR results

corroborated the expression patterns observed in the transcriptome data. Consequently, these findings hold significant value, providing insights and a reference framework for identifying and studying key genes involved in secondary metabolite biosynthesis pathways.

In addition to ferulate, 4-vinylphenol, and galangin were negatively correlated, and the rest of the shikimates and phenylpropanoids were positively correlated. Among them, the correlation was highest between isovitexin and *PGT1*. Isoorientin had the highest correlation with *CHS*. Loganin and hesperetin had the highest correlation with *CHS*. Sinapyl alcohol and ferulic acid had the highest correlation with *PGT1*. Acacetin had the highest correlation with *CYP93G1*.It suggests that these genes may be correlated with the biosynthesis of shikimates and phenylpropanoids, which are key pathways in the production of secondary metabolites. Additionally, the compound 6'-O-beta-D-glucosylgentiopicroside was found to have a positive correlation with the genes *CYP76F14* and *PGT1*, while showing a negative correlation with other genes. The transcription factors *HCT*, *CYP93G1*, *CHS*, and *PGT1* appear to be implicated in the flavonoid biosynthesis pathway. Their involvement suggests a regulatory role in the expression of genes that contribute to the synthesis of flavonoids, which are significant for the medicinal attributes of the plant.

3.4 Verification of RNA-Seq Sequencing Data by qRT-PCR Analysis

In order to verify the DEGs expression profile obtained by RNA-seq, 6 genes were selected for q PCR verification. The expression levels of six genes were consistent with the RNA-seq data (Fig. 7). This suggests that both the sampling method and RNA-seq used in this study are suitable for studying transcriptome patterns of *G. macrophylla* Pall.

4. Discussion

In this study, the wild population of *G. macrophylla* were used as the control group, while the cultivated population constituted the experimental group. The analysis of the up-regulation and down-regulation of metabolites showed that 15,991 were up-regulated and 9596 were down-regulated. The total up-regulation number was much higher than the down-regulation number. There are more metabolites in cultivation than in the wild. However, a single analysis of its differential metabolites showed that the difference between wild and cultivated was small, flavonoids, phenolic acids, and lignans were higher in cultivation than in wild, while terpenoids, isoflavonoids, stilbenoids, and coumarins were higher in wild than in cultivation. In the analysis of the total genes, it was found that 3438 genes were up-regulated and 3392 genes were down-regulated, that is, the expression level of 3438 genes in cultivated samples was higher than that in wild samples, and the difference was small. A joint analysis of differential metabolites and differential genes revealed that 37 metabolites and 6 DEGs were involved, mainly involving the KEGG pathway for monoterpene biosynthesis (map00902) and flavonoid biosynthesis (map00941) pathway, which was verified by RT-qPCR, and the results were consistent with RNA-seq data, and all are up-regulated.

By combining the metabolic and transcriptome analysis of wild and cultivated *G. macrophylla*, a comprehensive and systematic genetic information of *G. macrophylla* was obtained, which provided data support for the subsequent functional identification of *G. macrophylla* genes [30].

5. Conclusion

In conclusion, the functional annotation analysis revealed that wild *G. macrophylla* shows a heightened response to both biotic and abiotic stress compared to cultivated varieties. This study identified significant metabolic pathways that could guide the screening of germplasm resources and the breeding of superior *G. macrophylla* varieties. Moreover, the detailed annotation of secondary metabolite synthesis pathways provides essential insights for understanding the plant's secondary metabolic processes. These findings highlight the critical need for the conservation of wild *G. macrophylla* species and lay a theoretical foundation for future efforts to improve the quality of cultivated *G. macrophylla*.

Declarations

Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: PRJNA1132749.

Funding

This work was funded by China Agriculture Research System of MOF and MARA (CARS-21), Gansu Province Science and Technology Plan Project: Major Science and Technology Special (23ZDFA013-1), Gansu Provincial Science and Technology Plan Project: Science and Technology Special Touring Special (23CXNA0042), Gansu Provincial Rare Traditional Chinese Medicine Resources Evaluation and Protection and Utilization Engineering Research Center Open Fund (GSZXZY202203), Innovation Center for Traditional Chinese MedicineCo-constructed by Gansu Province & MOE of PRC, Strategic Research and Consulting Project of China Academy of Engineering (GS2021ZDA06), Improvement of traditional Chinese medicine guarantee and innovation capabilities (State Administration of Traditional Chinese Medicine).

Author Contributions

Juanjuan Liu: Methodology, Investigation, Data curation, Visualization, Formal analysis, Writing-original draft. Jialing Zhang: Conceptualization, Methodology, Investigation. Yiyang Chen: Methodology, Investigation, Data curation. Ke Li: Methodology, Supervision, Investigation. Liangcai Wang: Investigation, Software, Data curation. Xiaohui Ma: Resources, Supervision, Investigation. Ling Jin: Conceptualization, Methodology, Funding acquisition, Project administration, Writing-review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by China Agriculture Research System of MOF and MARA (CARS-21), Gansu Province Science and Technology Plan Project: Major Science and Technology Special (23ZDFA013-1), Gansu Provincial Science and Technology Plan Project: Science and Technology Special Touring Special (23CXNA0042), Gansu Provincial Rare Traditional Chinese Medicine Resources Evaluation and Protection and Utilization Engineering Research Center Open Fund (GSZXZY202203), Innovation Center for Traditional Chinese MedicineCo-constructed by Gansu Province & MOE of PRC, Strategic Research and Consulting Project of China Academy of Engineering (GS2021ZDA06), Improvement of traditional Chinese medicine guarantee and innovation capabilities (State Administration of Traditional Chinese Medicine).

References

1. Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China*. vol. Vol (China Medical Science Press, Beijing, 2020).
2. Cheng, Z. *et al.* Extraction optimization, characterization and antioxidant activity of polysaccharide from *Gentiana* scabra bge. *International Journal of Biological Macromolecules* **93**, 369–380 (2016).
3. Liang, G. cheng, Duan, W. gui & Chen, S. yin. Research progress on chemical composition and pharmacological activity of Gentianae Macrophyllae Radix. *Chinese Traditional and Herbal Drugs* 2472–2490 (2024) doi:10.7501/j.issn.0253-2670.2024.07.033.
4. Ramírez-Cisneros, M. Á. *et al.* In vitro COX-1 and COX-2 enzyme inhibitory activities of iridoids from Penstemon barbatus, Castilleja tenuiflora, Cresentia alata and Vitex mollis. *Bioorg Med Chem Lett* **25**, 4505–4508 (2015).
5. Yin, C., Xie, L. & Guo, Y. Phytochemical analysis and antibacterial activity of *Gentiana macrophylla* extract against bacteria isolated from burn wound infections. *Microbial Pathogenesis* **114**, 25–28 (2018).
6. Kou, Y. *et al.* A Comparative Transcriptomic with UPLC-Q-Exactive MS Reveals Differences in Gene Expression and Components of Iridoid Biosynthesis in Various Parts of Gentiana macrophylla. *Genes (Basel)* **13**, 2372 (2022).
7. Duan, B. zhong, Huang, L. fang & Li, W. tao. Macroscopic Identification of Cultivated and Wild Gentiana macrophylla and Preliminary Screening for High Quality Germplasm. *Journal of Chinese Medicinal Materials* 1889–1892 (2012) doi:10.13863/j.issn1001-4454.2012.12.005.
8. Gay, M., Lempereur, L., Francis, F. & Caparros Megido, R. Control of Dermanyssus gallinae (De Geer 1778) and other mites with volatile organic compounds, a review. *Parasitology* **147**, 731–739 (2020).

9. Yaqub, G. *et al.* Biomonitoring of Workers Exposed to Volatile Organic Compounds Associated with Different Occupations by Headspace GC-FID. *Journal of Chemistry* **2020**, e6956402 (2020).
10. Zhang, J. *et al.* Integrated metabolomics and transcriptomics reveal metabolites difference between wild and cultivated *Ophiocordyceps sinensis*. *Food Research International* **163**, 112275 (2023).
11. Tsaballa, A. *et al.* Comprehensive approaches reveal key transcripts and metabolites highlighting metabolic diversity among three oriental tobacco varieties. *Industrial Crops and Products* **143**, 111933 (2020).
12. Peng, H. sheng, Cheng, M. en, Zhang, L., Yao, yong & Han, B. xing. Analysis Odor of Rhizom a Atractylod is Macrocephalae Based on Electronic Nose. *Journal of Chinese Medicinal Materials* **33**, 503–506 (2010).
13. Doppler, M. *et al.* Stable Isotope-Assisted Evaluation of Different Extraction Solvents for Untargeted Metabolomics of Plants. *Int J Mol Sci* **17**, 1017 (2016).
14. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **6**, 1060–1083 (2011).
15. Want, E. J. *et al.* Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* **5**, 1005–1018 (2010).
16. Cai, Y., Weng, K., Guo, Y., Peng, J. & Zhu, Z.-J. An integrated targeted metabolomic platform for high-throughput metabolite profiling and automated data processing. *Metabolomics* **11**, 1575–1586 (2015).
17. Wang, J. *et al.* Serum metabolomics for early diagnosis of esophageal squamous cell carcinoma by UHPLC-QTOF/MS. *Metabolomics* **12**, 116 (2016).
18. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* **78**, 779–787 (2006).
19. Liu, Y. *et al.* Transcriptome and Metabolome Combined to Analyze Quinoa Grain Quality Differences of Different Colors Cultivars. *Int J Mol Sci* **23**, 12883 (2022).
20. Lin, M., Zhou, Z. & Mei, Z. Integrative Analysis of Metabolome and Transcriptome Identifies Potential Genes Involved in the Flavonoid Biosynthesis in Entada phaseoloides Stem. *Front Plant Sci* **13**, 792674 (2022).
21. Zixuan, Z. *et al.* The phenotypic variation mechanisms of Atractylodes lancea post-cultivation revealed by conjoint analysis of rhizomic transcriptome and metabolome. *Plant Physiol Biochem* **203**, 108025 (2023).
22. Xie, C. *et al.* KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res* **39**, W316-322 (2011).
23. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-280 (2004).

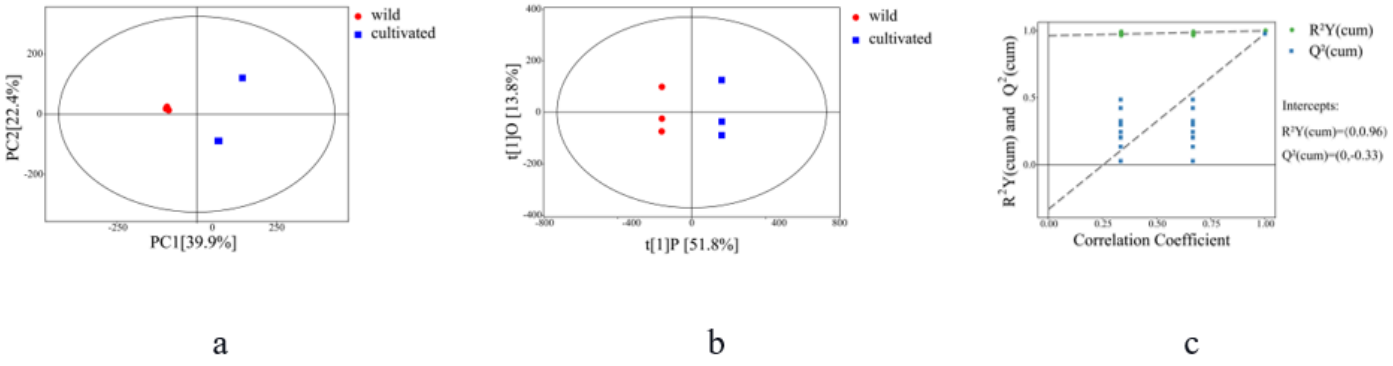
24. S H. *et al.* Integrative Analysis of the Transcriptome and Metabolome Reveals the Developmental Mechanisms and Metabolite Biosynthesis of the Tuberous Roots of Tetrastigma hemsleyanum. *Molecules (Basel, Switzerland)* **28**, (2023).
25. Xiao, J. F., Zhou, B. & Ressom, H. W. Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Analyt Chem* **32**, 1–14 (2012).
26. Sun, J. *et al.* Integrated metabolomics and proteomics analysis to study the changes in Scutellaria baicalensis at different growth stages. *Food Chem* **419**, 136043 (2023).
27. Zhang, T., Wang, M., Li, Z., Wu, X. & Liu, X. Transcriptome analysis and exploration of genes involved in the biosynthesis of secoiridoids in *Gentiana rhodantha*. *PeerJ* **11**, e14968 (2023).
28. Zhao, X. *et al.* Transcriptome and metabolome reveal the accumulation of secondary metabolites in different varieties of Cinnamomum longepaniculatum. *BMC Plant Biology* **22**, 243 (2022).
29. Wang, T. tan. Transcriptional factor analysis of Lonicera japonica Thunb. and its differencial transcriptome analysis with Lonicera macranthoids Hand.-Mazz. (2020) doi:10.27461/d.cnki.gzjdx.2019.001101.
30. Feng, W. meng *et al.* Transcriptomic data analyses of wild and cultivated Angelica sinensis root by high-throughput sequencing technology. *China Journal of Chinese Materia Medica* 1879–1886 (2020) doi:10.19540/j.cnki.cjcmm.20200208.101.

Table

Table 1. Primer sequence of candidate genes used for validation by real-time RT-qPCR

Gene	Primer sequence (5'—3')
CYP76F14	Forward: AGTCATCGGGAAAGGCAAGG
	Reverse: TTCTTTGATGGCACACCGGA
10HGO	Forward: CAGGGAGTGGTATTGGAGGG
	Reverse: GCTCCATTGCAGTGTTCACAT
CYP93G1	Forward: ACGTTTCGTCTTCACCCTCC
	Reverse: ATCCAGCCACTTGGATGTCG
PGT1	Forward: TGAAGTATGAAGACCGCCCG
	Reverse: ATACTTCCTTCACCGCTCCG
HCT	Forward: GTAAAAACTCGGCCCAAGCG
	Reverse: CCCCCATAAGCAACCCCATT
CHS	Forward: TCCGAACCGACTATCAACGC
	Reverse: CGTGGACCGAGTGAAACAGA

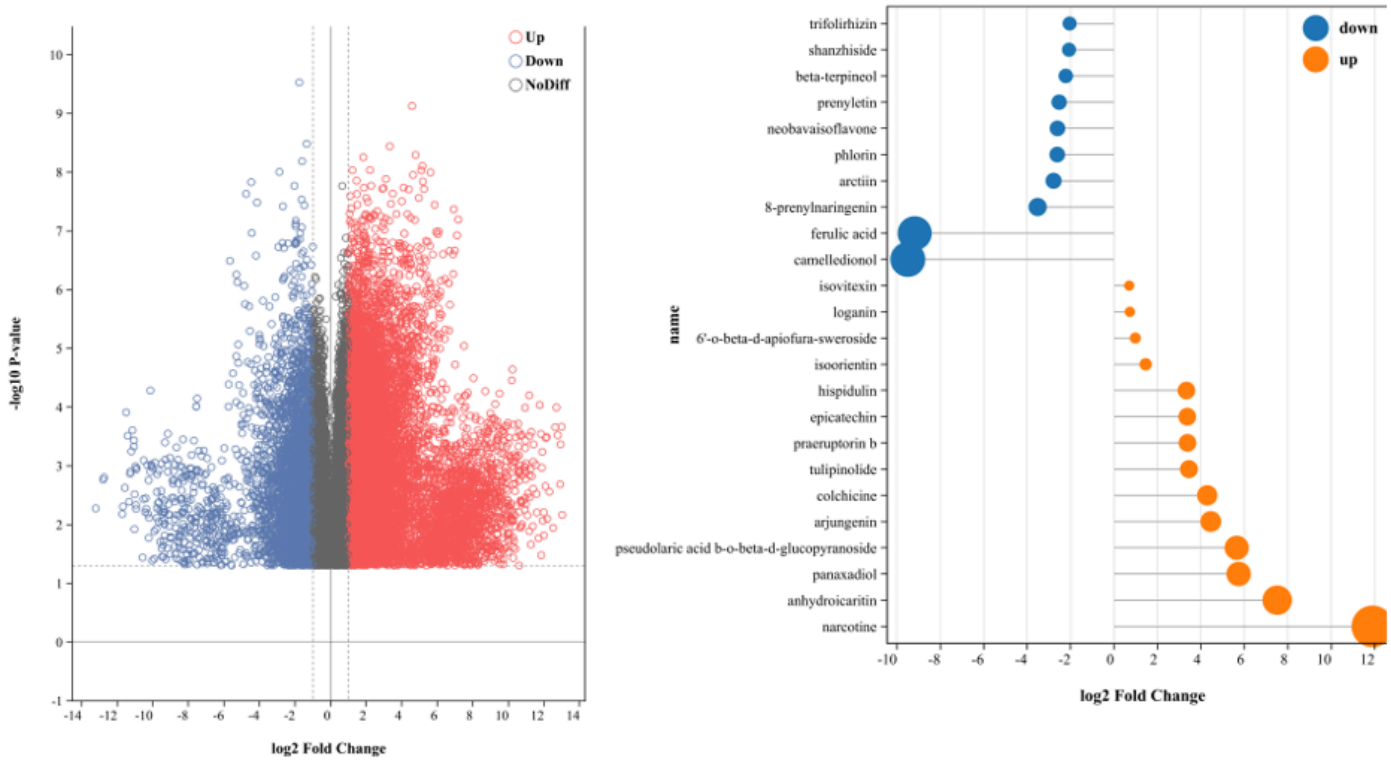
Figures



a b c

Figure 1

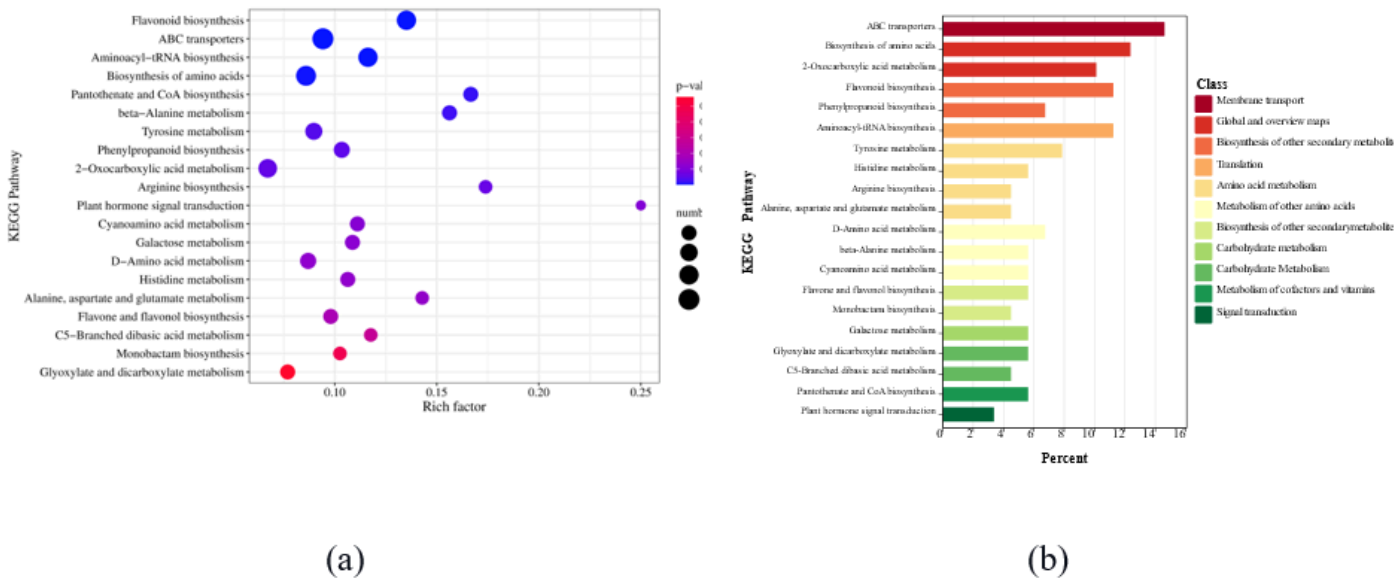
(a) PCA plot of differential metabolites. (b) OPLS-DA plot of differential metabolites. (c) Permutation plot test of OPLS-DA model.



a b

Figure 2

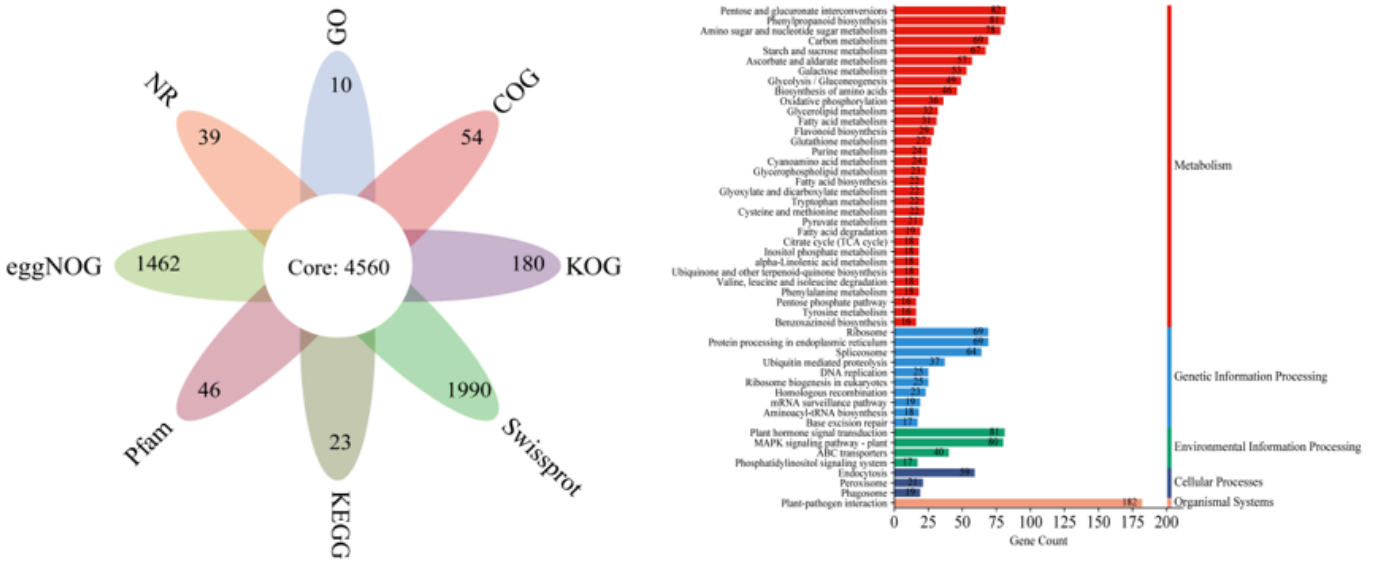
(a) Volcano plot of metabolites. (b) matchstick analysis of metabolites.



(a) (b)

Figure 3

(a) enrichment analysis of differentially accumulated metabolite pathways. (b) KEGG classification diagram of differential metabolites



(a) (b)

Figure 4

(a) All database petal figure (Each petal represents an annotated gene unique to the database, and the flower center represents an annotated gene shared by all databases). (b) KEGG annotated genes

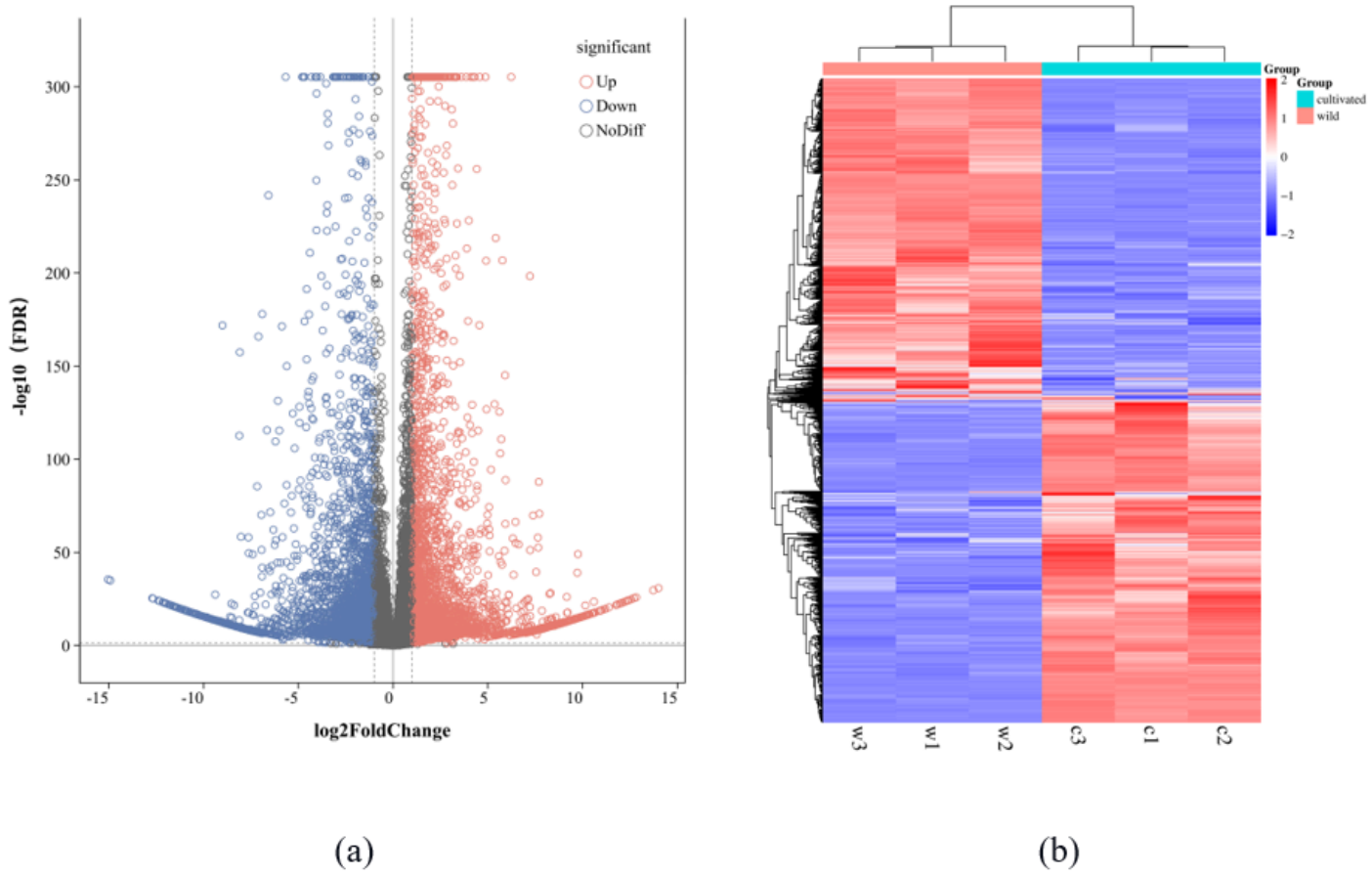


Figure 5

(a) Differentially expressed gene volcano map. (b) Heatmap of Differential Gene Clustering

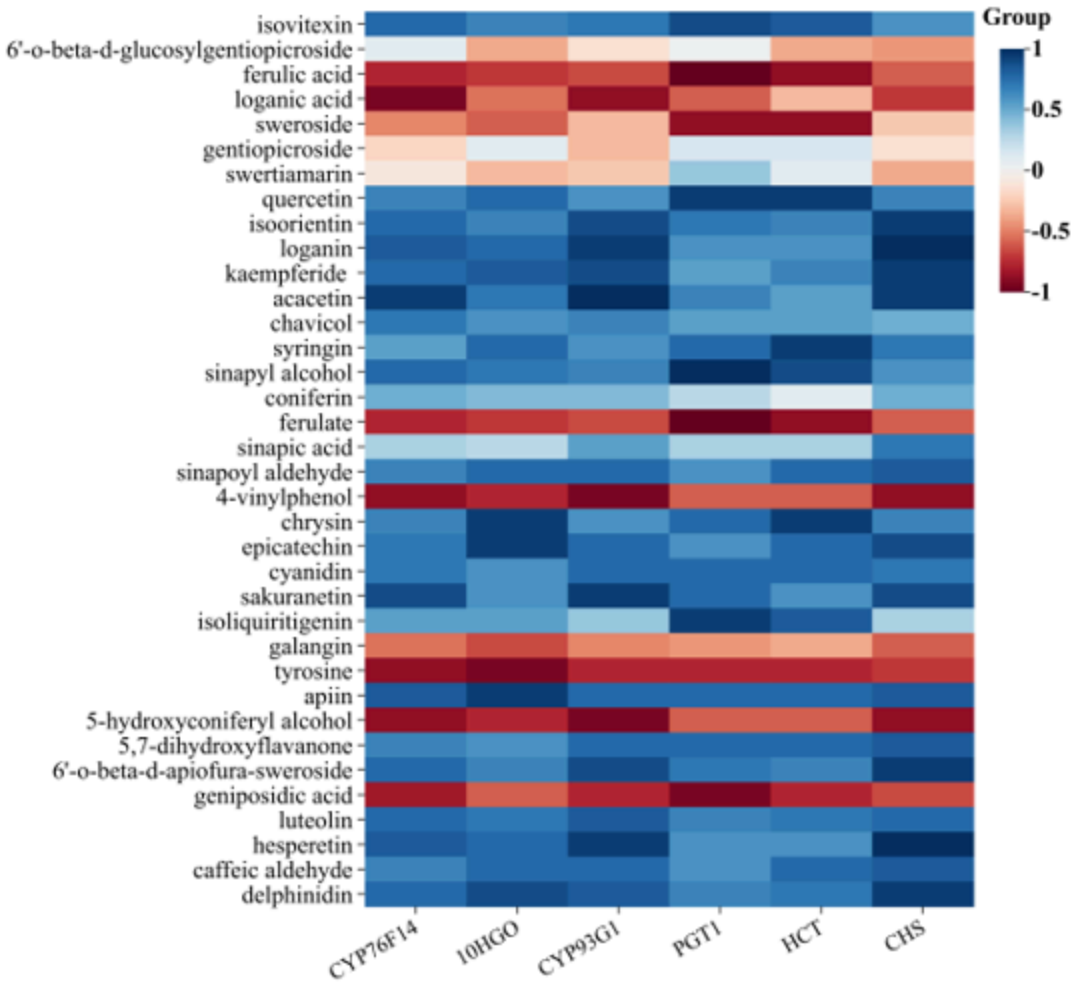


Figure 6

Correlation analysis of differential genes and differential metabolites

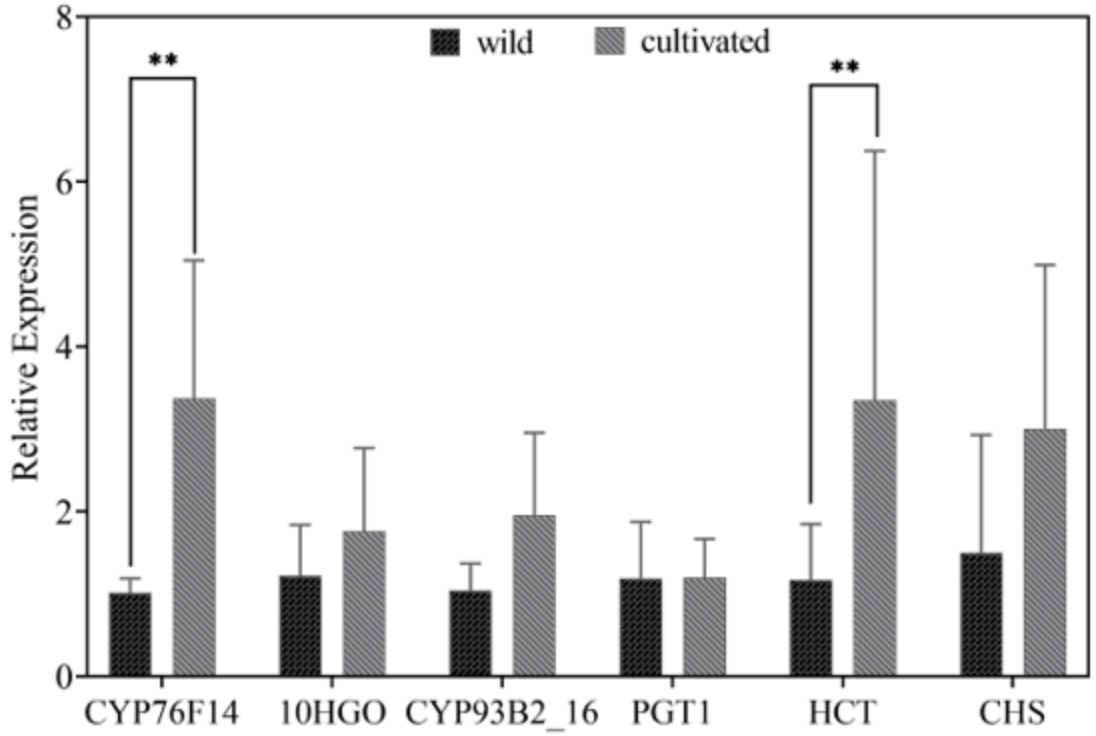


Figure 7

Verification of RNA-Seq sequencing data by the qRT-PCR assay, P≤0.05