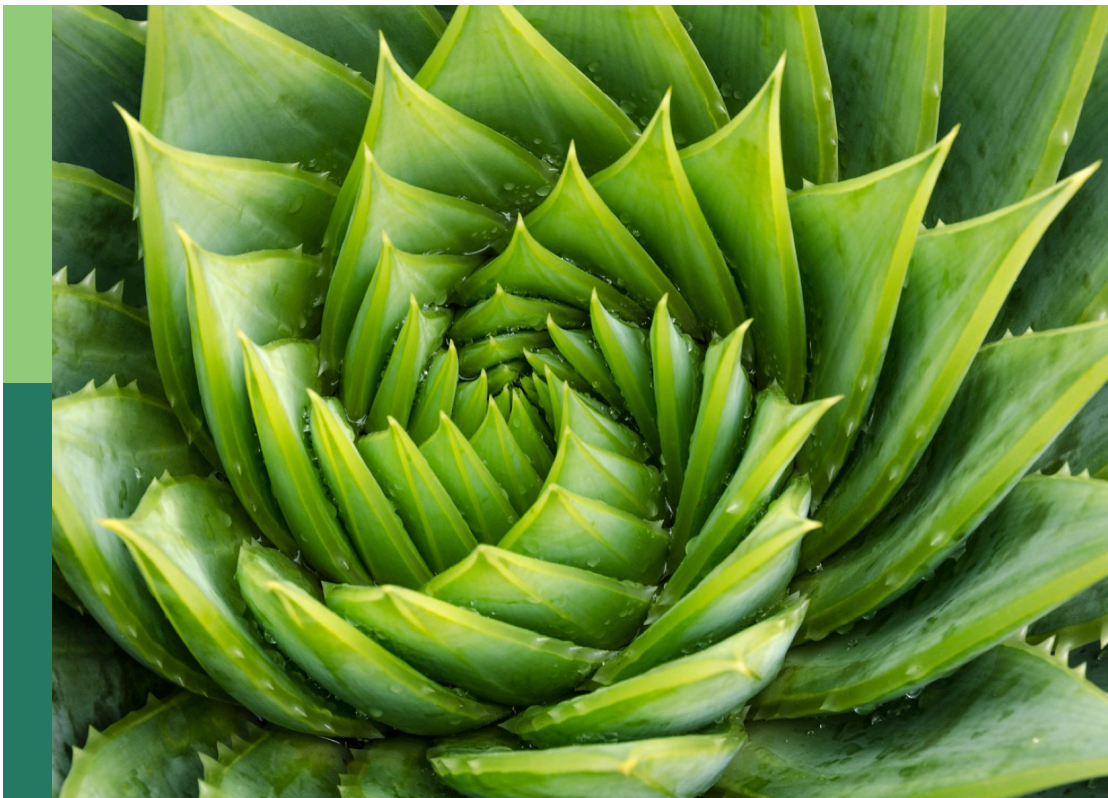
frontiers | Research Topics

Omics data-based identification of plant specialized metabolic genes

Edited by
Peipei Wang, Pengxiang Fan, Yan Bao, Li Wang and Wei Li

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-2803-7
DOI 10.3389/978-2-8325-2803-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Omics data-based identification of plant specialized metabolic genes

Topic editors

Peipei Wang — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

Pengxiang Fan — Zhejiang University, China

Yan Bao — Shanghai Jiao Tong University, China

Li Wang — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

Wei Li — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

Citation

Wang, P., Fan, P., Bao, Y., Wang, L., Li, W., eds. (2023). *Omics data-based identification of plant specialized metabolic genes*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2803-7

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 09 June 2023
DOI 10.3389/fpls.2023.1209334

OPEN ACCESS

EDITED AND REVIEWED BY
Yingzhen Kong,
Qingdao Agricultural University, China

*CORRESPONDENCE
Peipei Wang
peipeiw@msu.edu
Pengxiang Fan
pxfan@zju.edu.cn
Yan Bao
yanbao@sjtu.edu.cn
Wei Li
liwei11@caas.cn
Li Wang
wangli03@caas.cn

RECEIVED 20 April 2023
ACCEPTED 03 May 2023
PUBLISHED 09 June 2023

CITATION
Wang P, Fan P, Bao Y, Li W and Wang L (2023) Editorial: Omics data-based identification of plant specialized metabolic genes.
Front. Plant Sci. 14:1209334.
doi: 10.3389/fpls.2023.1209334

Editorial: Omics data-based identification of plant specialized metabolic genes

Peipei Wang[1,2]*, Pengxiang Fan[3]*, Yan Bao[4]*, Wei Li[1,2]* and Li Wang[1,2]*

[1]Traditional Chinese Medicine & Floriculture Research Center, Kunpeng Institute of Modern Agriculture at Foshan, Foshan, Guangdong, China, [2]Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, China, [3]Department of Horticulture, Zhejiang University, Hangzhou, China, [4]Shanghai Collaborative Innovation Center of Agri-Seeds, Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China

KEYWORDS

specialized metabolic genes, metabolome, transcriptome, plant natural products, evolutionary history

Editorial on the Research Topic
Omics data-based identification of plant specialized metabolic genes

Plant natural products, especially specialized metabolites, are major sources of nutrients, medicines and industrial materials. Identifying genes responsible for the biosynthesis of plant natural products has crucial significance but also has always been a recognized challenge in synthetic biology, due to the lineage-specific distribution and the fast evolution of plant natural products, the complex evolutionary history of underlying genes, the interweaving network of metabolic pathways, and the time-consuming and laborious experient processes, etc. In this Research Topic, 97 authors contributed 12 original research and review manuscripts, primarily focused on the identification of plant specialized metabolic genes via integrating multiple omics data.

Transcriptome and metabolome are dynamic and closely related. Comparing and integrating these two omics data has been widely used to identify key genes responsible for the biosysthesis of specialized metabolites. Phthalides from roots of the medicinal herb *Angelica sinensis* are the main chemical components for promoting blood circulation. By integrating metabolome and transcriptome from roots of two groups of *A. sinensis*, Feng et al. proposed the reaction pathway for phthalide biosynthesis, identified and validated six enzyme genes involved in this biosynthesis pathway. Li et al. used a similar strategy to compare volatile metabolome and transcriptome between two varieties of *Lonicera japonica* and gained insights of floral scents regulation in these two varieties. Yang et al. examined the different cumulation of benzylisoquinoline from different tissues of *Sinomenium acutum*, a medicinal plant used for treating rheumatoid arthritis for hundreds of years, and identified candidate genes responsible for benzylisoquinoline alkaloid biosynthesis via integrating the metabolome and full-length transcriptome data. Wang et al. selected two varieties of the high-value nut crop, *Macadamia integrifolia*, each exhibiting distinct floral traits, such as flower coloration and aroma formation. They revealed a metabolic network integrating genes associated with hormone signal transduction, starch and phenylpropanoid metabolism, which play roles in the development of flower coloration and aroma.

Aiming to reveal the formation mechanism of red stem in *Prunus mume*, a graceful horticultural plant known for its varied colors and postures, Qiu et al. identified metabolites in anthocyanin-related pathways, particularly cyanidin glycoside and paeoniflorin glycoside, that are only accumutated in the "Wuyuyu" accession with red stem rather than in the "Fei Lve" accession with green stem. They also identified several genes which are potentially involved in anthocyanin metabolic pathways for the red pigment formation using transcriptome data. Liu et al. identified several candidate genes for the biosynthesis of gastrodin, a main bioactive ingredient of a medicinal plant *Pholidota chinensis* Lindl., which is used to treat high blood pressure, dizziness and headache.

Lotus plumule is a green tissue in the middle of seeds that predominantly accumulates bisbenzylisoquinoline alkaloids and chlorophyll. Sun et al. identified potential enzyme genes responsible for bisbenzylisoquinoline alkaloids and chlorophyll biosynthesis in Lotus plumule by comparing time series profiles of these two types of metabolites and the transcriptome in this tissue. By conducting gene family evolution analysis and examining gene expression pattern in different tissues and under UV-B treatment, Ren et al. identified two flavonoid glucosyltransferases responsible for flavonoid glucosylation in Chinese bayberry *Morella rubra*. To elucidate the metabolic and transcriptomic basis for toon bud deterioration after harvest, Zhao et al. explored the metabolic regulation of *Toona sinensis* toon buds in different postharvest storage temperatures via metabolic profiling and gene expression analysis, and identified several metabolic pathways whose changes after harvest might contribute to the toon bud deterioration.

Tocopherols are widely recognized as vitamin E, which is an essential nutrient in the human diet. As a lipid-soluble antioxidant, vitamin E plays a broad and fundamental role in controlling seed longevity and viability, high-light acclimation, and cold response, among other functions. In a recent study, Niu et al. summarized recent progresses in the biosynthesis and response of vitamin E. Key discoveries in recent years include the identification of the seed-specific α/β hydrolase VTE7 for tocopherol biosynthesis, the genetic connection between vitamin E metabolism and miRNA biogenesis, and the critical role of vitamin E in mediating plant cold response. This review provides valuable information for tracking the progresses of vitamin E synthesis and signaling pathways.

The exploration of evolutionary history of genes involved in the secondary metabolite biosynthetic pathways can provide insights into the biological processes in land plants. At the macro-evolutionary scale, Kruse et al. investigated the patterns of sequence and expression evolution of BAHD acyltransferase family in land plants. They found that the gene family expansions are concordant with the prominence of metabolite classses and most co-expressed BAHDs in rice and *Arabidopsis* belong to distinct clades. At the micro-evolutionary scale, Marszalek-Zenczak et al. unveiled the evolution of four metabolic gene clusters (MGCs) in *Arabidopsis thaliana* populations covering around 1,000 accessions. Marneral and tirucalladienol MGCs are rather conserved, while thalianol and especially arabidiol/baruol MGCs display profound diversity among accessions. The arabidiol/baruol MGC contained divergent duplicates of both *CYP705A2* and *BARS1* genes in one-third of accessions, which was correlated with the root growth dynamics and adaptation to climate changes. These two studies indicate that the evolutionary history of metabolic genes sheds light on the gene functional prediction and phenotypic diversity of plants.

In summary, this Research Topic features diverse research efforts and methodologies aimed at identifying plant specialized metabolic genes, through integrating and comparing multiple omics data types. These studies have uncovered key genes involved in specialized metabolite biosynthesis across a wide variety of plants, providing insights into the molecular mechanisms that underlie the production of valuable plant-derived compounds. The knowledge gained from this Research Topic has the potential to pave the way for enhancing production and utilization of these valuable compounds in various applications, spanning the nutritional, pharmaceutical, and ornamental industries.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

Research in the lab of PW is supported by the Scientific Research Foundation for Principle Investigator, Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003 to PW) and the Funding of Major Scientific Research Tasks, Kunpeng Institute of Modern Agriculture at Foshan (KIMA-ZDKY2022004 to PW). Research in the lab of PF is supported by Natural Science Foundation of Zhejiang pvovince, China (Grant No. LZ22C150005) and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0011). Research in the lab of YB is supported by the Natural Science Foundation of Shanghai (23ZR1427700) and China Agriculture Research System of MOF and MARA. Research in the lab of WL is supported by National Natural Science Foundation of China (32170264), the National Key R&D Program of China (grant no. 2020YFA0907900 and 2022YFD1700200) and Science Technology and Innovation Commission of Shenzhen Municipality of China (ZDSYS 20200811142605017). Research in the lab of LW is supported by the National Natural Science Foundation of China (Grant No. 32070242) and Science Technology and Innovation Commission of Shenzhen Municipality of China (ZDSYS 20200811142605017).

Acknowledgments

Thanks to all the authors contributing the 12 manuscripts to this Research Topic and all the reviewers for their critical comments and valuable suggestions for these manuscripts.

Conflict of interest

The 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.

Publisher's note

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 06 May 2022
doi: 10.3389/fpls.2022.884883

Integration of Transcriptome and Metabolome Reveals the Formation Mechanism of Red Stem in *Prunus mume*

Like Qiu, Tangchun Zheng, Weichao Liu, Xiaokang Zhuo, Ping Li, Jia Wang, Tangren Cheng and Qixiang Zhang**

Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Engineering Research Center of Landscape Environment of Ministry of Education, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education, School of Landscape Architecture, Beijing Forestry University, Beijing, China

Prunus mume var. purpurea, commonly known as "Red Bone", is a special variety with pink or purple-red xylem. It is famous due to gorgeous petals and delightful aromas, playing important roles in urban landscaping. The regulation mechanism of color formation in *P. mume* var. purpurea stem development is unclear. Here, we conducted a comprehensive analysis of transcriptome and metabolome in WYY ('Wuyuyu' accession, red stem) and FLE ('Fei Lve' accession, green stem), and found a total of 256 differential metabolites. At least 14 anthocyanins were detected in WYY, wherein cyanidin 3,5-O-diglucoside and peonidin3-O-glucoside were significantly accumulated through LC-MS/MS analysis. Transcriptome data showed that the genes related to flavonoid-anthocyanin biosynthesis pathways were significantly enriched in WYY. The ratio of dihydroflavonol 4-reductase (*DFR*) and flavonol synthase (*FLS*) expression levels may affect metabolic balance in WYY, suggesting a vital role in xylem color formation. In addition, several transcription factors were up-regulated, which may be the key factors contributing to transcriptional changes in anthocyanin synthesis. Overall, the results provide a reference for further research on the molecular mechanism of xylem color regulation in *P. mume* and lay a theoretical foundation for cultivating new varieties.

Keywords: *Prunus mume*, transcriptome, metabolome, flavonoid biosynthesis, anthocyanin biosynthesis, red stem, xylem color

OPEN ACCESS

Edited by:
Yan Bao,
Shanghai Jiao Tong University, China

Reviewed by:
Haihai Wang,
Shanghai Institutes for Biological Sciences (CAS), China
Nan Jiang,
Michigan State University, United States

****Correspondence:***
Tangchun Zheng
zhengtangchun@bjfu.edu.cn
Qixiang Zhang
zqxbjfu@126.com

Specialty section:
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

Received: *27 February 2022*
Accepted: *25 March 2022*
Published: *06 May 2022*

Citation:
Qiu L, Zheng T, Liu W, Zhuo X, Li P, Wang J, Cheng T and Zhang Q (2022) Integration of Transcriptome and Metabolome Reveals the Formation Mechanism of Red Stem in Prunus mume. Front. Plant Sci. 13:884883. doi: 10.3389/fpls.2022.884883

INTRODUCTION

Color formation is one of the main quality traits of ornamental plants. The most widespread non-green pigments in flowers, fruits and various organs are generally classified into three categories: anthocyanins, carotenoids, and betalains (Boldt et al., 2014). Anthocyanins, a class of secondary metabolites, are important water-soluble pigments that are widely accumulated in vascular plants and the different substituents on the B ring of flavonoid basic skeleton result in a variety of colors ranging from red to blue (Tanaka et al., 2008). Besides providing beautiful pigmentation in flowers to attract pollinators and seed spreaders, anthocyanins also play a key role in signal transmission between plants and microorganisms (Harborne and Williams, 2000), photoprotection

during photosynthesis (Hughes et al., 2005), antioxidant activity (Wei et al., 2018), and UV protection (Jansen et al., 1998). Recent studies have indicated that anthocyanins can ameliorate drug-induced cognitive deficits (Jo et al., 2015) and anthocyanin-rich diets are associated with decreased cardiovascular diseases and mortality (Isaak et al., 2017). Therefore, on account of the functional diversity, anthocyanin have attracted much more attention, becoming a research hotspot in the field of secondary metabolism of horticultural crops.

So far, more than 700 compounds have been found in plants, which mainly derive from six anthocyanidins aglycones (Celli et al., 2018). Anthocyanin biosynthesis is rather conservative in plants. It occurs on the cytoplasmic surface of the endoplasmic reticulum. The biosynthetic pathway has been extensively studied in *Arabidopsis thaliana* (Martens et al., 2010), *Petunia hybrida* (Jonsson et al., 1984), *Zea Mays* (Harborne and Gavazzi, 1969) and *Dianthus* (Stich and Wurst, 1992). In the first step, phenylalanine is catalyzed by phenylalanine ammonia-lyase (PAL) to produce 4-coumaroyl-CoA and 3 malonyl-CoA. In the second stage, dihydroflavonols are generated from coumaroyl-CoA under the catalysis of chalcone synthase (CHS), chalcone isomerase (CHI) and flavanone 3-hydroxylase (F3H), which is a key step in the metabolism of flavonoids. Flavonoid 3′-hydroxylase (F3′H) and flavonoid 3′,5′-hydroxylase (F3′5′H) catalyze the hydroxylation of dihydrokaempferol to form dihydroquercetin and dihydromyricetin, and determine the hydroxylation pattern of flavonoid and anthocyanin B ring, which are important enzymes for the synthesis of cyanidin and delphinidin (Zhuang et al., 2019). Subsequently, under the action of dihydroflavonol 4-reductase (DFR), dihydroflavonols are reduced to the corresponding 3,4-cis-leucoanthocyanidins, which are then catalyzed by anthocyanidin synthase (ANS) to form colored anthocyanidins. Unstable colored anthocyanins are modified by glycosylation, methylation, and acetyltransferase to form different types of stable anthocyanin polymers that give plants different colors. The high expression level of *DFR*, *ANS* and *anthocyanidin 3-O-glucosyltransferase* (*UFGT*) genes in downstream of the anthocyanin biosynthetic pathway tends to promote flower or fruit coloring (Wang et al., 2017). The *UFGT* has been identified as a key gene for determining the concentration and accumulation of anthocyanins, causing red color petals in *Prunus mume* (Wu et al., 2017). Silencing of *FaDFR* in strawberry decreases the anthocyanin content, rendering pale fruit color (Lin et al., 2013).

In addition to the above-mentioned structural genes, transcription factors can activate or inhibit the temporal and spatial expression of structural genes through specific proteins, thus affecting the intensity and pattern of anthocyanin biosynthesis. Many studies have shown that the synthesis of anthocyanins is regulated by a protein complex formed by a MYB transcription factor, a basic helix-loop-helix protein (bHLH), and a WD-repeat protein, which bind to the promoters of structural genes to induce their transcription (Koes et al., 2005). In the model plant *Arabidopsis thaliana*, the AtTT2-AtTT8-AtTTG1 transcription complex can promote the expression of *DFR*, *ANS*, and *TT19* genes in the proanthocyanidin synthesis pathway (Xu et al., 2013) and AtPAP1-AtTT8/GL3-AtTTG1 can also actuate the expression of structural genes in the anthocyanin synthesis pathway (Tohge et al., 2005; Gonzalez et al., 2008). In petunia, PhAN11-PhAN1-PhAN2 regulates anthocyanin synthesis in petals by adjusting the expression of *DFR* and *CHS*, while AN11-AN1-AN4 regulates anther development (Quattrocchio et al., 2006). In *Gerbera hybrida*, *GMYB10* interacts with bHLH factor *GMYC1* to superintend anthocyanin synthesis in leaves, floral stems, and flowers by actuating the late gene *DFR* (Elomaa et al., 2003). The *GtMYB3* transcription factor and *GtbHLH1* work synergistically to affect the expression of the *F3'5'H* gene, which in turn promotes the enrichment of gentiodelphin in *Gentian trifloral* (Nakatsuka et al., 2008). The *MdMYB10* induces cyanidinin accumulation and slightly increases the expression of *DFR* in apples with overexpression of *MdMYB10*. Meanwhile, *MdMYB10* co-expressed with *MdbHLH3* and *MdbHLH33* promote the biosynthesis of anthocyanins (Espley et al., 2007).

Prunus mume (also named mei), a member of the Rosaceae family, is a traditional flower species in China that has high ornamental value due to its colorful corolla, wispy fragrance, and varying flower types. With the rapid development of high-throughput sequencing technology, the whole genome sequencing of *P. mume* was completed in 2012, which provided important data support for revealing the biological characteristics of bud dormancy, floral scent, and plant architecture. However, there are few systematic studies on the formation mechanism of wood color in *P. mume*, and the metabolic pathway and molecular regulatory mechanism of wood color are still unclear. Transcriptome and metabolome are combined to identify and analyze the interaction of single and multiple genes in metabolic pathways in many plants, including potato (Cho et al., 2016), grape hyacinth (Lou et al., 2014), turnip (Zhuang et al., 2019), and crape myrtle (Qiao et al., 2019), providing a powerful tool to analyze the mechanisms of plant tissue-specific metabolism and secondary metabolism. Combined transcriptome and metabolome studies can not only detect the abundance of transcripts, but also provide a new perception of the flow of metabolism.

Here, we used ultra-performance liquid chromatography and mass spectrometry to survey the differences of metabolites conferring the red stems of *P. mume* cultivar 'Wuyuyu' compared to 'Fei Lve', which feature red and green stems, respectively. Besides, we employed RNA-seq technique to identify the key candidate genes involved in anthocyanin metabolism of differential pigmentation and then verified by quantitative real-time polymerase chain (qRT-PCR). The outcomes of the study may provide valuable information for understanding the molecular mechanism of the wood color formation at the transcriptomic and metabolomic levels in *P. mume*.

RESULTS

Metabolome Profiling of LC-MS/MS Data

Two *P. mume* cultivars with different stem colors ['Fei Lve' (FLE) and 'Wuyuyu' (WYY)] were chosen for this study (**Figure 1A**). To evaluate the components of FLE and WYY, widely-targeted metabolomics was used to analyze the metabolic profiles based on UPLC-MS/MS. A total of 778 metabolites

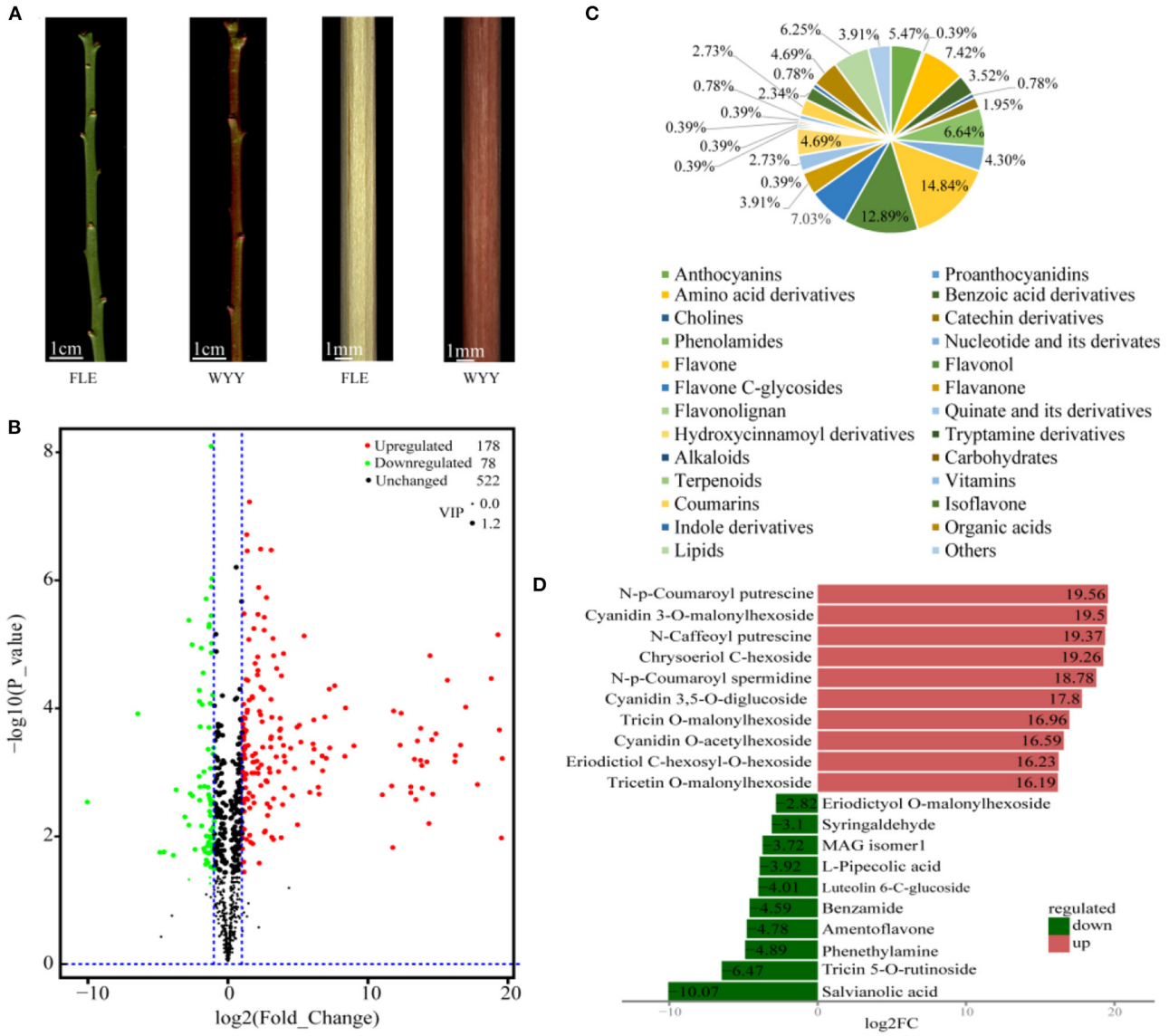


FIGURE 1 | Morphological characterization and differentially accumulating metabolites between WYY and FLE. **(A)** Green-stem cultivar 'Fei Lve' (FLE) and red-stem cultivar 'Wuyuyu' (WYY). **(B)** Volcano plot of the 778 metabolites identified. Differential metabolites were defined with fold change ≥ 2 or ≤ 0. 5, and VIP ≥ 1. **(C)** Pie chart depicting the biochemical categories of the differential metabolites identified between WYY and FLE. **(D)** The top 10 significantly differentially expressed metabolites.

grouped into 26 classes were identified from FLE and WYY samples and the most abundant metabolites were the flavonoids (**Supplementary Table S1**).

Principal component analysis (PCA) revealed the overall metabolic differences between WYY and FLE samples. It can be clearly seen that the two groups were separated into distinct clusters on the PC1 × PC2 score plot, where the variances of PC1 and PC2 were 83.11%, 5.23%, respectively (**Supplementary Figure S1**). Multivariate statistics of the metabolite concentration data was performed to access the differences in sample accumulation patterns. These two groups could be easily distinguished from each other on the heatmap, indicating significant biochemical differences in metabolites between WYY and FLE (**Supplementary Figure S2**).

Differentially Accumulated Metabolite Analysis Based on OPLS-DA

PCA is insensitive to the variables with small correlation, while the orthogonal projections to latent structures-discriminant analysis (OPLS-DA) can maximize the discrimination between groups, which is conducive to searching for differential metabolites. We used OPLS-DA models to filter differential compounds between two groups of samples by removing irrelevant differences. As described in materials and methods, the

variable importance in projection (VIP) values and the *P*-values of univariate statistical *t*-test were used to screen the metabolites with significant differences. Furthermore, the high R2 and Q2 values indicated that the multivariate model had good quality and predictive ability (**Supplementary Figure S3**). Overall, we putatively identified 256 differentially accumulated metabolites (DAMs), including 78 down-regulated and 178 up-regulated metabolites in WYY compared with FLE (**Figure 1B** and **Supplementary Table S2**). The DAMs can be divided into more than 20 different categories, and flavonols, amino acids, amino acid derivatives, lipids, and anthocyanins were significantly different between the two cultivars (**Figure 1C**). The most up-regulated metabolites were N-p-Coumaroyl putrescine, followed by cyanidin 3-O-malonylhexoside and N-caffeoyl putrescine. The top three down-regulated metabolites were salvianolic acid, tricin 5-O-rutinoside, and phenethylamine (**Figure 1D**).

To further understand the biological classification and pathways of these metabolites, the DAMs were assigned to KEGG database for enrichment analysis. Notably, the relative abundance of differential metabolites between FLE and WYY was mainly related to metabolic and secondary metabolites biosynthesis, including phenylpropanoid biosynthesis, flavonoid biosynthesis and anthocyanin biosynthesis (**Supplementary Figure S4**). Flavonoids occupied a large proportion in the composition of plant pigments.

Metabolites in the Anthocyanin Biosynthetic Pathway

Anthocyanins are water-soluble pigments existing in the vacuoles of cells that absorb different wavelengths of light and exhibit different colors. Twenty-four anthocyanins were identified in all samples, among which 14 were differentially expressed (**Table 1**). In general, the content and species of anthocyanins in the stems of WYY were higher than those of FLE. Quantitative profiles showed that cyanidin 3-O-malonylhexoside, cyanidin 3,5-O-diglucoside, cyanidin O-acetylhexoside, pelargonidin 3-O-malonylhexoside, pelargonidin O-acetylhexoside, which are the main source of the reddish-purple color in the plant kingdom, were only detected in WYY and their contents were significantly higher than FLE, while the contents of the other nine anthocyanins varied between the two cultivars. Anthocyanins of the same species were detected in the two cultivars, but their expression levels varied significantly. Notably, the level of peonidin in FLE was slightly higher than that in WYY, suggesting anthocyanin accumulation in the green stem.

TABLE 1 | Differentially expressed anthocyanins in WYY and FLE.

Metabolite name	Ion abundance	Ion abundance	log_2FC	VIP
	FLE	WYY	WYY/FLE	
Cyanidin 3-O-malonylhexoside	N/A	6.68E+06	19.50	1.20
Cyanidin 3,5-O-diglucoside	N/A	2.05E+06	17.80	1.28
Cyanidin O-acetylhexoside	N/A	8.88E+05	16.59	1.30
Pelargonidin 3-O-malonylhexoside	N/A	9.04E+04	13.29	1.29
Pelargonidin O-acetylhexoside	N/A	1.84E+04	11.00	1.27
Peonidin O-hexoside	3.81E+05	1.27E+08	8.37	1.31
Cyanidin O-syringic acid	1.93E+05	6.16E+07	8.32	1.29
Cyanidin 3-O-rutinoside	3.56E+05	6.90E+07	7.60	1.31
Pelargonidin 3-O-beta-D-glucoside	4.42E+05	6.53E+07	7.21	1.31
Cyanidin 3-O-glucoside	7.72E+04	9.81E+06	6.99	1.29
Delphinidin	1.43E+04	9.50E+05	6.05	1.29
Delphinidin 3-O-rutinoside	2.41E+03	7.40E+04	4.94	1.21
Malvidin 3,5-diglucoside	5.09E+04	1.11E+05	1.12	1.04
Peonidin	6.72E+05	2.06E+05	−1.70	1.26

HPLC-MS Analysis of Pigmentation in Stems

To further confirm the role of anthocyanins in stem pigmentation, the contents of anthocyanins in several *P. mume* cultivars were determined. In addition to WYY and FLE, we selected two other cultivars, 'Fenhong Zhusha' (FHZS, red stem) and 'Zaohua Lve' (ZHLE, green stem) for quantitative determination. Anthocyanins in the red and green samples were identified and relatively quantified by HPLC-MS based on retention time and mass spectrometry. The content of anthocyanins in samples was calculated by the percentage of peak area between sample solution and standard solution.

HPLC showed that the maximum absorption wavelength of WYY and FHZS was 520 nm, which is the characteristic peak of anthocyanins pigments that is different from other flavonoid compounds, while FLE and ZHLE did not detect the specific peak of anthocyanins. Since WYY and FHZS belong to the same cultivar group, we found the same anthocyanin species between them, but the amount of anthocyanins varied greatly. There were five obvious peaks in WYY extractions with retention times of 21.62, 27.72, 31.97, 36.31, and 41.49 min, respectively (**Figure 2A**). These peaks were further identified by mass spectrometry (**Table 2**). Quantification of these peaks indicated that both cultivars contained three typical red-purple anthocyanin compounds: 70.82% for Cy, 16.47% for Pn and 12.71% for Dp in WYY; and 68.39% for Cy, 20.89% for Pn and 10.72% for Dp in FHZS (**Figure 2C**). This shows that Cy3G5G and Cy3Ru are the most important anthocyanins in the red stem formation of WYY and FHZS. Since anthocyanins were detected in FLE in the metabolome, the absolute contents of Cy and Pn in the four cultivars were tested to verify the correctness of the metabolome data (**Figure 2B**). Between red stem cultivars, FHZS had the highest anthocyanin accumulation level, which was 615 μg/g of fresh weight (FW). However, the content in WYY (265 μg/g of FW) was much higher than FLE (1.29 μg/g of FW). Only trace amounts of anthocyanins were detected in FLE and ZHLE, indicating that different statistical methods may produce certain errors in the results under the condition of too low content (**Figure 2B**).

Global Analysis of RNA-seq Data

To investigate the difference of gene expression level in anthocyanin metabolism of WYY, total RNA isolated from the stems of FLE and WYY were sequenced for transcriptomic

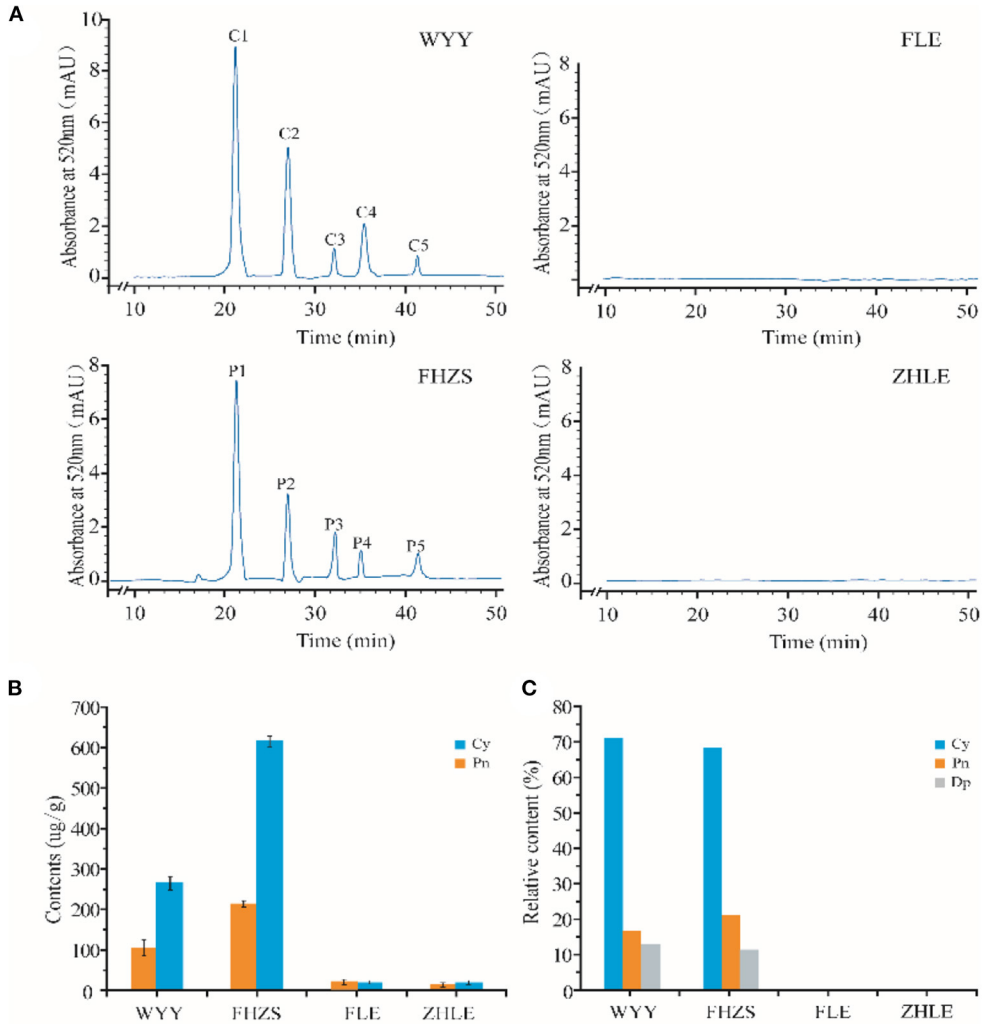


FIGURE 2 | HPLC-MS analysis of pigmentation in stems. **(A)** Results of anthocyanins composition in mei cultivars with different stem colors by HPLC-MS analysis. **(B)** The absolute quantification of anthocyanins content in four mei cultivars. **(C)** The relative quantification of anthocyanins content in four mei cultivars. WYY and FHZS were red stem cultivars, FLE, and ZHLE were green stem cultivars.

analysis, which is consistent with samples used for metabolomic analysis. A total of 61.01 Gb clean data was generated with average 6.52 Gb per sample, and the percentage of Q30 base rates was more than 92.86% and the rate of total mapping ranged from 85.88% to 91.19%. A total of 15,933 and 15,458 core genes were found in the FLE and WYY groups, respectively (**Figures 3A,B** and **Supplementary Table S3**). The PC1 displays the distinct separation between red-colored and green-colored samples, which is consistent with the results of the metabolome. All the assembled unigene sequences were aligned to the public databases using the BLAST program for gene function annotations and protein prediction. Finally, 1,631 novel genes were discovered, of which 1,424 had functional annotations.

GO and KEGG Analysis of DEGs

To identify alterations in gene expression levels, differential gene screening was performed based on a false discovery rate (FDR)$\leq$0.01 and |log2 (foldchange)|$\geq$1. The Pearson's correlation coefficient analysis was used to evaluate the reproducibility of the differential gene expression library. The FLE4 (T08) library had a poor correlation with others (**Supplementary Figure S5**). To improve the repeatability between samples, the T08 library was removed and the heatmap clustering analysis was performed again (**Figure 3E**). Finally, a total of 392 DEGs including 255 up-regulated genes and 137 down-regualted genes were obtained from WYY/FLE libraries (**Figure 3C** and **Supplementary Table S4**). Functional annotation analyses revealed that 385 (98.21%), 150 (38.26%), 95 (24.23%), 189 (48.21%), 311 (79.34%), 283 (72.19%), and 117 (29.84%) unigenes were significantly enriched in the NR, COG, GO, KOG, PFAM, Swiss-Prot and KEGG databases, respectively. We obtained about 95 genes, which were classified into three main categories, and then grouped them into 44 sub-categories according to the GO classification, namely

TABLE 2 | Structure identification of anthocyanidins in WYY.

	Ingredient	Retention time (min)	Molecular ion (*m/z*)	MS^2 (*m/z*)	Annotation
WYY	C1	21.62	949, 757, 595, 449, 287	287.1	Cyanidin 3-O-rutinoside
	C2	27.72	611, 449, 287,	287.1	Cyanidin 3,5-diglucoside
	C3	31.97	465, 303	303.1	Delphinidin 3-O-glucoside
	C4	36.31	463, 301	301.7	Peonidin 3-O-glucoside
	C5	41.49	609, 463, 447, 301	301.7	Peonidin 3-O-rutinoside

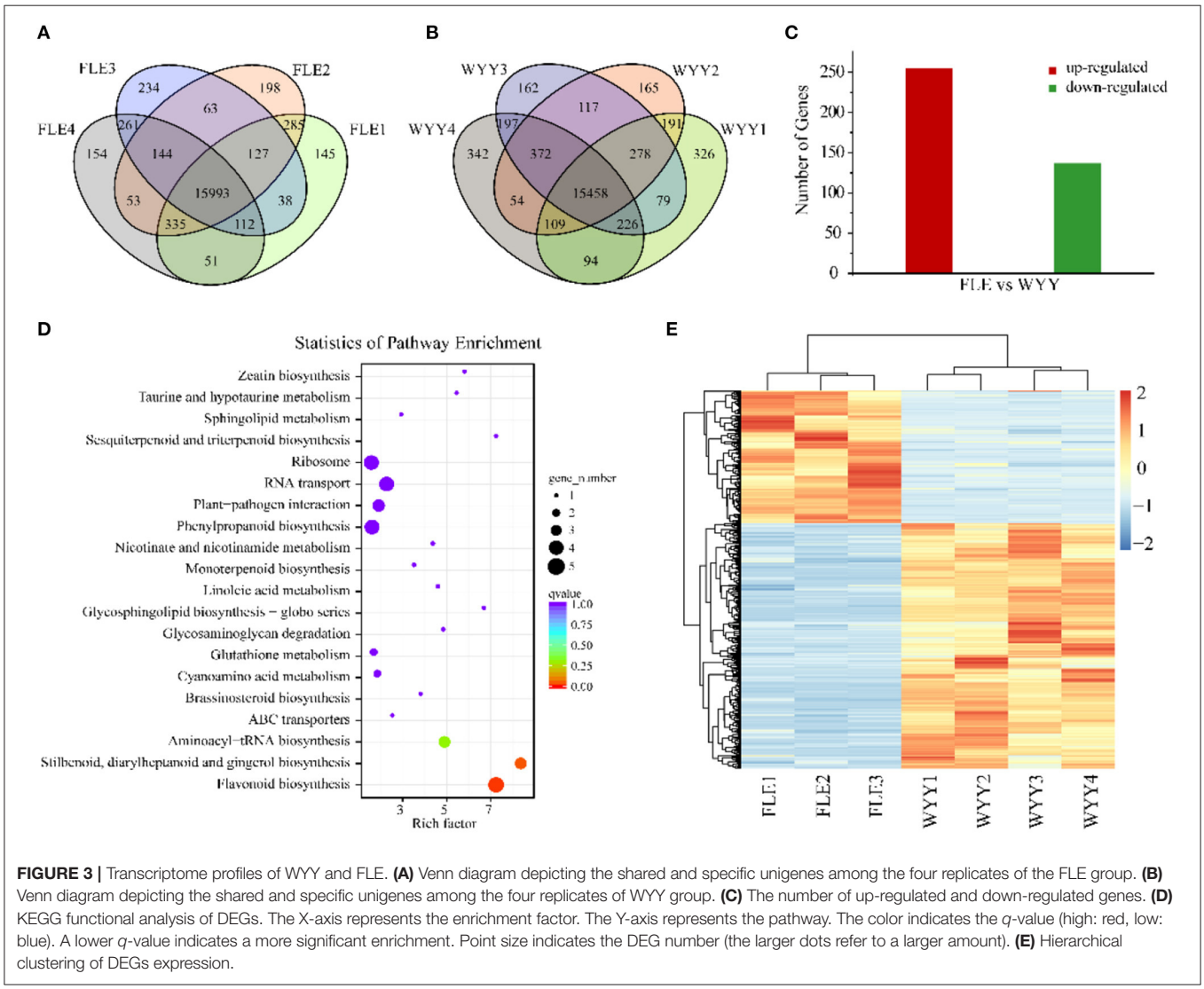


FIGURE 3 | Transcriptome profiles of WYY and FLE. **(A)** Venn diagram depicting the shared and specific unigenes among the four replicates of the FLE group. **(B)** Venn diagram depicting the shared and specific unigenes among the four replicates of WYY group. **(C)** The number of up-regulated and down-regulated genes. **(D)** KEGG functional analysis of DEGs. The X-axis represents the enrichment factor. The Y-axis represents the pathway. The color indicates the *q*-value (high: red, low: blue). A lower *q*-value indicates a more significant enrichment. Point size indicates the DEG number (the larger dots refer to a larger amount). **(E)** Hierarchical clustering of DEGs expression.

biological processes, molecular functions and cellular component (**Supplementary Figure S6**). In the biological process, GO terms were mainly enriched in the metabolic and the cellular processes, followed by the single-organism process. For the cellular component category, DEGs associated with cell and cell part were the most abundant. Within molecular function, the main sub-categories were catalytic activity and binding, followed by molecular function regulator, indicating that transcription factors and high enzymatic activity are closely related to the regulation of WYY stem coloration.

We then compared the DEGs against the KEGG pathways to obtain significantly enriched pathways. A total of 117 DEGs were mapped into 43 pathways, while only 4 pathways were significantly enriched. The top 20 enriched pathways were used to draw the enrichment map (**Figure 3D**). The most represented pathways comprised of one flavonoid biosynthesis

pathway (ko00941), one phenylpropanoid biosynthesis pathway (ko00940), followed by stilbenoid, diarylheptanoid and gingerol biosynthesis (ko00945) and RNA transport (ko03013), which can form a metabolic network. Basically, naringenin produced by the "phenylpropanoid biosynthesis" pathway is used by the "flavonoid biosynthesis pathway" to produce dihydroflavonols. Then, the flavonoid metabolic pathway produces leucoanthocyanins as the substrate of the anthocyanin biosynthesis pathway, which is modified to produce various types of anthocyanins and transported to the vacuole for stable existence.

DEGs Related to Color Development

Considering the differences between WYY and FLE in the major anthocyanin types, and the KEGG enrichment showed that DEGs were enriched in the flavonoid metabolic pathway, 55 genes involved in anthocyanin biosynthesis were investigated, including *PAL*, *4CL*, *CHS*, *F3'5'H*, *F3'H*, *DFR*, *LDOX*, *UFGT*, and *GST*. Genes encoding anthocyanin synthesis, modification and transport (*DFR*, *UFGT* and *GST*) showed higher expression levels in WYY, implying that these genes may be vital for the accumulation of anthocyanins. In contrast, *chalcone synthases* (*CHS*), *coumaricacid3-hydroxylase* (*C3H*), and *flavonol synthase* (*FLS*) were down-regulated in WYY compared with FLE (**Figure 4**). Two homologous genes of hydroxycinnamoyl-CoA shikimate/quinate *hydroxycinnamoyl transferases* (*HCT*) showed opposite expression patterns, one was up-regulated and the other was down-regulated.

DEGs Related to Flavonoid Transport

Anthocyanin glycosides are usually transported into the vacuole *via* transporters for storage or isolation. We searched all these transcripts encoding proteins implicated in transport and catabolism pathways in the functional annotations. The results showed that 27 key unigenes were differentially expressed, of which 18 were up-regulated and 9 were down-regulated. Among the 27 genes, cytochrome P450 families, ABC transporter family, and glutathione S-transferase-like protein had the most members (**Figure 4**, **Supplementary Table S5**). The expression levels of four cytochrome P450 genes were up-regulated and four genes were down-regulated in WYY. Interestingly, three glutathione S-transferase were expressed in all four WYY samples, but not in any of the FLE samples. Moreover, a multidrug and toxic efflux transporter protein (TT12) was significantly up-regulated in WYY.

DEGs Related to Sugar and Hormones Metabolism

Among the 12 putative functional homologous genes implicated in hormone and sugar metabolism, homologs of *SAUR-like auxin-responsive protein* (*SAUR32*), *auxin efflux carrier component 5* (*PIN5*), and *brassinosteroid insensitive 1* (*BI1*), and *cytokinin dehydrogenase 3* (*CKX3*) were significantly up-regulated in WYY. Three transcripts were significantly down-regulated in WYY, including homologs of *salicylic acid-binding protein* (*SABP2*), *sucrose synthase 6* (*SUS6*), and an NADPH: *quinone oxidoreductase* (**Supplementary Table S6**).

Transcription Factors Related to the Synthesis of Flavonoids

Transcription factors play an important role in the regulation of flavonoid biosynthesis by regulating the expression level of structural genes. In our study, 17 TFs of 7 TF families were found. Among the TF families, MYB (3 unigenes) and B3 (2 unigenes) were the most prominent, followed by NAC (2 unigenes), WRKY (1 unigene), HD-ZIP (1 unigene), bZIP (1 unigene) and bHLH (1 unigene) (**Supplementary Table S7**). Almost all MYB DEGs belong to the R2R3 MYB family, which has been reported to participate in regulating anthocyanin synthesis in multiple species. Two MYB genes encoding *MYB75* and *MYB108* were identified, and transcription analysis showed that *MYB108* had a higher expression level in the red stems. We infer that these TF families play an important role in the structural gene regulation of WYY stem coloration. To verify the accuracy of RNA-seq data, 16 genes, including the candidate structural genes together with the key transcription factors, were selected and a quantitative real-time PCR (qRT-PCR) was performed. The results showed that the gene expression profiles were well consistent with the RNA-seq data, which further demonstrated the credibility of the data generated in our study (**Figure 5**, **Supplementary Table S8**).

Comprehensive Analysis of Metabolome and Transcriptome

To investigate the relationship of DEMs and DEGs involved in the same biological process (KEGG pathway), the co-expression analysis of metabolome and transcriptome was performed using Pearson's correlation coefficient. There were many pathways for simultaneous annotation of differential metabolites and differential genes. We picked genes and metabolic pathways with $p < 0.05$ for priority analysis, which can save the time of data screening and quickly find the pathways for subsequent analysis. These pathways are summarized in **Figure 6A**, and the results indicated that the most representative category was the flavonoid biosynthesis pathway.

To obtain the potential relationship between genes and metabolites, the FPKM of transcription level and metabolic level was calculated, and then screened according to the canonical correlation analysis (CCA), with the following parameters: $|CC|>0.8$ and CCP <0.05. Further analysis showed that 37% of the genes belonged to the congruent groups (group c and group g), and their expression trends were consistent at the transcriptional and metabolic levels, which may positively regulate the expression of metabolites. Meanwhile, 63% of the genes were located in the incongruent groups (group a and i), and the expression trends of genes and metabolites were opposite, which meant that genes and metabolites were negatively correlated (**Figure 6B**).

Next, we focused on metabolites and genes with consistent trends. All significantly differentially expressed metabolites and genes were classified by K-means method, and four clusters were identified (**Figures 7A–D**). Compared with Metabolites cluster 1 and Genes cluster 2, which showed an overall downward trend, Metabolites cluster 2 and Genes cluster 1 displayed an overall upward trend. Genes cluster 1 encompassed 255 genes,

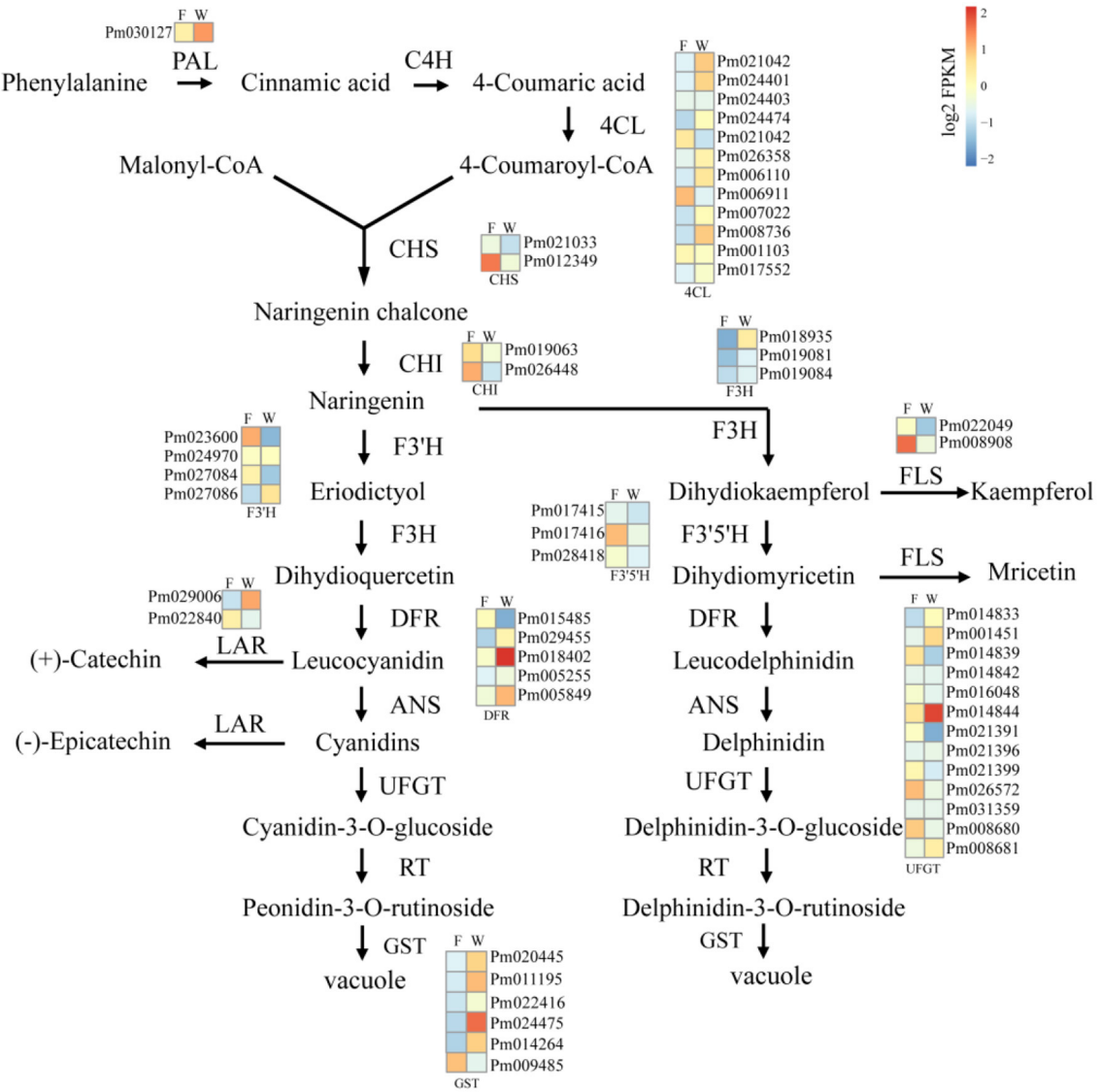


FIGURE 4 | The analysis of unigenes involved in anthocyanin biosynthesis of WYY. Phenylalanine ammonia-lyase (PAL), cinnamic acid 4-hydroxylase (C4H), 4 coumarate-CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonol synthase (FLS), anthocyanidin 3-O-glucosyltransferase (UFGT), Dihydroflavonol 4-reductase (DFR), anthocyanidin reductase (ANR), leucoanthocyanidin dioxygenase (LAR). F:FLE, W:WYY. Squares represent different transcripts of genes. In the heat maps, blue-yellow-red scale represent low and high transcript expression, respectively.

which were highly expressed, suggesting a positive correlation with red stem formation in WYY. Conversely, 137 genes in Genes cluster 2 were down-regulated, which might be negatively associated with the red coloration. KEGG analysis also revealed that flavonoid biosynthesis and phenylpropanoid biosynthesis were significantly enriched in Genes cluster 1.

The Coexpression Analysis of DEGs and DEMs in Flavonoid Biosynthesis Pathway

The co-expression networks of DEGs and DEMs were mainly enriched in phenylpropanoid biosynthesis (**Figure 7E**) and flavonoids biosynthesis (**Figure 7F**). We found that p-Coumaric acid, p-Coumaryl alcohol, sinapyl alcohol, and coniferin were positively correlated with the transcription expression of *Pm008812* encoding HCT (hydroxycinnamoyl-CoA shikimate transferases), and *Pm003887*, *Pm003886*, both encoding C3H (cytochrome P450 98A2). In contrast, *Pm008809*, encoding HCT, and *Pm028093* (beta-glucosidase 12) were negatively related to these metabolites. The enzymes encoded by these genes catalyzed the conversion between p-coumaroyl CoA and p-coumaroyl shikimic/quinic acid, which would be critical for the synthesis of caffeoyl shikimic/quinic acids. C3H/HCT determines the flow of carbon sources in plants and plays an important role in the phenylpropanoids pathway. Whether p-coumarin-CoA forms H-lignin or G/S-lignin depends on the activity of C3H/HCT. The DEGs and DEMs, such as *CHS* (*Pm012349*), *C3H*(*Pm003886*),

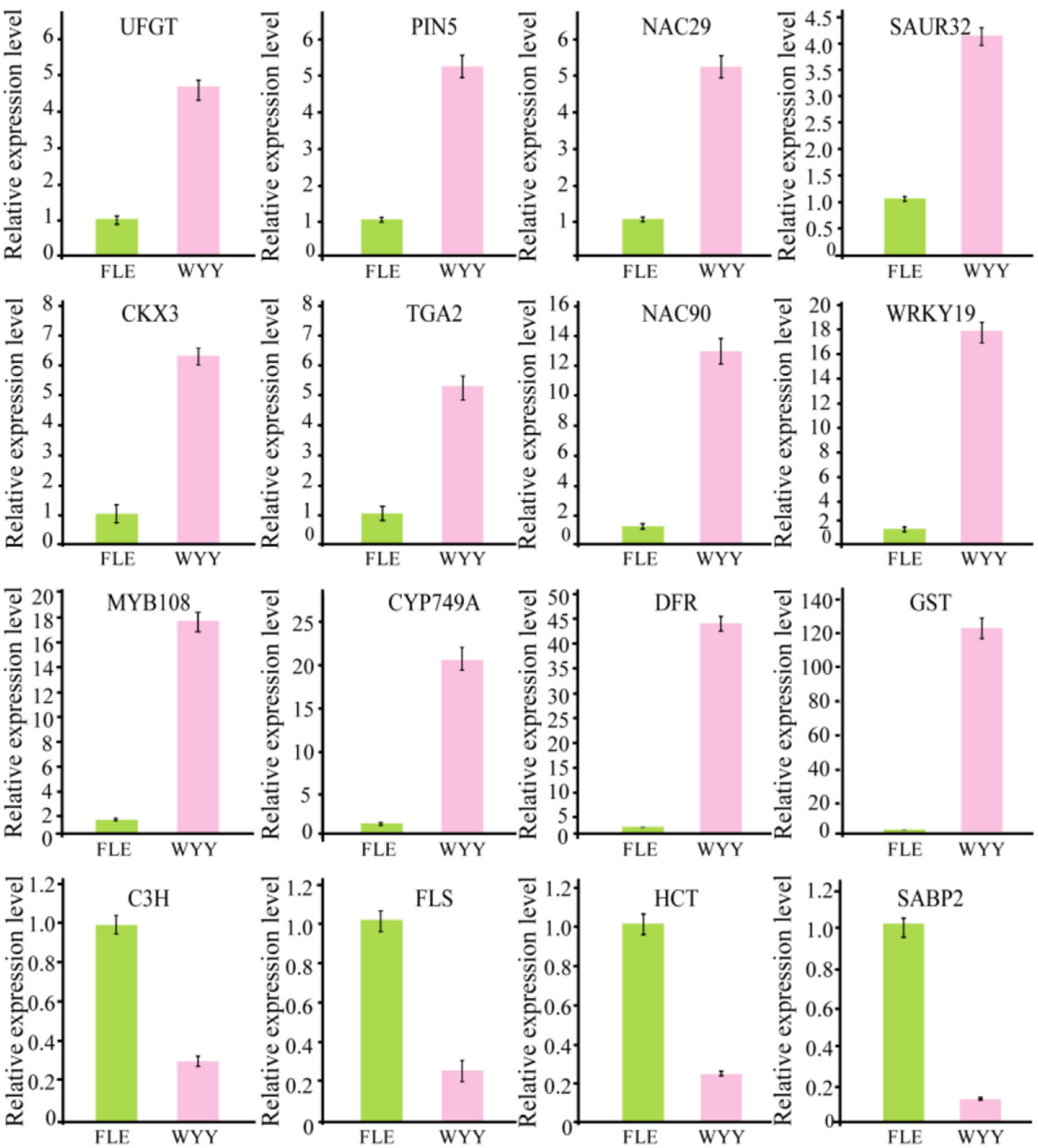


FIGURE 5 | Validation of RNA-seq data by qRT-PCR.

kaempferol and myricetin, quercetin, delphinidin were found to be involved in the biosynthesis of flavonoids.

DISCUSSION

Deciphering the mechanism of plant coloration has always been a hot topic in ornamental horticulture. However, previous studies mainly focused on the leaves, flowers, fruits, seeds, epidermis and other organs or tissues of herbaceous plants, and there was no report on the formation mechanism of xylem color trait of woody plants. In this work, as an effort to reveal the underlying molecular mechanisms of the formation of different colors in stems of *P. mume*, a combined metabolome and transcriptome study was designed. Upon comparison of the differentially accumulated metabolites in the different cultivars, the contents of flavonoids and anthocyanins were the main reason for the difference in stem color. A hypothetical model for red stem formation in *P. mume* is summarized in **Figure 8**. Higher anthocyanin content was detected in WYY, while a very small amount of anthocyanins were accumulated in FLE, which was consistent with the HPLC results in other cultivars. Twenty-four anthocyanins were identified in WYY and the high

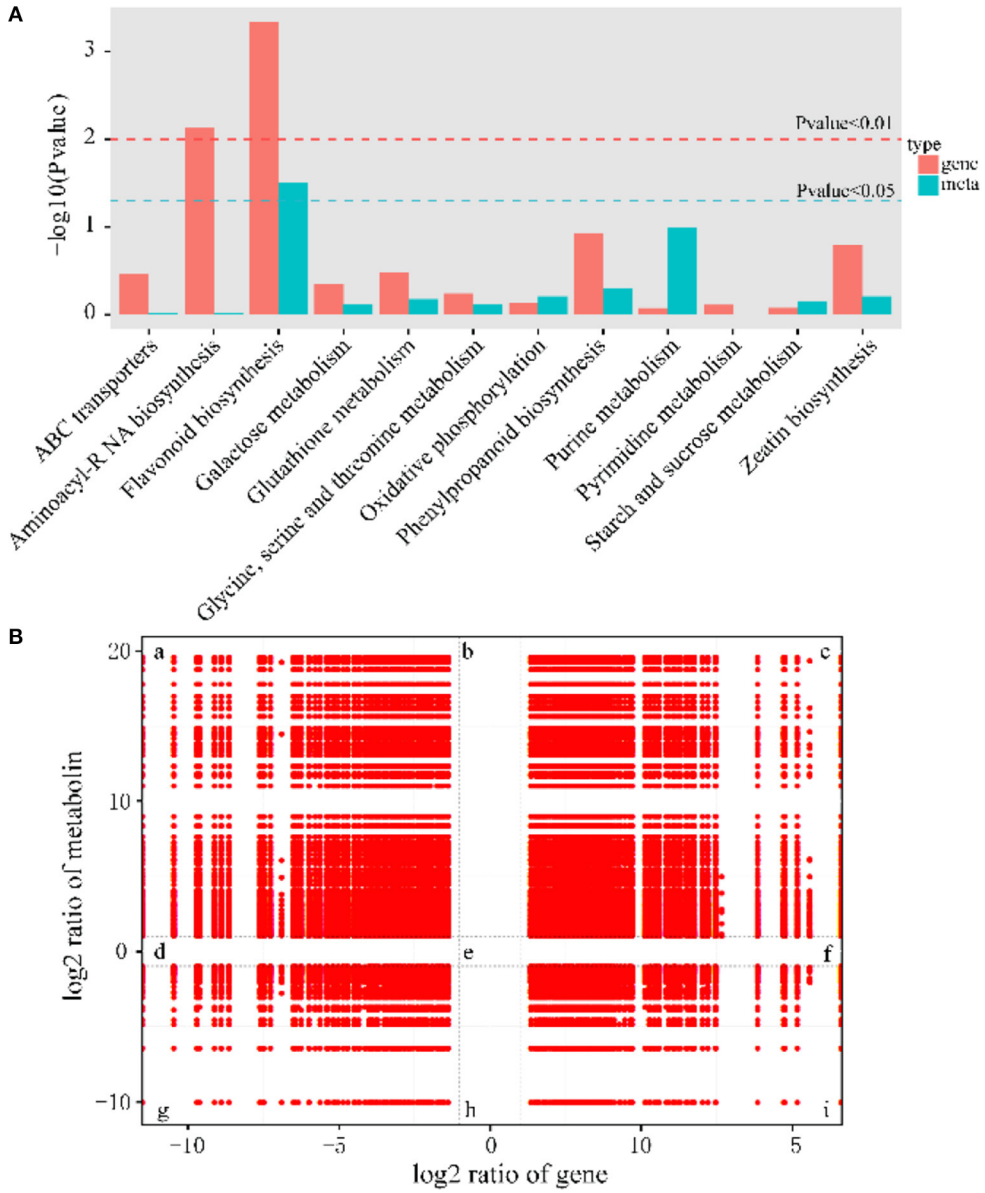


FIGURE 6 | Significantly related KEGG pathways and nine quadrant diagrams. **(A)** Integrated analysis of KEGG pathways of metabolome and transcriptome. **(B)** Nine squares indicated nine responsive groups: (a) Transcriptionally down-regulated and metabolically up-regulated genes. (b) Transcriptionally unchanged and metabolically up-regulated genes. (c) Transcriptionally and metabolically up-regulated genes. (d) Transcriptionally down-regulated and metabolically unchanged genes. (e) Transcriptionally and metabolically unchanged genes. (f) Transcriptionally up-regulated and metabolically unchanged genes. (g) Transcriptionally and metabolically down-regulated genes. (h) Transcriptionally unchanged and metabolically down-regulated genes. (i) Transcriptionally up-regulated and metabolically down-regulated genes. The X axis represents log2 expression ratios of gene and Y-axis represents log2 expression ratios of metabolin.

accumulation levels of cyanidin 3,5-O-diglucoside, cyanidin 3-O-rutinoside, and peonidin 3-O-glucoside were considered to be the main components of red pigmentation. In addition, a large amount of procyanidin A1, quercetin, tricetin, and other important secondary metabolites were accumulated in the biosynthetic pathway of phenylpropanoid and flavonoids. It is worth noting that some anthocyanin compounds were accumulated only in WYY, in particular cyanidin glycosides and pelargonidin glycosides. This is in line with the previous studies, which found that red flowers and white flowers contained the same non-red flavonoids, and there was a significant positive correlation between red pigmentation and anthocyanin content in *P. mume* (Zhao et al., 2004). Ma et al. similarly found that cyanidin 3-O-glucoside, cyanidin 3,5-O-diglucoside, and peonidin 3-O-glucoside were the primary anthocyanins in pink petals, while no such substances were detected in *P. mume*

with white petals (Ma et al., 2018). We also found that several anthocyanins were present simultaneously in WYY and FLE, which may be because, in addition to the amount and type of total anthocyanins, the intensity of anthocyanins were affected by a variety of complex factors, such as vacuolar pH, co-pigmentation, and metal chelation (Manetas, 2006). The same predominant anthocyanins were found in three different colored poinsettia bracts (Slatnar et al., 2013). In our study, quercetin and kaempferol were significantly accumulated in the flavonoid synthesis pathway, which is the upstream of the anthocyanin pathway. Low levels of epicatechin, which compete with anthocyanins for the same substrate, and high levels of flavonoids provide sufficient substrates for anthocyanin accumulation.

On the basis of GO and KEGG analysis of transcriptome data, we suggest that phenylpropanoid synthesis and flavonoid biosynthetic pathway may strongly influence the formation of red stems. The increased gene transcription levels in these pathways strongly support our metabolome results. Anthocyanin biosynthetic pathway has been found to regulate color formation in many plants, such as *Camellia sinensis* (Zhou et al., 2020), *Paeonia suffruticosa* (Gu et al., 2019), waterlily (Wu et al., 2016), and *Senecio cruentus* (Jin et al., 2016). Based on the expression levels and fold changes, some genes related to anthocyanin biosynthesis were exploited. In terms of the upstream genes, CHS, as a key enzyme affecting the accumulation of flavonoids, can catalyze the condensation of coumaric acid CoA into chalcone derivatives, which form the basic skeleton of downstream flavonoids. However, in our study, we found that a *CHS* homologous unigene was down-regulated in WYY and negatively related to the accumulation of kaempferol and myricetin, which may be affected by the feedback mechanism of the flavonoid pathway in plants. The transcriptome expression pattern revealed that compared with the flavonols biosynthesis, WYY preferentially flows to the anthocyanin pathway by up-regulating the expression of *DFR* and down-regulating one *FLS* gene, which can catalyze dihydroflavanols to produce leucocyanidin and flavonols, respectively. It has been reported that the metabolic balance in the flavonoid biosynthetic pathway is affected by the ratio of *FLS/DFR* (Gou et al., 2011). Here, one differentially expressed *DFR* homologous unigene was identified in WYY, and its expression level was 4-fold higher than that in FLE. Conversely, the *FLS* was highly expressed in FLE. Furthermore, correlation analysis between metabolites and transcripts showed that anthocyanin content was positively correlated with *DFR* and negatively correlated with *FLS*. The silencing of *McDFR* increased the accumulation of flavonols, whereas inactivation of *McFLS* elevated anthocyanin content in crabapple (Tian et al., 2015). In grape hyacinth, a highly expressed *FLS* along with a lowly expressed *DFR* lead to the fading of blue pigmentation (Lou et al., 2014). It provides a new perspective for breeders to cultivate new varieties of ornamental plants with novel colors.

On the other hand, DFR is a rate-limited enzyme, which can specifically catalyze one or more dihydroflavonols to produce the corresponding leucoanthocyanidins due to the different enzymatic sites in the conserved domain. In peanuts, DFR does not catalyze dihydrokaempferol and therefore fails to produce pelargonidin, resulting in lack of brick-red flowers (Johnson et al., 1999). Down-regulating endogenous *DFR* expression levels and over-expressing the *iris DFR*, which preferred dihydromyricetin, yielded a blue rose rich in delphinidin (Katsumoto et al., 2007). In our study, WYY was able to accumulate all three anthocyanins, indicating that DFR belonged to the non-specific DFR enzymes and can convert all types of dihydroflavonols. However, there were significant differences in the contents of different types of anthocyanins. Therefore, how *DFR* affect the formation of different colors needs further study.

Anthocyanins are easily degraded, so glycosylation is an indispensable part of the process of anthocyanin accumulation in plants. Under the catalysis of UFGT, the hydrophilicity and stability of aromatic rings were increased by adding sugar moieties to the 3-OH position of anthocyanin (Lo Piero, 2015). The *GST* catalyzed the formation of a relatively stable complex between glutathione and anthocyanin and transported it to vacuoles for storage (Mol et al., 1998; Grotewold, 2006). In apple, the activity of UFGT was positively correlated with the accumulation of anthocyanins (Ji et al., 2015). In our study, we detected that one *UFGT* gene and several *GST* genes were strongly up-regulated, possible contributing to the high accumulation of anthocyanins. This is consistent with findings in grape, which found that *UFGT* was highly expressed in red-skin grapes, but not in white cultivar tissues, and *UFGT* was up-regulated at veraison (Boss et al., 1996).

Glutathione s-transferase (GST) is a family of proteins with abundant physiological functions, which plays an important role in plant detoxification and secondary metabolism. There are two hypotheses for the mechanism of anthocyanin transport from endoplasmic reticulum to vacuoles: vesicle-trafficking model and transporter model (Liu et al., 2019). Both models infer that GST was significant for the efficiency of anthocyanin transport. Our results also suggest that *GST* may be indispensable for anthocyanin pigmentation. We detected three *GSTs* in all four WYY samples, which were almost undetectable in FLE samples, and a *multidrug and toxic efflux transporter protein* (*TT12*) was significantly up-regulated in WYY. The effect of GST on conjugation of glutathione and anthocyanins was first demonstrated in maize *bz2* mutant (Marrs et al., 1995). In petunia, *AN9*, which was regulated by transcriptional activator of anthocyanin pathway, performed a similar function to *bz2* and complemented the phenotype of *bz2* mutant (Alfenito et al., 1998). Based on research in peach, a small indel mutation in *Riant*, which encoded a GST, made the red petals to fade or even turn white (Cheng et al., 2015). *Arabidopsis* MATE family contained 56 genes, each of which performed different functions, and their transport substances were different. *AtTT12* is located in the tonoplast and function to transport proanthocyanidin precursors epicatechin-3-O-glucoside into the vacuole (Marinova et al., 2007). More recently, MATE2 from *Medicago truncatula*, homologous *to TT12,* was found to efficiently transport cyanidin 3-O-glucoside (Zhao et al., 2011). On the whole, high expression levels of anthocyanin-related genes including *DFR*, *UFGT*, *GST*, and *TT12*, may play key roles in the direction and distribution of flavonoids metabolic flux and the accumulation of anthocyanins in WYY.

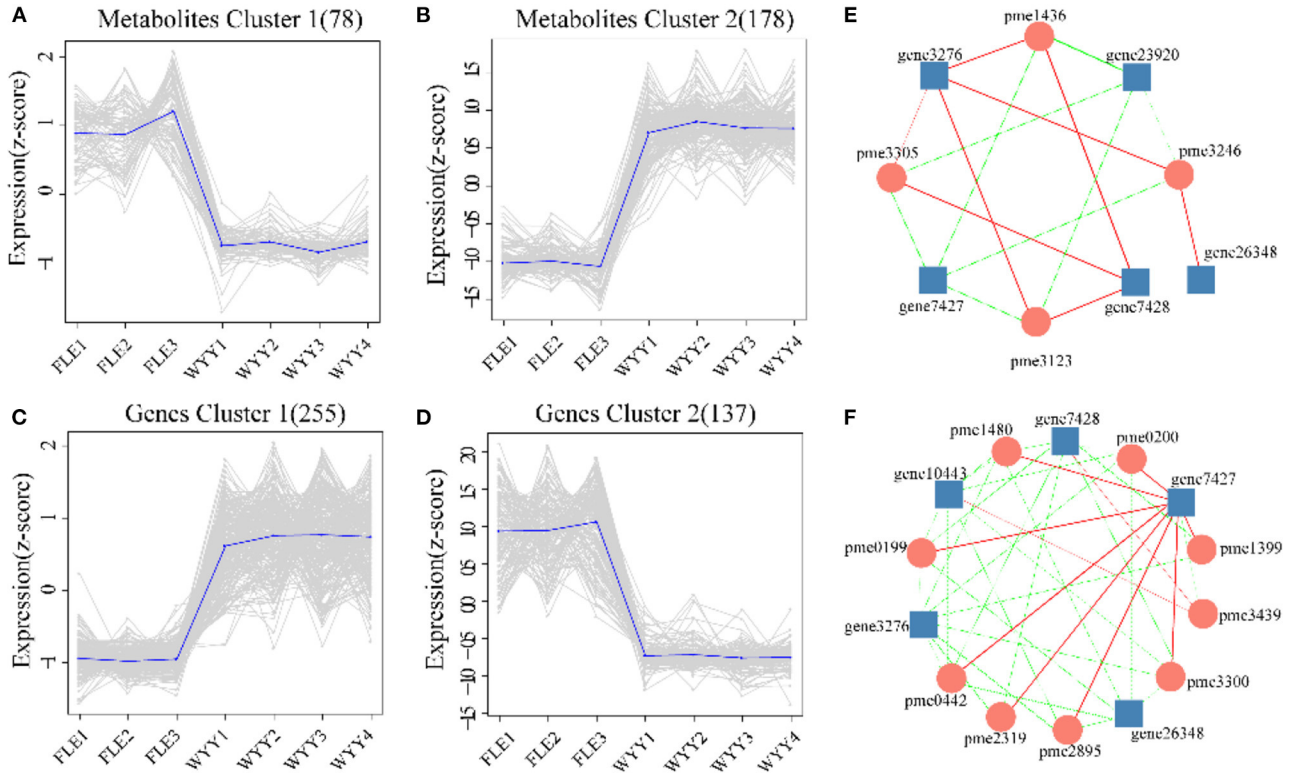


FIGURE 7 | Trend analysis and the co-expression analysis of DEGs and DEMs. **(A,B)** K-means clustering analysis of the DEMs into two clusters according to their expression profile. **(C,D)** K-means clustering analysis of the DEGs into two clusters according to their expression profile. **(E)** Interaction network of DEGs and DEMs involved in phenylpropanoid biosynthesis. **(F)** Interaction network of DEGs and DEMs associated with flavonoids biosynthesis. Metabolites are shown in circles and gene IDs are shown in squares. Edges colored in "red" and "green" represent positive and negative correlations, respectively.

Anthocyanin synthesis has specific synergies among structural genes and is regulated by a transcription complex formed by R2R3-MYB, bHLH, and WD40 transcription factors (MBW). The function of the MBW complex has been widely verified in plants, such as *Prunus persica* (Uematsu et al., 2014), *Lilium* (Yamagishi, 2010), and *Actinidia chinensis* (*Peng et al., 2019*). In peony, *PsMYB12* interacted with a bHLH and a WD40 protein in a regulatory complex that directly activated *PsCHS* expression (Gu et al., 2019). Furthermore, *LcR1MYB1* and *LcNAC13* acted together to antagonistically regulate anthocyanin biosynthesis during litchi fruit ripening (Jiang et al., 2019). We identified several members of these transcription factors. One unigene encoding *MYB108* exhibited high expression levels in WYY, and several members of this gene family had been reported to mediate flower color in plants (Takahashi et al., 2011; Zhu et al., 2015). However, *MYB75*, highly homologous to *A. thaliana MYB113*, was down-regulated. Besides, we also observed that a bHLH transcription factor was up-regulated, which was involved in tapetal cell development. However, no annotated gene related to WD40 was found in DEGs. Our recent studies showed that a petal color-associated QTL was located in the same area of *MYB108* on chromosome Pa4. Transcriptome data also indicated that *MYB108* was specially expressed in the red petals of WYY (Zhang et al., 2018). Furthermore, we found that the expression of some hormones and sugar-related proteins, such as *efflux carrier component 5* (*PIN5*), *auxin response protein SAUR32*, *brassinosteroid insensitive 1* (*BRI1*), and sucrose synthase 6, changed significantly in WYY. The production of anthocyanin in *O. linearis* callus cultures was regulated by the feedback of auxin concentration, which was inhibited at high concentrations and promoted at low concentrations (Meyer and Van Staden, 1995). Liu et al. (2014) showed that auxin increased the expression of *DFR* and *ANS* genes by stimulating TTG1-TT8-PAP1 complex, thereby regulating the accumulation of anthocyanins in red pap1-D *Arabidopsis* cells. It has been proved that sugar-phosphorylation interacting with related signal transduction can induce gene expression and anthocyanin accumulation in developing petunia flowers (Solfanelli et al., 2006). We hypothesize that these differentially expressed transcription factors can be candidate regulators of WYY anthocyanin biosynthesis, but their molecular and physiological functions are still unclear and need further study.

CONCLUSION

In this research, we systematically studied the formation mechanism of red stem in WYY through multi-omics analysis.

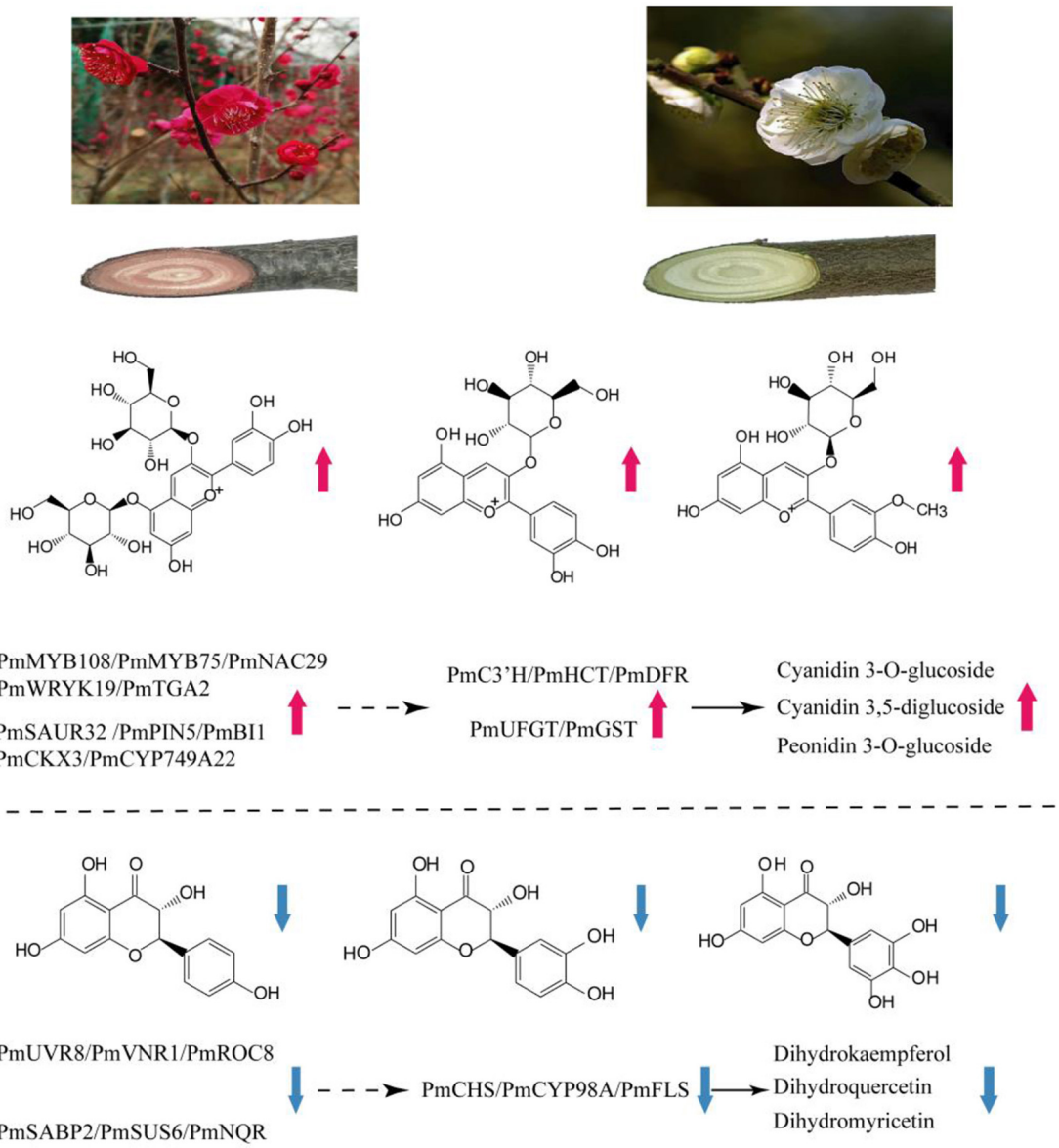


FIGURE 8 | Summary of transcription-level regulation of the red stem formation in *P. mume*. Transcripts and metabolites show obvious differences in WYY and FLE. Total anthocyanins content and expression of *C3'H, HCT, DFR, UFGT, GST* genes are increased, resulting in preferential anthocyanin pathway in WYY. Moreover, the increased expression of *CHS, CYP98A and FLS* causes a higher proportion of flavonols in FLE. The solid black arrow indicates direct control; the dashed arrow indicates indirect regulation.

The high accumulation of metabolites in anthocyanin-related synthetic pathways, especially cyanidin glycoside and paeoniflorin glycoside, were considered to be the main sources of red pigmentation. The transcriptome and the correlation analyses revealed that differentially expressed structural genes and transcription factors, such as *FLS*, *DFR*, *UFGT*, *MYB75*, and *MYB108*, may contribute to modulating the formation of the red stems in WYY. Our results provide a reference for molecular mechanism of xylem color trait in *P. mume* and lay a theoretical foundation for cultivating new varieties.

MATERIALS AND METHODS

Plant Materials

Two *P. mume* cultivars with different stem colors ('Wuyuyu' with red stem and 'Fei Lve' with green stem) were chosen for transcriptome and metabolome analyses. Both cultivars were grown in the same nursery of Beijing Forestry University. Stems are selected from three different individuals of each cultivar based on size, length, and degree of lignification. The samples were frozen in liquid nitrogen for subsequent transcriptome and metabolome analyses. Transcriptome sequencing was performed

on three biological replicates, and metabolic profile analysis was performed on four biological replicates. The cultivars 'Wuyuyu' and 'Fei Lve' were abbreviated as WYY and FLE, respectively.

Metabolomics

For metabolomic analysis, the metabolites were extracted, identified, and quantitatively analyzed by Biomarker Technologies (Beijing, China). In brief, the freeze-dried stem was crushed and 100 mg of powder was dissolved in 1.0 mL of 70% methanol solution overnight at 4°C. After centrifugation and filtration, the extracted solution was analyzed using an LC-ESI-MS/MS system (HPLC, Shim-pack UFLC SHIMADZU CBM30A system; MS, Applied Biosystems 6500 Q TRAP). Before the data analysis, a quality control sample, which was a mixture of sample extracts, was prepared to monitor the repeatability of the sample under the same detection method. Mass spectrometry data were processed by Analyst 1.6.1 software (AB Sciex). Principal component analysis (PCA) and Hierarchical cluster analysis (HCA) were used to evaluate the metabolic differences between different samples. Metabolites were considered to be differentially accumulated if the fold change was≥2 and variable importance in project (VIP) was≥1. Enrichment analysis of differential metabolites was compared to the KEGG database and clustered by using cluster Profiler in R.

Anthocyanin Content Measurement

Four cultivars were used to determine the content of anthocyanins, including 'Wuyuyu', 'Fenhong Zhusha', 'Fei Lve' and 'Zaohua Lve'. First, 0.25 g of powder sample per cultivar was added into 5 mL of methanol, distilled water, formic acid, and trifluoroacetic acid (70:27:2:1, v/v/v/v), and extracted at 4°C for 24 h. After centrifugation, the supernatants were filtered by a medium-speed filter paper and a 0.22 μm syringe filter (Millipore, Bedford, MA, USA) before subjecting it to HPLC-MS/MS analysis. The mobile phase was 2% formic acid in water (phase A) and 0.1% formic acid acetonitrile solution (phase B) at a flow rate of 0.6 mL/min. The linear gradient of phase B was as follows: 0 min 5% B, 20 min 28% B, 30 min 60 B%, 45 min 28% B, 45–60 min 5% B. The UV-visible light detector wavelength was set at 520 nm for detecting anthocyanins. The mass spectrometry analysis conditions were as follows: electrospray ionization (ESI^+); ion trap analyzer; scan mode: total ion scanning; scanning range (m/z):100–1,000; capillary voltage: 4,000 V; sprayer pressure: 35 psi; dry gas: N2; dry temperature: 350°C. For preparation of the standard solution, we accurately weighed the cyanidin 3,5-O-diglycoside and diluted it to several different concentrations. The quantitative analysis was carried out according to the procedures described by Sun et al. (2009). Each sample was repeated three times under the same conditions.

Transcriptomics

Total RNA of all stem samples was extracted using a RNAprep Pure Plant Kit (DP432, Tiangen, China). The quality and purity of RNA were evaluated by 1% agarose gel and NanoDrop 2000 (Thermo fisher Scientific, USA). Qualified RNA samples were sent to Biomarker Technology for cDNA library construction and sequencing. At least four biological repeats were designed per cultivar. Eight cDNA libraries were constructed and sequenced on an Illumina HiSeq 2500 platform. After removing the adapter and low-quality sequencing data, the obtained high-quality clean reads were aligned to the *P. mume* genome sequence using HISAT2, and then assembled and quantified using StringTie (Pertea et al., 2015). For functional annotation, all assembled unigenes were aligned to the public database by using BLASTX (Altschul et al., 1997), including Nr, Pfam, KOG/COG, Swiss-Prot, and KEGG. The differentially expressed genes (DEGs) were detected with DEGseq (Wang et al., 2010), and fold change≥2 and false discovery rate (FDR) < 0.01 were used as threshold in the detection process. GOseq R package and KOBAS software were used for GO enrichment and KEGG analysis of the DEGs, respectively.

Integrative Metabolome and Transcriptome Analysis

Pearson correlation coefficient was used to screen the correlation between metabolites and genes, with correlation coefficients> 0.8 and $P < 0.05$ as the selection criteria. The co-expression analysis between differential expression metabolites (DEMs) and differential expression genes (DEGs) in phenylpropanoid and flavonoids biosynthesis were visualized in Cytoscape (v3.3.0).

Expression Pattern Analysis

The transcript levels of 16 genes were subjected to qRT-PCR. The specific gene primers were designed by Integrated DNA Technologies (https://sg.idtdna.com) based on reference sequences from *P. mume* genome. All reactions were conducted using the TB Green® Premix Ex Taq™ II (TaKaRa, Beijing, China) on the PikoReal Real–Time PCR System (Thermo Fisher Scientific). The 2-delta-delta Ct method was used to calculate gene expression, and *Protein phosphatase 2A* (*Pm006362*) was selected as an internal reference control. All analytical procedures were carried out with three biological replicates.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://ngdc.cncb.ac.cn/search/?dbId=gsa&q=CRA006277.

AUTHOR CONTRIBUTIONS

LQ and TZ conceived and drafted the manuscript. LQ, TZ, WL, XZ, and PL analyzed the data. JW and TC provided of plant resources. TZ and QZ contributed to the conception of the study and finalized the manuscript. All authors read and approved the final manuscript.

FUNDING

The research was supported by the National Natural Science Foundation of China (Nos. 32071816 and 31800595), and the Special Fund for Beijing Common Construction Project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.884883/full#supplementary-material

Supplementary Figure S1 | Principal component analysis of the metabolome samples.

Supplementary Figure S2 | Heat map of the metabolic profiling.

Supplementary Figure S3 | OPLS-DA model plots.

Supplementary Figure S4 | KEGG Pathway classification and functional enrichment of DEMs.

Supplementary Figure S5 | Pearson correlation between eight libraries.

Supplementary Figure S6 | GO functional analysis and classification of DEGs.

Supplementary Table S1 | All metabolite information identified by widely-targeted metabolomics.

Supplementary Table S2 | Differentially expressed metabolites between WYY and FLE.

Supplementary Table S3 | Statistics of RNA-seq reads and genome alignment in 8 samples.

Supplementary Table S4 | Differentially expressed genes between WYY and FLE.

Supplementary Table S5 | DEGs related to flavonoid transport.

Supplementary Table S6 | DEGs related to sugar and hormones metabolism.

Supplementary Table S7 | Differentially expressed transcription factors.

Supplementary Table S8 | Primers used for qRT-PCR.

REFERENCES

Alfenito, M. R., Souer, E., Goodman, C. D., Buell, R., Mol, J., Koes, R., et al. (1998). Functional complementation of anthocyanin sequestration in the vacuole by widely divergent glutathione S -transferases. *Plant Cell* 10, 1135–1149. doi: 10.1105/tpc.10.7.1135

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Boldt, J. K., Meyer, M. H., and Erwin, J. E. (2014). Foliar anthocyanins: A Horticultural Review. *Hortic. Rev.* 42. doi: 10.1002/9781118916827.ch04

Boss, P. K., Davies, C., and Robinson, SP (1996). Analysis of the expression of anthocyanin pathway genes in developing *Vitis vinifera* L. cv Shiraz grape berries and the implications for pathway regulation. *Plant Physiol* 0.111, 1059–1066. doi: 10.1104/pp.111.4.1059

Celli, G. B., Tan, C., and Selig, M. J. (2018). "Reference module in food science || anthocyanidins and anthocyanins," in *Encyclopedia of Food Chemistry* (Academic Press), 218–223. doi: 10.1016/B978-0-08-100596-5.21780-0

Cheng, J., Liao, L., Zhou, H., Gu, C., Wang, L., and Han, Y. (2015). A small indel mutation in an anthocyanin transporter causes variegated colouration of peach flowers. *J. Exp. Bot.* 66, 7227–7239. doi: 10.1093/jxb/erv419

Cho, K., Cho, K. S., Sohn, H. B., Ha, I. J., Hong, S. Y., Lee, H., et al. (2016). Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. *J. Exp. Bot.* 67, 1519–1533. doi: 10.1093/jxb/erv549

Elomaa, P., Uimari, A., Mehto, M., Albert, V. A., Laitinen, R. A., and Teeri, T. H. (2003). Activation of anthocyanin biosynthesis in *Gerbera hybrida* (Asteraceae) suggests conserved protein-protein and protein-promoter interactions between the anciently diverged monocots and eudicots. *Plant Physiol.* 133, 1831–1842. doi: 10.1104/pp.103.026039

Espley, R. V., Hellens, R. P., Putterill, J., Stevenson, D. E., Kutty-Amma, S., and Allan, A. C. (2007). Red colouration in apple fruit is due to the activity of the MYB transcription factor, MdMYB10. *Plant J.* 49, 414–427. doi: 10.1111/j.1365-313X.2006.02964.x

Gonzalez, A., Zhao, M., Leavitt, J. M., and Lloyd, A. M. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* 53, 814–827. doi: 10.1111/j.1365-313X.2007.03373.x

Gou, J. Y., Felippes, F. F., Liu, C. J., Weigel, D., and Wang, J. W. (2011). Negative regulation of anthocyanin biosynthesis in *Arabidopsis* by a miR156-targeted SPL transcription factor. *Plant Cell* 23, 1512–1522. doi: 10.1105/tpc.111.084525

Grotewold, E. (2006). The genetics and biochemistry of floral pigments. *Annu. Rev. Plant Biol.* 57, 761–780. doi: 10.1146/annurev.arplant.57.032905.105248

Gu, Z., Zhu, J., Hao, Q., Yuan, Y. W., Duan, Y. W., Men, S., et al. (2019). A novel R2R3-MYB transcription factor contributes to petal blotch formation by regulating organ-specific expression of *PsCHS* in tree peony (*Paeonia suffruticosa*). *Plant Cell Physiol.* 60, 599–611. doi: 10.1093/pcp/pcy232

Harborne, J. B., and Gavazzi, G. (1969). Effect of Pr and pr alleles on anthocyanin biosynthesis in *zea mays*. *Phytochemistry* 8, 999–1001. doi: 10.1016/S0031-9422(00)86345-1

Harborne, J. B., and Williams, C. A. (2000). Advances in flavonoid research since 1992. *Phytochemistry* 55, 481–504. doi: 10.1016/S0031-9422(00)00235-1

Hughes, N. M., Neufeld, H. S., and Burkey, K. O. (2005). Functional role of anthocyanins in high-light winter leaves of the evergreen herb *Galax urceolata*. *New Phytol* 168, 575–587. doi: 10.1111/j.1469-8137.2005.01546.x

Isaak, C. K., Petkau, J. C., Blewett, H., Karmin, O., and Siow, Y. L. (2017). Lingonberry anthocyanins protect cardiac cells from oxidative-stress-induced apoptosis. *Can. J. Physiol. Pharmacol.* 8, 904–910. doi: 10.1139/cjpp-2016-0667

Jansen, M. A.K., Gaba, V., and Greenberg, B. M. (1998). Higher plants and UV-B radiation: balancing damage, repair and acclimation. *Trends Plant Sci.* 3, 131–135. doi: 10.1016/S1360-1385(98)01215-1

Ji, X. H., Zhang, R., Wang, N., Yang, L., and Chen, X. S. (2015). Transcriptome profiling reveals auxin suppressed anthocyanin biosynthesis in red-fleshed apple callus (*Malus sieversii* f. niedzwetzkyana). *Plant Cell* 123, 389–404. doi: 10.1007/s11240-015-0843-y

Jiang, G., Li, Z., Song, Y., Zhu, H., Lin, S., Huang, R., et al. (2019). LcNAC13 physically interacts with LcR1MYB1 to coregulate anthocyanin biosynthesis-related genes during litchi fruit ripening. *Biomolecules* 9, 135. doi: 10.3390/biom9040135

Jin, X., Huang, H., Wang, L., Sun, Y., and Dai, S. (2016). Transcriptomics and metabolite analysis reveals the molecular mechanism of anthocyanin biosynthesis branch pathway in different *Senecio cruentus* cultivars. *Front. Plant Sci.* 7, 1307. doi: 10.3389/fpls.2016.01307

Jo, Y. N., Jin, D. E., Jeong, J. H., Kim, H. J., Kim, D. O., and Heo, H. J. (2015). Effect of anthocyanins from rabbit-eye blueberry (*Vaccinium virgatum*) on cognitive function in mice under trimethyltin-induced neurotoxicity. *Food Sci. Biotechnol.* 24, 1077–1085. doi: 10.1007/s10068-015-0138-4

Johnson, E. T., Yi, H., Shin, B., Oh, B., Cheong, H., and Choi, G. (1999). Cymbidium hybrida dihydroflavonol 4-reductase does not efficiently reduce dihydrokaempferol to produce orange pelargonidin-type anthocyanins. *Plant J.* 19, 81–85. doi: 10.1046/j.1365-313X.1999.00502.x

Jonsson, L. M.V., Donker-Koopman, W. E., and Schram, A. W. (1984). Turnover of anthocyanins and tissue compartmentation of anthocyanin biosynthesis in flowers of *Petunia hybrida*. *J. Plant Physiol.* 115, 29–37. doi: 10.1016/S0176-1617(84)80048-6

Katsumoto, Y., Fukuchi-Mizutani, M., Fukui, Y., Brugliera, F., Holton, T. A., Karan, M., et al. (2007). Engineering of the rose flavonoid biosynthetic pathway successfully generated blue-hued flowers accumulating delphinidin. *Plant Cell Physiol.* 48, 1589–1600. doi: 10.1093/pcp/pcm131

Koes, R., Verweij, W., and Quattrocchio, F. (2005). Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. *Trends Plant Sci.* 10, 236–242. doi: 10.1016/j.tplants.2005.03.002

Lin, X., Xiao, M., Luo, Y., Wang, J., and Wang, H. (2013). The effect of RNAi-induced silencing of *FaDFR* on anthocyanin metabolism

in strawberry (*Fragaria*×*ananassa*) fruit. *Sci. Hortic.* 160, 123–128. doi: 10.1016/j.scienta.2013.05.024

Liu, Y., Qi, Y., Zhang, A., Wu, H., Liu, Z., and Ren, X. (2019). Molecular cloning and functional characterization of *AcGST1*, an anthocyanin-related glutathione S-transferase gene in kiwifruit (*Actinidia chinensis*). *Plant Mol. Biol.* 100, 451–465. doi: 10.1007/s11103-019-00870-6

Liu, Z., Shi, M. Z., and Xie, D. Y. (2014). Regulation of anthocyanin biosynthesis in *Arabidopsis thaliana* red pap1-D cells metabolically programmed by auxins. *Planta* 239, 765–781. doi: 10.1007/s00425-013-2011-0

Lo Piero, A. R. (2015). The state of the art in biosynthesis of anthocyanins and its regulation in pigmented sweet oranges [(*Citrus sinensis*) L. *Osbeck*]. *J. Agric. Food Chem.* 63, 4031–4041. doi: 10.1021/acs.jafc.5b01123

Lou, Q., Liu, Y., Qi, Y., Jiao, S., Tian, F., Jiang, L., et al. (2014). Transcriptome sequencing and metabolite analysis reveals the role of delphinidin metabolism in flower colour in grape hyacinth. *J. Exp. Bot.* 65, 3157–3164. doi: 10.1093/jxb/eru168

Ma, K., Zhang, Q., Cheng, T., Yan, X., Pan, H., and Wang, J. (2018). Substantial epigenetic variation causing flower color chimerism in the ornamental tree *Prunus mume* revealed by single base resolution methylome detection and Transcriptome Sequencing. *Int. J. Mol. Sci.* 19, 2315. doi: 10.3390/ijms19082315

Manetas, Y. (2006). Why some leaves are anthocyanic and why most anthocyanic leaves are red? *Flora Morphol. Distribut. Funct. Ecol. Plants* 201, 163–177. doi: 10.1016/j.flora.2005.06.010

Marinova, K., Pourcel, L., Weder, B., Schwarz, M., Barron, D., Routaboul, J. M., et al. (2007). The *Arabidopsis* MATE transporter TT12 acts as a vacuolar flavonoid/H+ -antiporter active in proanthocyanidin-accumulating cells of the seed coat. *Plant Cell* 19, 2023–2038. doi: 10.1105/tpc.106.046029

Marrs, K. A., Alfenito, M. R., Lloyd, A. M., and Walbot, V. (1995). A glutathione S-transferase involved in vacuolar transfer encoded by the maize gene Bronze-2. *Nature* 375, 397–400. doi: 10.1038/375397a0

Martens, S., Preuss, A., and Matern, U. (2010). Multifunctional flavonoid dioxygenases: flavonol and anthocyanin biosynthesis in *Arabidopsis thaliana* L. *Phytochemistry* 71, 1040–1049. doi: 10.1016/j.phytochem.2010.04.016

Meyer, H. J., and Van Staden, J. (1995). The *in vitro* production of an anthocyanin from callus cultures of *Oxalis linearis*. *Plant Cell Tissue Organ Cult.* 55–58 doi: 10.1007/BF00041119

Mol, J., Grotewold, E., and Koes, R. (1998). How genes paint flowers and seeds. *Trends Plant Sci.* 3, 212–217. doi: 10.1016/S1360-1385(98)01242-4

Nakatsuka, T., Haruta, K. S., Pitaksutheepong, C., Abe, Y., Kakizaki, Y., Yamamoto, K., et al. (2008). Identification and characterization of R2R3-MYB and bHLH transcription factors regulating anthocyanin biosynthesis in gentian flowers. *Plant Cell Physiol.* 49, 1818–1829. doi: 10.1093/pcp/pcn163

Peng, Y., Lin-Wang, K., Cooney, J. M., Wang, T., Espley, R. V., and Allan, A. C. (2019). Differential regulation of the anthocyanin profile in purple kiwifruit (*Actinidia species*). *Hortic Res* 6, 3. doi: 10.1038/s41438-018-0076-4

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Qiao, Z., Liu, S., Zeng, H., Li, Y., Wang, X., Chen, Y., et al. (2019). Exploring the molecular mechanism underlying the stable purple-red leaf phenotype in *Lagerstroemia indica* cv. Ebony Embers. *Int. J. Mol. Sci.* 20, 5636. doi: 10.3390/ijms20225636

Quattrocchio, F. M., Verweij, C. W., Kroon, A. R., Spelt, C. E., Mol, J. N.M., and Koes, R. E. (2006). PH4 of *Petunia* is an R2R3 MYB protein that activates vacuolar acidification through interactions with basic-helix-loop-helix transcription factors of the anthocyanin pathway. *Plant Cell* 18, 1274–1291. doi: 10.1105/tpc.105.034041

Slatnar, A., Mikulic-Petkovsek, M., Veberic, R., Stampar, F., and Schmitzer, V. (2013). Anthocyanin and chlorophyll content during poinsettia bract development. *Sci. Hortic.* 150, 142–145. doi: 10.1016/j.scienta.2012.10.014

Solfanelli, C., Poggi, A., Loreti, E., Alpi, A., and Perata, P. (2006). Sucrose-specific induction of the anthocyanin biosynthetic pathway in *Arabidopsis*. *Plant Physiol.* 140, 637–646. doi: 10.1104/pp.105.072579

Stich, K, E. T., and Wurst, F. (1992). Flavonol synthase activity and the regulation of flavonol and anthocyanin biosynthesis during flower development in *Dianthus caryophyllus* L. (Carnation). *Ztschrift Für Naturforschung C* 47. doi: 10.1515/znc-1992-7-811

Sun, W., Li, C., Wang, L., Dai, S., and Xu, Y. (2009). Anthocyanins present in flowers of *Senecio cruentus* with different colors. *Acta Hortic. Sin.* 36, 1775–1782. doi: 10.16420/j.issn.0513-353x.2009.12.011

Takahashi, R., Benitez, E. R., Oyoo, M. E., Khan, N. A., and Komatsu, S. (2011). Nonsense mutation of an MYB transcription factor is associated with purple-blue flower color in soybean. *J. Hered.* 102, 458–463. doi: 10.1093/jhered/esr028

Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant J.* 54, 733–749. doi: 10.1111/j.1365-313X.2008.03447.x

Tian, J., Han, Z. Y., Zhang, J., Hu, Y., Song, T., and Yao, Y. (2015). The balance of expression of dihydroflavonol 4-reductase and flavonol synthase regulates flavonoid biosynthesis and red foliage coloration in crabapples. *Sci. Rep.* 5, 12228. doi: 10.1038/srep12228

Tohge, T., Nishiyama, Y., Hirai, M. Y., Yano, M., Nakajima, J., Awazuhara, M., et al. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* 42, 218–235. doi: 10.1111/j.1365-313X.2005.02371.x

Uematsu, C., Katayama, H., Makino, I., Inagaki, A., Arakawa, O., and Martin, C. (2014). *Peace*, a MYB-like transcription factor, regulates petal pigmentation in flowering peach 'Genpei' bearing variegated and fully pigmented flowers. *J. Exp. Bot.* 65, 1081–1094. doi: 10.1093/jxb/ert456

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wang, Z., Cui, Y., Vainstein, A., Chen, S., and Ma, H. (2017). Regulation of fig (*Ficus carica* L.) fruit color: metabolomic and transcriptomic analyses of the flavonoid biosynthetic pathway. *Front. Plant Sci.* 8, 1990. doi: 10.3389/fpls.2017.01990

Wei, J., Wu, H., Zhang, H., Li, F., Chen, S., Hou, B., et al. (2018). Anthocyanins inhibit high glucose-induced renal tubular cell apoptosis caused by oxidative stress in db/db mice. *Int. J. Mol. Med.* 41, 1608–1618. doi: 10.3892/ijmm.2018.3378

Wu, Q., Wu, J., Li, S. S., Zhang, H. J., Feng, C. Y., Yin, D. D., et al. (2016). Transcriptome sequencing and metabolite analysis for revealing the blue flower formation in waterlily. *BMC Genomics* 17, 897. doi: 10.1186/s12864-016-3226-9

Wu, X., Gong, Q., Ni, X., Zhou, Y., and Gao, Z. (2017). UFGT: the key enzyme associated with the petals variegation in Japanese apricot. *Front. Plant Sci.* 8, 108. doi: 10.3389/fpls.2017.00108

Xu, W., Grain, D., Le Gourrierec, J., Harscoet, E., Berger, A., Jauvion, V., et al. (2013). Regulation of flavonoid biosynthesis involves an unexpected complex transcriptional regulation of TT8 expression, in *Arabidopsis*. *New Phytol.* 198, 59–70. doi: 10.1111/nph.12142

Yamagishi, M. (2010). Oriental hybrid lily *Sorbonne* homologue of LhMYB12 regulates anthocyanin biosyntheses in flower tepals and tepal spots. *Mol. Breed.* 28, 381–389. doi: 10.1007/s11032-010-9490-5

Zhang, Q., Zhang, H., Sun, L., Fan, G., Ye, M., Jiang, L., et al. (2018). The genetic architecture of floral traits in the woody plant *Prunus mume*. *Nat. Commun.* 9, 1702. doi: 10.1038/s41467-018-04093-z

Zhao, C., Guo. W., and Chen J. (2004). Research advances in the flower color of *Prunus mume*. *Journal of Beijing Forestry University* S1, 123–127.

Zhao, J., Huhman, D., Shadle, G., He, X. Z., Sumner, L. W., Tang, Y., et al. (2011). MATE2 mediates vacuolar sequestration of flavonoid glycosides and glycoside malonates in *Medicago truncatula*. *Plant Cell* 23, 1536–1555. doi: 10.1105/tpc.110.080804

Zhou, C., Mei, X., Rothenberg, D. O., Yang, Z., Zhang, W., Wan, S., et al. (2020). Metabolome and transcriptome analysis reveals putative genes involved in anthocyanin accumulation and coloration in white and pink tea (*Camellia sinensis*) flower. *Molecules* 25, 190. doi: 10.3390/molecules25010190

Zhu, Q., Sui, S., Lei, X., Yang, Z., Lu, K., Liu, G., et al. (2015). Ectopic expression of the coleus R2R3 MYB-type proanthocyanidin regulator gene *SsMYB3* alters the flower color in transgenic tobacco. *PLoS ONE* 10, e0139392. doi: 10.1371/journal.pone.0139392

Zhuang, H., Lou, Q., Liu, H., Han, H., Wang, Q., Tang, Z., et al. (2019). Differential regulation of anthocyanins in green and purple turnips revealed by combined *de novo* transcriptome and metabolome analysis. *Int. J. Mol. Sci.* 20, 4387. doi: 10.3390/ijms20184387

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflictof interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 23 May 2022
doi: 10.3389/fpls.2022.885503

Transcriptome-Wide Characterization of Alkaloids and Chlorophyll Biosynthesis in Lotus Plumule

Heng Sun[1,2], *Heyun Song*[1,3], *Xianbao Deng*[1,2], *Juan Liu*[1,2], *Dong Yang*[1,2], *Minghua Zhang*[1,3], *Yuxin Wang*[1,3], *Jia Xin*[1,3], *Lin Chen*[4], *Yanling Liu*[1,2]* and *Mei Yang*[1,2]*

[1] *Aquatic Plant Research Center, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China,* [2] *Hubei Key Laboratory of Wetland Evolution and Ecological Restoration, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China,* [3] *University of Chinese Academy of Sciences, Beijing, China,* [4] *Center of Applied Biotechnology, Wuhan Institute of Bioengineering, Wuhan, China*

Lotus plumule is a green tissue in the middle of seeds that predominantly accumulates bisbenzylisoquinoline alkaloids (bis-BIAs) and chlorophyll (Chl). However, the biosynthetic mechanisms of these two metabolites remain largely unknown in lotus. This study used physiological and RNA sequencing (RNA-Seq) approaches to characterize the development and molecular mechanisms of bis-BIAs and Chl biosynthesis in lotus plumule. Physiological analysis revealed that exponential plumule growth occurred between 9 and 15 days after pollination (DAP), which coincided with the onset of bis-BIAs biosynthesis and its subsequent rapid accumulation. Transcriptome analysis of lotus plumule identified a total of 8,725 differentially expressed genes (DEGs), representing ∼27.7% of all transcripts in the lotus genome. Sixteen structural DEGs, potentially associated with bis-BIAs biosynthesis, were identified. Of these, 12 encoded *O*-methyltransferases (OMTs) are likely involved in the methylation and bis-BIAs diversity in lotus. In addition, functionally divergent paralogous and redundant homologous gene members of the BIAs biosynthesis pathway, as well as transcription factors co-expressed with bis-BIAs and Chl biosynthesis genes, were identified. Twenty-two genes encoding 16 conserved enzymes of the Chl biosynthesis pathway were identified, with the majority being significantly upregulated by Chl biosynthesis. Photosynthesis and Chl biosynthesis pathways were simultaneously activated during lotus plumule development. Moreover, our results showed that light-driven Pchlide reduction is essential for Chl biosynthesis in the lotus plumule. These results will be useful for enhancing our understanding of alkaloids and Chl biosynthesis in plants.

Keywords: lotus plumule, bisbenzylisoquinoline alkaloids, chlorophyll, biosynthetic mechanism, transcriptome analysis

OPEN ACCESS

Edited by:
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Fei Zhou,
Nanjing University, China
Haiyang Xu,
Chongqing University, China

***Correspondence:**
Mei Yang
yangmei815815@wbgcas.cn
Yanling Liu
liuyanling@wbgcas.cn

Specialty section:
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

Received: *28 February 2022*
Accepted: *20 April 2022*
Published: *23 May 2022*

Citation:
Sun H, Song H, Deng X, Liu J, Yang D, Zhang M, Wang Y, Xin J, Chen L, Liu Y and Yang M (2022) Transcriptome-Wide Characterization of Alkaloids and Chlorophyll Biosynthesis in Lotus Plumule. Front. Plant Sci. 13:885503. doi: 10.3389/fpls.2022.885503

INTRODUCTION

Lotus is a perennial aquatic plant in the family Nelumbonaceae that contains a single genus, *Nelumbo*, with two extant species: *Nelumbo nucifera* Gaertn. and *Nelumbo lutea* Pers (Wang et al., 2013). In Asia, lotus is an old domesticated herbaceous crop with versatile uses that are classified as seed-, rhizome-, and flower-lotus based on varieties (Yang et al., 2015). Lotus seed is not only

an important reproductive organ consisting of the pericarp, seed coat, cotyledon, and plumule, but is also a rich source of nutrients and bioactive compounds with medicinal properties. The plumule, also known as *Lianzixin*, is a common traditional Chinese medicine with important pharmacological properties, such as antihypertensive, antiarrhythmic, and diuretic (Liu et al., 2017).

Alkaloids are a class of alkaline organic nitrogen compounds in plants (Liu et al., 2019). Lotus tissues, such as leaf, plumule, and petal are rich in benzylisoquinoline alkaloids (BIAs) (Deng et al., 2016). Bisbenzylisoquinoline alkaloids (bis-BIAs) are structural dimers of 1-benzylisoquinolines and are important bioactive components that are predominantly accumulated in the lotus plumule especially liensinine, isoliensinine, and neferineare (Deng et al., 2016). Previous studies have only focused on the identification, separation, purification, and pharmacological effects of bis-BIAs in lotus plumule (Deng et al., 2016; Chen et al., 2019; Liu et al., 2019). However, the molecular mechanisms underlying the biosynthesis of bis-BIAs in the lotus plumule remain largely unknown. BIAs are synthesized through a common pathway derived from the L-tyrosine substrate in plants. The substrate is subsequently catalyzed by tyrosine/DOPA decarboxylase (TYDC), norcoclaurine synthases (NCS), norcoclaurine 6-O-methyltransferase (6OMT), coclaurine N-methyltransferase (CNMT), (*S*)-N-methylcoclaurine-3′-hydroxylase (CYP80B), and 3′-hydroxy-N-methylcoclaurine 4′-O-methyltransferase (4′OMT) to produce (*S*)-reticuline, which is the common precursor of most BIAs (Ziegler and Facchini, 2008; Hagel and Facchini, 2013). In addition, bis-BIAs are produced *via* the catalysis of *N*-methylcoclaurine by the P450 enzyme CYP80A1 (Ziegler and Facchini, 2008). Notwithstanding, the alkaloids biosynthetic pathways vary greatly in different plants; thus, identification of key structural genes and determination of the biosynthetic mechanism of bis-BIAs in lotus plumule is necessary.

Unlike in many angiosperms, lotus plumule displays a dim-light photosynthetic capacity and can synthesize Chl while still being enclosed by dense layers of seed integuments, such as pericarp, seed coat, and cotyledon (Shen-Miller, 2007). As the most abundant pigment in the plant kingdom, Chl is crucial for light harvesting and energy transduction during photosynthesis (Tripathy and Pattanayak, 2012). The Chl biosynthetic pathway has been well-elucidated in higher plants, with over 16 enzymes and enzymatic steps responsible for this process identified and characterized (Tripathy and Pattanayak, 2012). Chl is synthesized through a complex pathway derived from the biosynthesis of the 5-aminolevulinic acid (ALA) precursor. Of the 16 enzymes reported to be involved in Chl biosynthesis, the conversion of protochlorophyllide (Pchlide) to chlorophyllide (Chlide) by the light-dependent Pchlide oxidoreductase (LPOR) is the only light-requiring reaction in angiosperms (Yamamoto et al., 2017). The LPOR encoding genes are nuclear-encoded and are distributed throughout oxygenic photosynthetic organisms. In contrast, gymnosperms employ an alternative Pchlide reduction reaction catalyzed by light-independent Pchlide oxidoreductase (DPOR) (Reinbothe et al., 2010). DPOR is encoded in the chloroplast genome by three genes, *chlL*, *chlN,* and *chlB*. To date, little is still known about Chl biosynthesis in basal eudicots including lotus. The plumule provides a model system for studying and improving our understanding of the mechanism of Chl biosynthesis in lotus and other basal eudicots.

The RNA-Seq has become a popular tool for uncovering the underlying molecular mechanisms of biological processes, including development, stress response, and metabolism processes in recent years (Fracasso et al., 2016; Goyal et al., 2016; Yang et al., 2017; Lanver et al., 2018; Xia et al., 2020; Sun et al., 2021). For example, the molecular mechanism of alkaloids biosynthesis has been clarified in numerous plants using RNA-seq (Guo et al., 2013; Cui et al., 2015; He et al., 2017; Deng et al., 2018). Due to its efficiency, this study used RNA-Seq technology to reveal the dynamic changes in gene expression during lotus plumule development. The result showed significant variations in alkaloid contents during plumule development and identification of structural genes likely associated with bis-BIAs biosynthesis, which were analyzed. In addition, the results clarified that the light-dependent Chl biosynthetic pathway in the lotus plumule. This study will expand our understanding of BIAs and Chl biosynthesis in plants.

MATERIALS AND METHODS

Lotus Plumule Collection

Lotus cultivars were grown in the experimental field at Wuhan Botanical Garden (Wuhan, China). Lotus plumules were collected at 9, 12, 15, and 18 DAP from the cultivar "Jianxuan17" (JNP) and at 12, 15, and 18 DAP from the cultivar "China Antique" (CNP). Samples were immediately frozen in liquid nitrogen and then stored at −80°C until use.

RNA Extraction and Sequencing

The plumule samples were ground into powder in liquid nitrogen, and total RNA was extracted using the Plant Total RNA Isolation Kit (Beijing Zoman Biotechnology Co., Ltd., Beijing, China). Illumina platform was used to sequence 21 high-quality RNA libraries at Biomarker Technologies Corporation (Beijing, China). The resulting clean data has been deposited at the National Center for Biotechnology Information (NCBI) with accession number, PRJNA747903.

Analysis of RNA Sequencing Data

After removing adaptors and low-quality sequence reads, clean reads were mapped to the lotus reference genome sequence (Ming et al., 2013). The fragments per kilobase of transcript per million fragments mapped (FPKM) was calculated to quantify gene expression levels, and genes that met the Fold Change (FC) ≥ 2 and False Discovery Rate (FDR) <0.01 criteria were assigned as differentially expressed (DEGs).

Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed by KOBAS 3.0, and Gene Ontology (GO) enrichment analysis was implemented by the GOseq R packages (Bu et al., 2021). K-means analysis of gene expression was performed using Genesis software

(Sturn et al., 2002). Principal component analysis (PCA), correlation analysis of libraries, and gene expression between the 510 TFs and 8,725 DEGs identified were performed using BMKCloud programs at www.biocloud.net. TFs were identified using PlantTFDB v5.0 (http://planttfdb.gao-lab.org/). Venn diagram, gene chromosomal location, synteny analysis, and heatmaps were visualized with TBtools software (Chen et al., 2020). The phylogenetic tree was constructed using MEGA7 software (Kumar et al., 2016). All protein sequences of bis-BIAs and Chl biosynthesis genes are listed in **Supplementary Table 1**.

qRT-PCR Analysis

High-quality RNAs were reverse transcribed to cDNA using TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix Kit (Lot#M31212, Beijing TransGen Biotech Co., Ltd., Beijing, China). Primers were designed using Primer Premier 5.0 and synthesized commercially (Huayu Gene, Wuhan, China). The qRT-PCR was performed using StepOnePlus Real-time PCR System (Applied Biosystems, USA) according to the protocol described by Deng et al. (2018). The *NnACTIN* (Gene ID NNU_24864) was used as the internal control to normalize the gene expression level. All primer sequences used are listed in **Supplementary Table 2**.

Measurement of Alkaloid Content

Extraction and quantification of alkaloids in the lotus plumule were performed according to the protocol described by Deng et al. (2016). Briefly, fresh lotus plumule samples were ground to a fine powder in liquid nitrogen followed by extraction of alkaloids using 0.3 M HCl-methanol, 1:1, v/v extraction buffer. Quantification of alkaloid extracts was performed using high-performance liquid chromatography (HPLC, Agilent Technologies, USA).

Measurement of Chl Content

Chl extraction and quantification were performed as previously described (Morley et al., 2020). Fresh lotus plumule samples were ground to a fine powder using liquid nitrogen followed by Chl extraction using 80% aqueous acetone. The absorption wavelength was set to 663 nm (A663) and 646 nm (A646), and detection was performed with an Infinite M200 Luminometer (Tecan, Mannerdorf, Switzerland). Chlorophyll *a* and Chlorophyll *b* were calculated according to the following equations, which were then summed to represent the total leaf Chl content.

$$\text{Chlorophyll } a = 12.21 * \text{A663} - 2.81 * \text{A646}$$
$$\text{Chlorophyll } b = 20.13 * \text{A646} - 5.03 * \text{A663}$$

Light and Dark Treatment

The seed-lotus cultivar, "Jianxuan17," was grown in the experimental field at Wuhan Botanical Garden (Wuhan, China). Aluminum foil was then used to tightly wrap pods at 3 DAPs in July, and then, the procedure was repeated in August. The unwrapped pods were used as a control group. Pods were collected at 12, 15, and 18 DAP to analyze the effects of light/dark treatment on Chl biosynthesis in the lotus plumule.

Statistical Analysis

Physiological data were statistically assessed by one-way ANOVA *via* IBM SPSS Statistics 20.0 software (SPSS Inc, USA), and significant differences in means were assessed with the least significant difference (LSD) test at $p = 0.05$.

RESULTS

Morphological Changes, Chlorophyll, and Alkaloids Content During Lotus Plumule Development

The plumule is located within the lotus seed, enclosed by layers of integuments, including pericarp, seed coat, and cotyledon (**Figure 1A**). Significant morphological changes in size, weight, and color were observed between 9 and 24 DAP (**Figure 1B**). For example, the plumule of lotus cv. "Jianxuan 17" (hereafter abbreviated, JNP) showed a rapid increase in length from 9 to 15 DAP, followed by a slow increase to about 1.74 cm at 24 DAP (**Figure 1C**). JNP weight increased continuously to a highest of 0.117 g at 21 DAP and then decreased to 0.099 g at 24 DAP (**Figure 1D**). In contrast, the plumule of lotus cv. "China Antique" (hereafter abbreviated, CNP) showed a comparatively smaller size, with a recorded weight of 0.063 g at 21 DAP, which represented a 53.9% decrease in comparison to that of JNP (**Supplementary Figures 1A–C**). In addition, rapid synthesis and accumulation of Chl were detected from 12 to 21 DAP, which was consistent with the observed change in plumule color (**Figure 1E**).

The HPLC identification of alkaloid components in lotus plumule showed that JNP mainly accumulated liensinine, isoliensinine, and neferine, with isoliensinine and neferine being detected at 15 DAP but liensinine being detected at 18 DAP (**Figure 1F** and **Supplementary Figure 2**). The total alkaloid content in JNP increased from 187.24 μg/g at 15 DAP to 5,130.1 μg/g at 21 DAP, with neferine as the most dominant bis-BIA. In contrast, the total alkaloid content in CNP increased from 67.28 μg/g at 15 DAP to 4,254.6 μg/g at 21 DAP, with liensinine and neferine as the main components detected (**Supplementary Figures 1D, 2**). These results indicate obvious variation in the plumule alkaloid components of the two tested lotus varieties.

Transcriptome Profiling of Lotus Plumule During Development

To investigate the molecular mechanisms of lotus plumule development, 21 RNA libraries, including JNP at 9, 12, 15, and 18 DAP and CNP at 12, 15, and 18 DAP were constructed. A total of ∼588.65 million paired-end clean reads were obtained after conducting quality control of sequencing data. The average GC content was 46.39%, and the average ≥ Q30 (the percentage base which the quality value of clean data is ≥30) of each library was 95.59% (**Supplementary Table 3**). Approximately 95.56% of the clean reads were mapped to

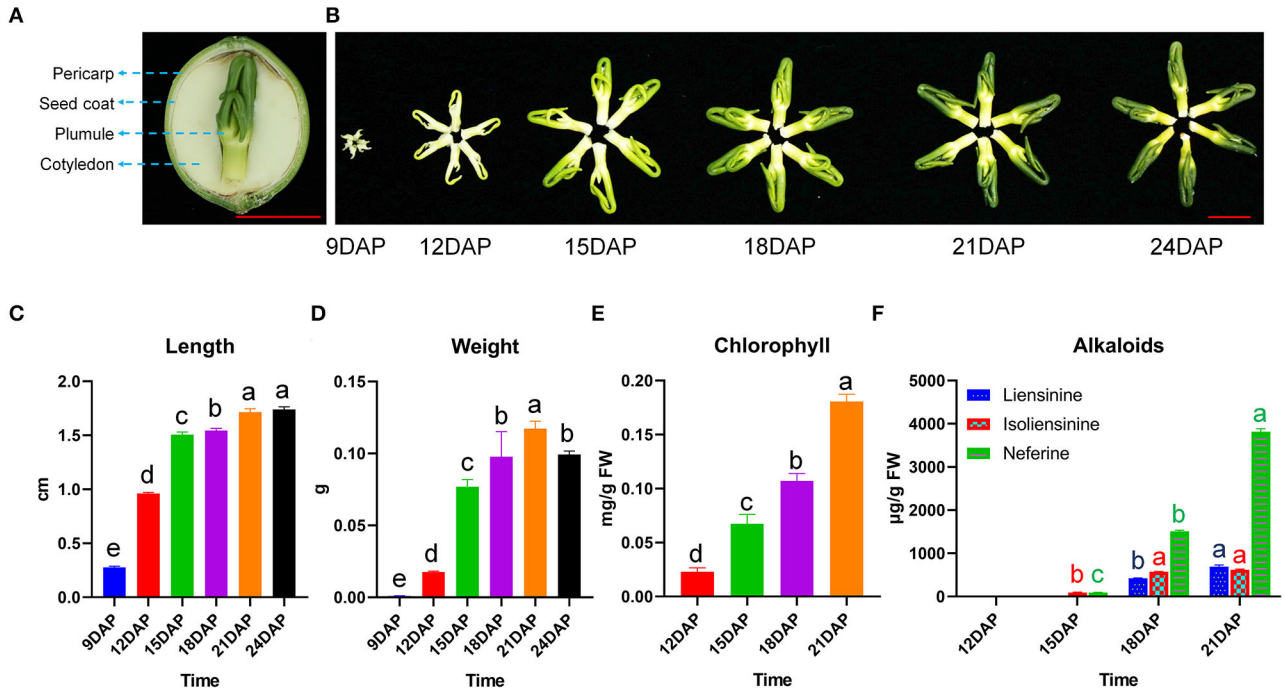


FIGURE 1 | Phenotypic changes during lotus plumule development in cv. "Jianxuan 17". **(A)** The structure of lotus seed. **(B)** Morphological changes in lotus plumule during development. The red bars represent 1 cm. **(C)** Length, and **(D)** Weight of lotus plumule at different developmental stages. Error bars represent means ± standard error ($n = 3$). **(E)** Chlorophyll, and **(F)** Alkaloid contents in lotus plumule at different developmental stages. Error bars represent means ± standard error ($n = 4$). Statistical significance is based on the Least Significant Difference (LSD) test at $P < 0.05$, one-way ANOVA. Different letters over bars indicate significant difference.

29,568 genes in the reference genome of "Chinese Antique" (Ming et al., 2013), and 20,455 genes with FPKM expression > 1 in at least one sample were identified. Analysis of the overall distribution of gene expression levels in each sample revealed that FPKM of most genes were in the ranges of 1–10 and 1– 100 (**Figure 2A**). Principal component analysis (PCA) and correlation coefficient heatmap of samples showed that the three biological replicates were closely clustered (**Figure 2B** and **Supplementary Figure 3**).

Identification of Differentially Expressed Genes in Lotus Plumule

For JNP, a total of 1,578, 5,292, 6,786, 2,787, 4,936, and 1,307 DEGs were identified in the 9_vs._12 DAP, 9_vs._15 DAP, 9_vs._18 DAP, 12_vs._15 DAP, 12_vs._18 DAP, and 15_vs._18 DAP comparison groups, respectively (**Figure 2C**). Venn diagram showed that the highest number of common DEGs of 4,544 was between 9_vs._15 DAP and 9_vs._18 DAP groups, and the lowest number of common DEGs of 205 was between 9_vs._12 DAP and 15_vs._18 DAP groups (**Supplementary Figure 4A**). In addition, 100 common DEGs were identified in all the six comparison groups (**Supplementary Figure 4A**). For CNP, a total of 773, 4,693, and 3,001 DEGs were identified in the 12_vs._15 DAP, 12_vs._18 DAP, and 15_vs._18 DAP comparison groups, respectively (**Figure 2C**). In addition, 403 common DEGs were identified in all the three groups, with the highest DEG overlap of 2,748 being observed between 12_vs._18 DAP and 15_vs._18 DAP groups (**Supplementary Figure 4B**). Moreover, 3,726 and 610 DEGs were exclusively identified in JNP and CNP, respectively, while 4,389 common DEGs, accounting for 87.8% of all DEGs in CNP, were identified between the two lotus varieties, suggesting a high similarity in their plumule developmental process (**Figure 2D**).

Functional Enrichment of DEGs

The KEGG analysis was used to analyze the functional enrichment of common and exclusive DEGs between JNP and CNP. As a result, 4,389 common DEGs were significantly enriched in 42 KEGG pathways ($P \leq 0.05$), such as metabolic pathways, biosynthesis of secondary metabolites, carbon fixation in photosynthetic organisms, and photosynthesis (**Supplementary Table 4**). Specific DEGs from JNP or CNP were shown to be significantly enriched in 19 or 7 pathways, respectively, with those from JNP being involved in metabolic pathways, plant hormone signal transduction, purine metabolism, and arginine biosynthesis, while those from CNP, being involved in fatty acid elongation, fatty acid metabolism, and brassinosteroid biosynthesis (**Supplementary Table 4**).

A total of 8,115 DEGs identified in JPN were used to investigate the enriched pathways of DEGs with different expression patterns. These DEGs could be classified into three

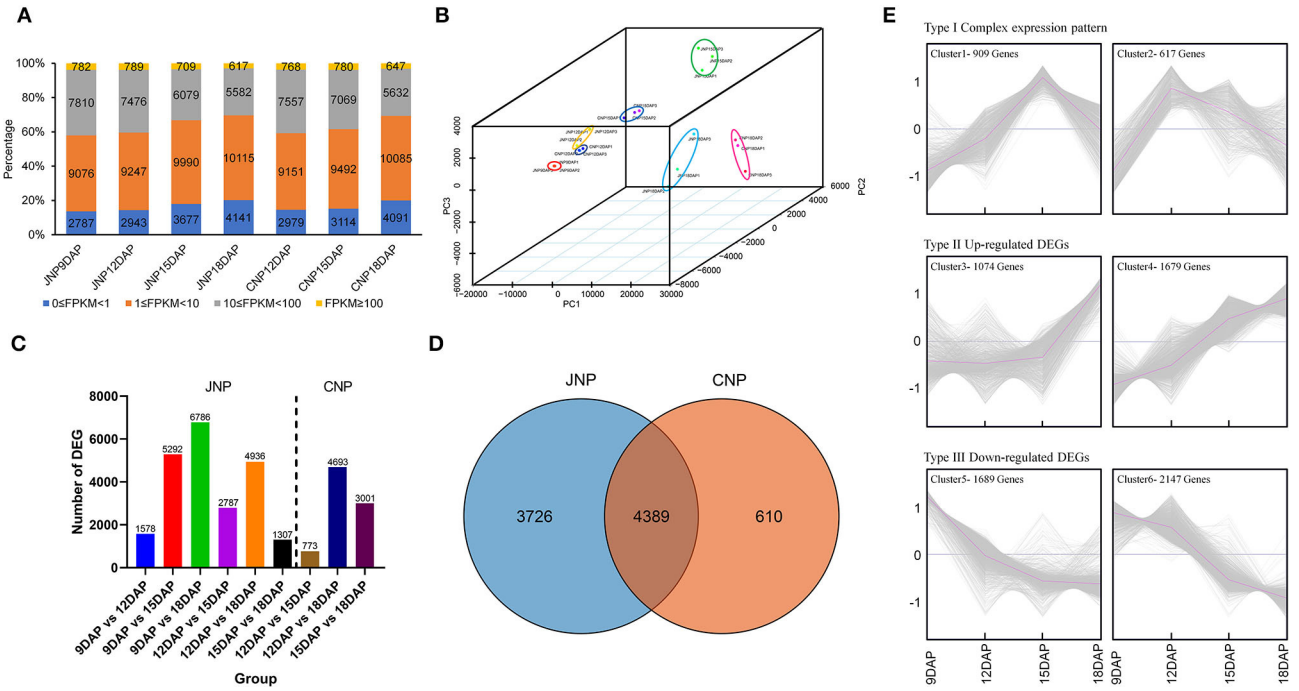


FIGURE 2 | Transcriptome profiling and identification of differentially expressed genes (DEGs) in lotus plumule. **(A)** The gene FPKM expression levels of each sample. **(B)** 3D-PCA plot showing sample clusters based on gene expression levels. **(C)** Number of DEGs between different comparison groups. JNP and CNP represent lotus plumule of cv. "Jianxuan 17" and cv. "China Antique," respectively. **(D)** Venn diagram showing the number of overlapping DEGs between JNP and CNP. **(E)** Cluster analysis of 8,115 DEGs identified in JPN based on the K-means method. The gray and purple lines in the panels indicate the expression levels of individual genes and the consensus of all genes in a specific subcluster.

categories based on expression patterns and six clusters using K-means clustering (**Figure 2E**). In the type I category, 1,526 DEGs showed complex expression patterns, with the highest expression abundance at 15 and 12 DAP in clusters 1 and 2, respectively. In cluster 1, 909 DEGs were enriched in 22 pathways, including metabolic pathways, photosynthesis, and isoquinoline alkaloid biosynthesis (**Supplementary Figure 5**). Most DEGs in cluster 2 were involved in metabolic pathways, biosynthesis of secondary metabolites, and biosynthesis of amino acids (**Supplementary Figure 5**). In the type II category, 2,753 DEGs showed an overall upregulated expression pattern, with cluster 4 showing a more continuous upregulated expression than cluster 3 (**Figure 2E**). Gene functional enrichment analysis indicated that more DEGs in the type II category were enriched in multiple metabolic pathways, such as biosynthesis of secondary metabolites, phenylpropanoid biosynthesis, and carotenoid biosynthesis for cluster 3, and in flavonoid biosynthesis, ubiquinone and another terpenoid-quinone biosynthesis, and starch and sucrose metabolism for cluster 4 (**Supplementary Figure 5**). In the type III category, 3,836 DEGs showed an overall downregulated expression pattern, with 10 and 29 KEGG pathways being, respectively enriched in clusters 5 and 6, such as metabolic pathways, plant hormone signal transduction, and purine metabolism for cluster 5, and DNA replication, cysteine, and methionine metabolism, and citrate cycle (TCA cycle) for cluster 6 (**Figure 2E** and **Supplementary Figure 5**).

Identification of Key Structural Genes in bis-BIAs Biosynthetic Pathway

The common bis-BIAs biosynthetic pathway is derived from L-tyrosine metabolism (Hagel and Facchini, 2013). Thus, we initially analyzed the expression patterns of key genes involved in the tyrosine biosynthetic pathway (**Supplementary Figure 6**). As a result, upregulated expression of a key enzyme of the glycolysis pathway, *NnPFK* (NNU_10589), encoding ATP-dependent 6-phosphofructokinase, and a key regulatory enzyme of the pentose phosphate pathway (PPP), *NnG6PD* (NNU_02159), encoding glucose 6-phosphate dehydrogenase were observed during lotus plumule development (**Supplementary Figure 6B**). Similarly, upregulated expression of *NnSK* (NNU_20134), *NnCS* (NNU_13158), and *NnADH* (NNU_08507), encoding shikimate kinase, chorismate synthase, and arogenate dehydrogenase, respectively, were observed. In contrast, *NnDHD-SDH* (NNU_06891) and *NnPPA-AT* (NNU_20211), encoding bifunctional 3-dehydrogenate dehydratase/shikimate dehydrogenase and prephenate aminotransferase, respectively, were downregulated between 12 and 15 DAP. Notably, *NnCM* (NNU_04572) encoding chorismate mutase, which catalyzes the first committed step in the assembly of tyrosine, showed the

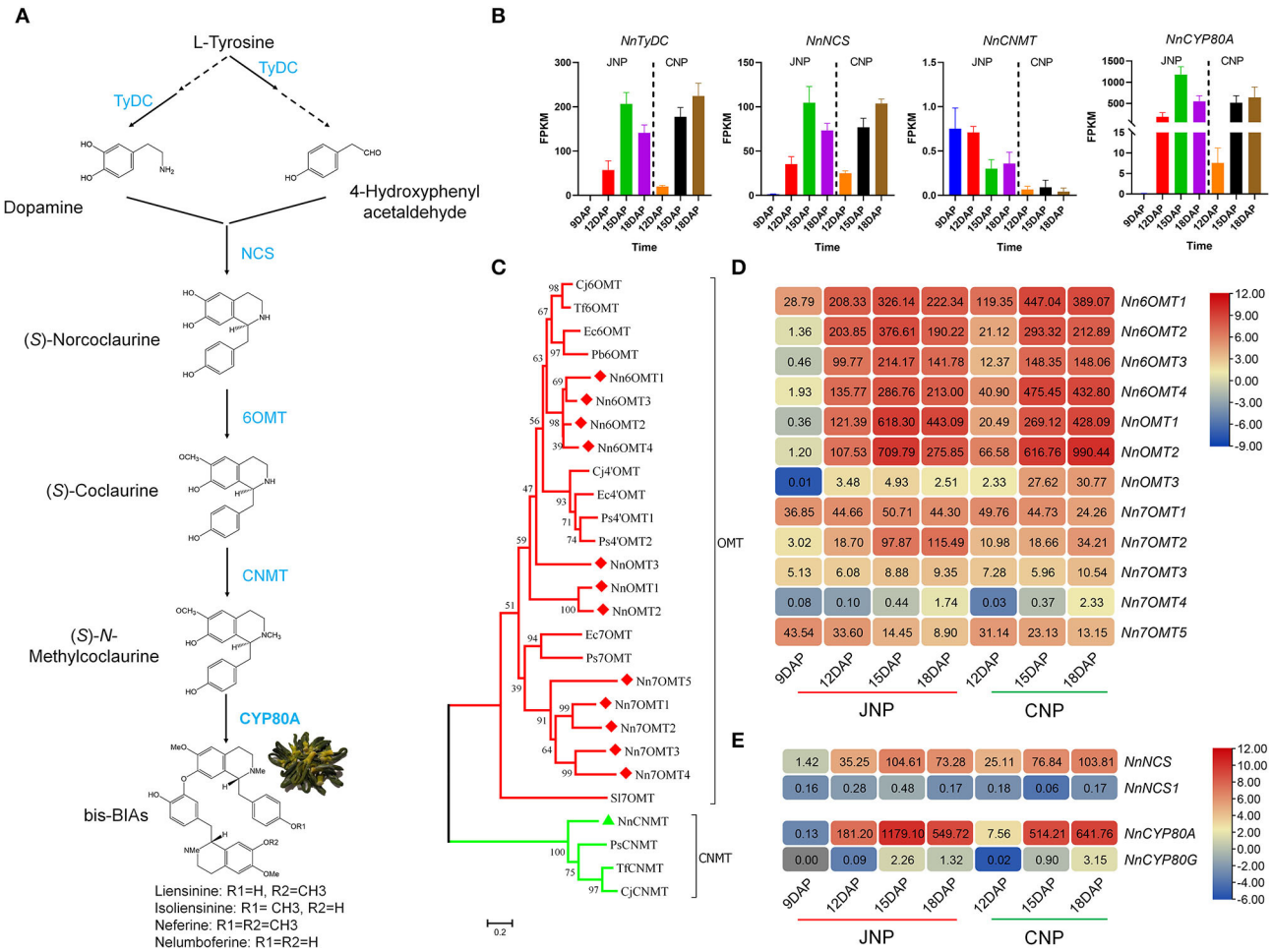


FIGURE 3 | Identification and expression analysis of bis-BIAs biosynthesis genes. **(A)** Key steps in the predicted bis-BIAs biosynthesis pathway in lotus plumule. **(B)** Expression patterns of *NnTyDC* (NNU_22559), *NnNCS* (NNU_14334), *NnCNMT* (NNU_11880), and *NnCYP80A* (NNU_21373) during lotus plumule development. **(C)** Phylogenetic tree depicting the relationship between differentially expressed *O*-methyltransferase (OMT) genes. GenBank Accession Number: Cj6OMT (BAB08004.1), Tf6OMT (ACO90238.1), Ec6OMT (BAM37634.1), Pb6OMT (ACO90225.1), Cj4'OMT (Q9LEL5.1), Ec4'OMT (BAM37633.1), Ps4'OMT1 (XP_026430708.1), Ps4'OMT2 (XP_026440860.1), Ec7OMT (BAE79723.1), Ps7OMT (XP_026440002), PsCNMT (Q7XB08.1), TfCNMT (Q5C9L6.1), and CjCNMT (BAB71802.1). **(D)** Expression patterns of OMT genes during lotus plumule development. **(E)** A heatmap showing expression differences between *NnNCS* and *NnCYP80* homolog members.

highest expression abundance at 12 and 15 DAP in JNP and CNP (**Supplementary Figure 6B**).

The putatively common BIAs biosynthetic pathway in the lotus plumule is shown in **Figure 3A**. DEGs associated with key lotus BIAs' biosynthetic pathways such as, *NnTyDC* (NNU_22559), *NnNCS* (NNU_14334), and *NnCYP80A* (NNU_21373) encoding TyDC, NCS, and CYP80A, respectively, were highly expressed in the plumule (**Figure 3B**). Phylogenetic analysis was performed to determine the evolutionary relationships with their homologous genes from other plant species (**Supplementary Figure 7A**). *NnTyDC*, *NnNCS*, and *NnCYP80A* were initially upregulated from 9 to 15 DAP and then downregulated at 18 DAP in JNP, whereas they were continuously upregulated until 18 DAP in CNP (**Figure 3B**). Notably, only a single gene copy of *NnCNMT* (NNU_11880) encoding CNMT was identified in the lotus genome with extremely low FPKM expression < 1 in the plumule (**Figure 3B**).

The *O*-methylation is a crucial step that catalyzes the *O*-methyltransferase (OMT) transfer of a methyl group to a hydroxyl group of an alkaloid substrate leading to the structural diversity of lotus BIAs (Morris and Facchini, 2019; Menendez-Perdomo and Facchini, 2020). Analysis of the structural formula showed that bis-BIAs were *O*-methylated at C6, C7, and C4′, suggesting that 6OMT, 7OMT, and 4′OMT could be involved in the biosynthesis of bis-BIAs in lotus plumule. Twelve differentially expressed OMT genes were identified in the lotus by homology alignment and confirmed by phylogenetic analysis (**Figure 3C**). Of these, four OMTs, the Nn6OMT1 (NNU_19035), Nn6OMT2 (NNU_23168), Nn6OMT3 (NNU_03166), and Nn6OMT4 (NNU_03165), were closely paired with the 6OMT genes from *Coptis japonica*, *Thalictrum flavum*, *Eschscholzia*

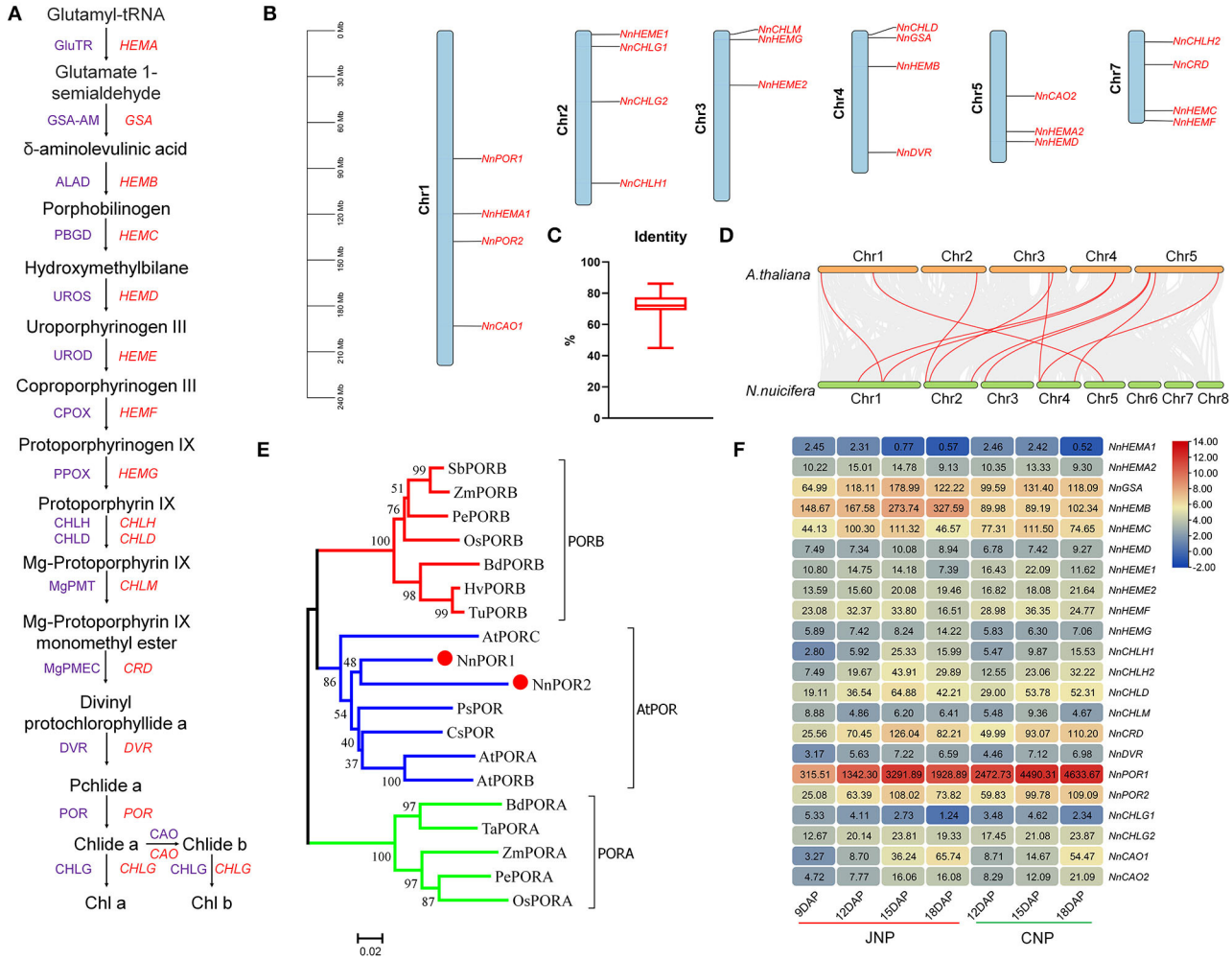


FIGURE 4 | Genome-wide identification and expression analysis of Chl biosynthesis genes in lotus. **(A)** Chl biosynthesis pathway in higher plants. Purple and red letters represent enzymes and genes, respectively. **(B)** Chromosomal localization of Chl biosynthesis genes in lotus. **(C)** The amino acid sequence identity between *Arabidopsis* and lotus Chl biosynthesis genes. **(D)** Synteny analysis of Chl synthesis genes between *Arabidopsis* and lotus. **(E)** Phylogenetic tree showing the relationship of lotus POR genes. GenBank Accession Number: SbPORB (XP_002467010.1), ZmPORB (NP_001149903.1), PePORB (PH01000789G0260), OsPORB (Q8W3D9.1), BdPORB (XP_003570527.1), HvPORB (Q42850.1), TuPORB (TuPORB), AtPORC (NP_001030948.1), PsPOR (CAA44786.1), CsPOR (BAA21089.1), AtPORA (AAC49043.1), AtPORB (NP_001031731.1), BdPORA (XP_010240614.1), TaPORA (Q41578.1), ZmPORA (NP_001167683.1), PePORA (PH01000087G0190), and OsPORA (Q7XKF3.1). **(F)** Expression patterns of Chl biosynthesis genes during lotus plumule development.

californica, and *Papaver bracteatum*, respectively. Similarly, five OMTs, including Nn7OMT1 (NNU_04966), Nn7OMT2 (NNU_04906), Nn7OMT3 (NNU_20903), Nn7OMT4 (NNU_20253), and Nn7OMT5 (NNU_16993), were closely clustered with 7OMT gene clusters from *Eschscholzia californica* and *Papaver somniferum*. The remaining three differentially expressed OMTs, including NnOMT1 (NNU_15801), NnOMT2 (NNU_15809), and NnOMT3 (NNU_25948), were both clustered with 6OMT and 4'OMT genes. Gene expression analysis showed that all differentially expressed OMTs except *Nn7OMT5* were upregulated from 9 to 15 DAP in JNP, with *Nn7OMT5* showing a continuously downregulated expression from 12 to 18 DAP in both JNP and CNP (**Figure 3D**). In addition, *NnOMT3* and *Nn7OMT2* genes exhibited obvious expression differences between JNP and CNP, with *NnOMT3* having a 12.3-fold increase in expression abundance in CNP than in JNP at 18 DAP (**Figure 3D**). Similarly, a 3.4-fold increase in the expression abundance of *Nn7OMT2* (NNU_04906) was observed in JNP at 18 DAP.

Using five BIAs' biosynthesis-related genes, the qRT-PCR analysis, including *NnTyDC*, *Nn6OMT1*, *Nn6OMT2*, *Nn6OMT3*, and *NnCYP80A*, was further conducted to validate the RNA-Seq data. As a result, a higher correlation between RNA-Seq data and qRT-PCR results was observed, suggesting strong RNA-Seq data reliability in this study (**Supplementary Figure 7B**). In addition, the determination of expression levels of bis-BIAs biosynthesis genes in other lotus tissues, including leaf, petiole, rhizome, and root using the publicly available transcriptome data (Shi et al.,

2020), revealed that most genes were highly expressed in leaf (**Supplementary Figure 8**). Interestingly, highly expressed genes in leaf tissues included *Nn6OMT2*, *NnCYP80A*, and *NnOMT1*, thus, suggesting that some bis-BIAs biosynthesis genes could also be involved in the biosynthesis of aporphine-type BIAs in lotus leaves.

BIAs' Biosynthesis Gene Pairs Show Functional Redundancy and Divergence Between Paralogous Members

Gene mapping analysis identified 16 bis-BIAs biosynthesis genes distributed across six lotus chromosomes (Chr), with seven genes localized on Chr1 (**Supplementary Figure 9A**). Notably, possible gene duplication events in some OMTs were observed. For example, *NnOMT1* and *NnOMT2* genes located in a 42.2 kb region on Chr1 shared about 83.4% amino acid sequence identity (**Supplementary Figure 9B**). Similarly, *Nn6OMT2*, *Nn6OMT3*, and *Nn6OMT4* genes located in a 354.7kb interval on Chr1 shared about 79.6% identity (**Supplementary Figure 9C**). The observed high transcript abundance of these OMT genes during plumule development suggested their functional redundancy and their synergistic interactions to accumulate bis-BIAs in lotus.

In addition, paralogs of two structural genes involved in alkaloid synthesis were identified, suggesting their likely functional divergence in lotus. For example, *NnNCS1* (NNU_21731) and its *NnNCS* homolog shared about 59.3% amino acid sequence identity (**Supplementary Figure 9D**). However, the expression of *NnNCS1* was extremely low in the lotus plumule (**Figure 3E**). Moreover, *NnCYP80A* and its homolog *NnCYP80G* (NNU_21372) located within a 26.1-kb interval shared about 59.18% identity, with the latter showing an extremely low expression level in lotus plumule (**Figure 3E** and **Supplementary Figure 9E**). Overall, these results suggest a degree of functional specialization between the paralogs of *NnNCS* and *NnCYP80* genes.

Genome-Wide Identification of Chl Biosynthesis Genes in Lotus

The Chl biosynthesis is crucial for lotus plumule development, and the process was accompanied by changes in color from light yellow at 9 DAP to dark green at 18 DAP (**Figures 1B,E**). Twenty-two genes encoding 16 key enzymes in the Chl biosynthesis pathway were found distributed across six chromosomes in the lotus genome (**Figures 4A,B**). The average amino acid sequence identity between the lotus and *Arabidopsis* homologous Chl genes was 71.6% (**Figure 4C**). In addition, collinearity analysis between lotus and *Arabidopsis* genes identified 11 colinear gene pairs, encoding GSA-AM (*NnGSA*, NNU_22236), UROD (*NnHEME1*, NNU_12265), CHLG (*NnCHLG1*, NNU_12622), POR (*NnPOR1*, NNU_01188; *NnPOR2*, NNU_16195), CHLH (*NnCHLH1*, NNU_03121), PPOX (*NnHEMG*, NNU_02121), DVR (*NnDVR*, NNU_06919), and CAO (*NnCAO2*, NNU_24327), which suggested that the Chl biosynthesis pathway is conserved between the two plants (**Figure 4D**).

The Pchlide reduction is catalyzed by pchlide oxidoreductase (POR) and presents the penultimate step in the Chl biosynthesis pathway. Here, two genes, *NnPOR1* and *NnPOR2*, encoding light-dependent Pchilde oxidoreductase (LPOR) were identified in the lotus, and phylogenetic analysis showed their close pairing with *Arabidopsis* POR genes (**Figure 4E**). In addition, multiple sequence alignments revealed that all *NnPOR* genes contained a conserved NADPH-binding motif, TGASSGLG, and an active YKDSK site motif (**Supplementary Figure 10**).

The expression patterns of lotus Chl biosynthesis genes were similar between JNP and CNP, with 20 genes showing differential expression profiles during plumule development (**Figure 4F**). Thirteen genes exhibited upregulated profiles from 9 to 15 DAP, which was later downregulated at 18 DAP, such as *NnGSA*, *NnHEMC* (NNU_01206), and *NnCRD* (NNU_10837). Four genes, including *NnHEMB* (NNU_04375), *NnHEMG,* and *NnCAO* (NNU_24327, NNU_19596) were continuously upregulated throughout the tested lotus plumule developmental stages (**Figure 4F**). Using six Chl biosynthesis genes, the qRT-PCR analysis, including *NnPOR1*, *NnPOR2*, *NnCAO1 NnCHLG2*, *NnHEMB*, and *NnCHLD*, revealed a higher correlation between RNA-Seq data and qRT-PCR results at 9, 12, and 15 DAP (**Supplementary Figure 11A**). Tissue expression analysis showed that most Chl biosynthesis genes were preferentially expressed in lotus leaf, while *NnHEMF* (NNU_00797) and *NnHEMG* were highly expressed in non-photosynthetic tissues (**Supplementary Figure 11B**).

Activated Expression of Photosynthesis-Related Genes During Lotus Plumule Development

Photosynthesis in green plants is the process of transforming light energy into chemical energy. With the biosynthesis of Chl, the photosynthesis pathway was activated during lotus plumule development in JNP and CNP (**Figure 5** and **Supplementary Figure 12**). Forty-eight differentially expressed photosynthesis-related genes were detected, including 15 genes involved in light-harvesting chlorophyll-protein complex, 11 Photosystem I genes, 13 Photosystem II genes, four photosynthetic electron transport genes, four F-type ATPase genes, and one Cytochrome b6/f complex gene. Notably, these genes exhibited continuous upregulated profiles from 9 to 15 DAP in JNP, such as photosystem I subunit VI *NnPSAH* (NNU_21431, NNU_26150), photosystem II oxygen-evolving enhancer protein *NnPSBO* (NNU_05490, NNU_23333), and ferredoxin *NnPETF* (NNU_05621, NNU_06707, and NNU_09587). The upregulated levels of these genes were consistent with the expression patterns of most Chl biosynthesis genes (**Figure 5** and **Supplementary Figure 12**). These results indicated a simultaneous activation of photosynthesis and Chl biosynthesis pathways during lotus plumule development.

Chl Is Synthesized by the Light-Dependent Reaction in Lotus Plumule

To investigate the relationship between light and Chl biosynthesis in lotus plumule, a light-controlled experiment was performed

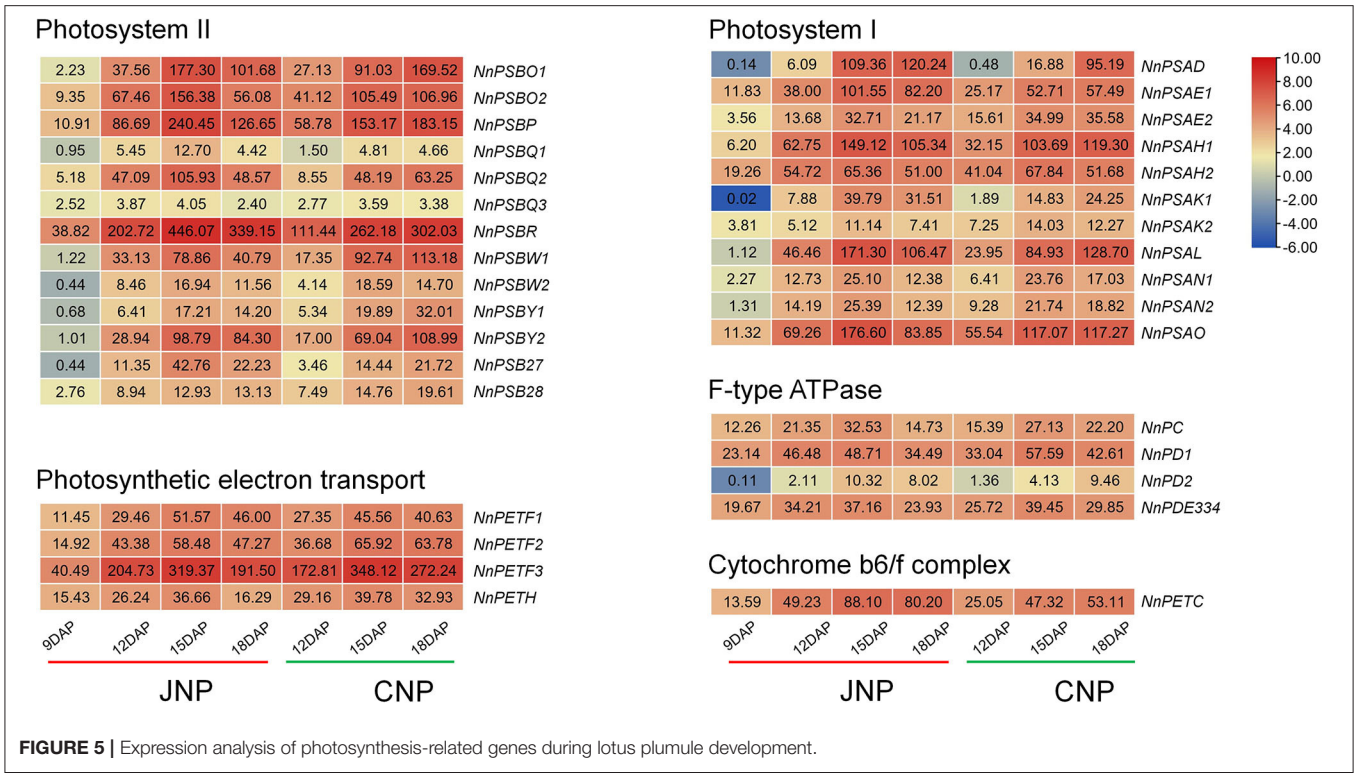


FIGURE 5 | Expression analysis of photosynthesis-related genes during lotus plumule development.

in the pods of seed-lotus cultivar "Jianxuan17". As a result, Chl biosynthesis in lotus plumule was strongly inhibited under the dark condition with plumule color turning yellowish, and Chl content decreased by 75.5%, relative to the unwrapped pods at 18 DAP (**Figures 6A,B**). In gymnosperms, algae, and photosynthetic bacteria, light-independent Pchilde reductase (DPOR) is responsible for Chl biosynthesis in dark conditions. Screening for homologous DPOR genes using the published lotus chloroplast (Wu et al., 2014) and nuclear (Ming et al., 2013) genome data revealed no hits (**Supplementary Figure 13**). In contrast, two LPOR genes were significantly upregulated from 9 to 15 DAP. For example, the FPKM expression of *NnPOR1* showed a significant increase from 351.51 at 9 DAP to 3291.89 at 15 DAP (**Figure 4F**). Overall, these results demonstrate that Chl biosynthesis in lotus plumule is light-dependent, catalyzed by the LPOR reduction of Pchlide. Notably, plumule exposure to dark conditions had no significant effect on the expression of Chl biosynthesis genes. For example, no variation in the expression of five Chl biosynthesis genes, including *NnHEMB*, *NnCHLD*, *NnPOR1*, *NnPOR2*, and *NnCAO1*, was observed in samples under light and dark treatments at 12 and 15 DAP (**Figure 6C**).

Light is a key factor affecting the biosynthesis of secondary metabolites (Coelho et al., 2007; Setiawati et al., 2018). The content of bis-BIAs in lotus plumule under dark treatment decreased by 18.4% relative to those under normal light exposure at 18 DAP, which was consistent with the decrease in the expression of BIAs pathway genes, such as *NnNCS*, *Nn6OMT*, and *NnCYP80A* (**Figures 6D,E**).

Identification of Transcription Factors Co-Expressed With bis-BIAs and Chl Biosynthesis Genes

Transcription factors (TFs) are master regulators of gene expression (Mitsis et al., 2020). A total of 510 differently expressed TFs from 50 TF families were identified in this study, with the majority being bHLH, ERF, MYB, and C2H2 TF family genes (**Figure 7A**). Varied expression patterns were observed among these TFs, for example, of the 50 bHLH TFs identified, 11 were continuously upregulated, and 16 were downregulated, whereas the remaining 23 had irregular expression patterns (**Supplementary Figure 14A**). Functional enrichment analysis of the 510 TFs showed that plant hormone signal transduction and DNA-binding transcription factor activity were the most enriched KEGG and GO terms, respectively. In addition, "response to chitin" and "cell differentiation" were the most enriched terms in biological process classification, and genes involved in these two processes showed varied expression patterns (**Supplementary Figure 14B**).

To investigate the potential functions of TFs involved in bis-BIAs and Chl biosynthesis, the correlation between their expression profiles and of the identified bis-BIAs and Chl biosynthesis structural genes was calculated (**Figure 7B** and **Supplementary Table 5**). For example, 37, 30, and 12 TFs were co-expressed with *NnTyDC*, *NnCYP80A*, and *Nn6OMT1*, while, 27, 18, and 17 TFs were co-expressed with *NnHEMB*, *NnGSA*, and *NnPOR1* ($r \geq 0.8$), respectively (**Figures 7C,D**). Interestingly, TFs, such as *NnMYB16* (NNU_07316) and *NnGLK1* (NNU_11191), showed co-expression with multiple

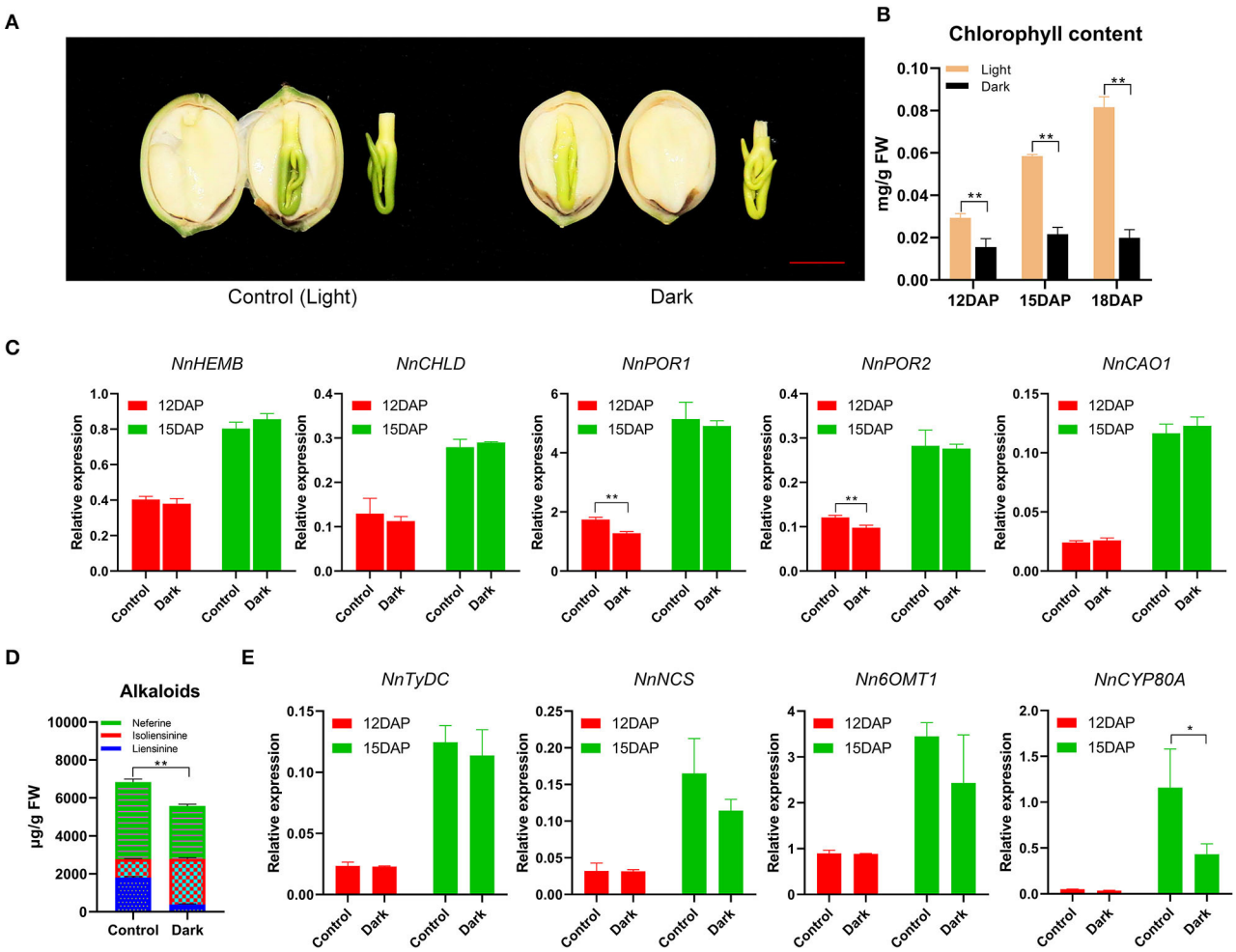


FIGURE 6 | The biosynthesis of Chl and alkaloids in lotus plumule is inhibited under dark treatment. **(A)** Phenotypes of 18 DAP lotus plumule under control (left) and dark treatments (right). The red bar represents 1 cm. **(B)** Chl content in lotus plumule under light and dark treatments. Bars represent means ± standard error ($n = 4$). $^{**}P \leq 0.01$. **(C)** Expression patterns of Chl biosynthesis genes under light and dark conditions. Bars represent means ± standard error ($n = 3$). $^{**}P \leq 0.01$. **(D)** Alkaloid content in lotus plumule under light and dark conditions. Bars represent means ± standard error ($n = 4$). $^{**}P \leq 0.01$. **(E)** Expression patterns of selected alkaloid biosynthesis genes under light and dark conditions. Bars represent means ± standard error ($n = 3$). $^{*}P \leq 0.05$.

structural genes (**Figure 7E**). MYB TFs are key regulators of plant secondary metabolite biosynthesis (Chezem et al., 2017; Kishi-Kaboshi et al., 2018). A candidate *NnMYB16* gene showed a continuous upregulated expression from 9 to 15 DAP, which was subsequently downregulated at 18 DAP in JNP. In addition, *NnMYB16* was co-expressed with nine bis-BIAs biosynthetic genes and 11 Chl biosynthetic genes. Similarly, GARP-type GLK TFs are key regulators of Chl biosynthesis, and *NnGLK1*, a homolog of *AtGLK1* (AT2G20570), has been shown to bind to the promoter of some Chl biosynthetic genes and regulate their expression (Waters et al., 2009; Sakuraba et al., 2017). In this study, a correlation was observed between the expression of *NnGLK1* and 11 Chl biosynthetic genes, including *NnGSA*, *NnCHLD*, and *NnPOR2* during lotus plumule development. In addition, *NnGLK1* also was co-expressed with eight bis-BIAs biosynthetic genes. Overall, these results provide important references for further research on the transcriptional regulation mechanisms of bis-BIAs and Chl biosynthesis in the lotus.

DISCUSSION

Recent studies on the development of lotus seeds have mainly focused on cotyledons, the main edible seed tissue (Wang et al., 2016; Li et al., 2018; Sun et al., 2020). However, despite the pharmacological significance of lotus plumule, its development process remains largely unknown. Physiological analysis in this study determined the rapid growth stage of lotus plumule to be from 9 to 15 DAP, while the onset and rapid accumulation of BIAs biosynthesis were detected at 15 DAP (**Figure 1**). The observed variation in plumule color during development occurred due to biosynthesis of Chl, which was rapidly accumulated between 12 and 18 DAP.

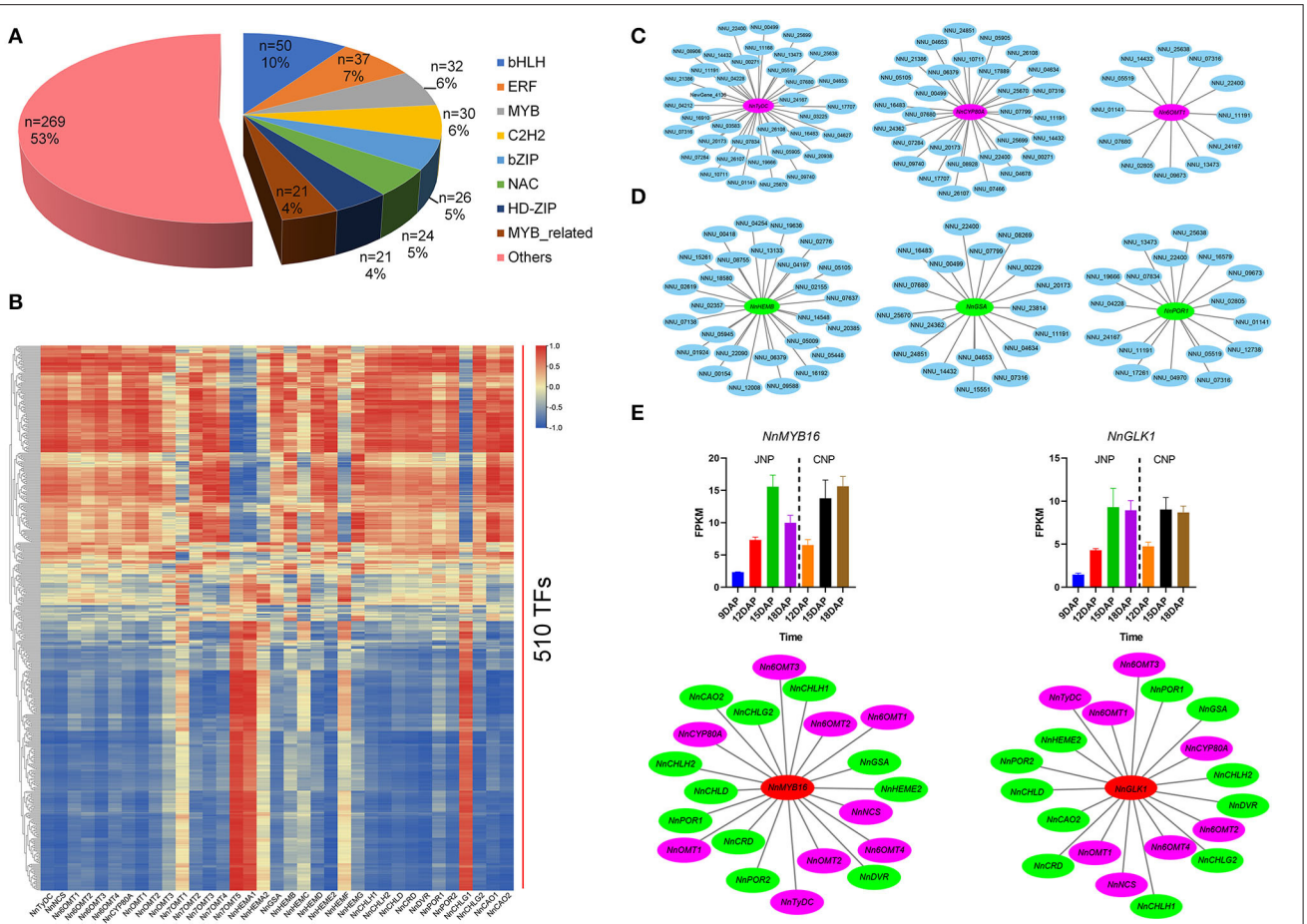


FIGURE 7 | Identification of transcription factors co-expressed with bis-BIAs and Chl biosynthesis genes. **(A)** The proportion of differentially expressed TFs. **(B)** Visual representation of the co-expression relationships between 510 TFs and bis-BIAs/ Chl biosynthesis genes. **(C)** Selected bis-BIAs, and **(D)** Chl biosynthesis genes co-expressed with TFs. **(E)** The FPKM expression of *NnMYB16* (NNU_07316) and *NnGLK1* (NNU_11191), and their co-expression with multiple bis-BIAs and Chl biosynthesis genes, respectively.

Dynamic Characteristics of BIAs Biosynthesis in Lotus

Most parts of the lotus plant have traditionally been used for various medicinal purposes due to its ability to accumulate abundant bioactive compounds, such as alkaloids and flavonoids (Mukherjee et al., 2009; Chen et al., 2012; Deng et al., 2016; Limwachiranon et al., 2018). To date, over 20 alkaloids categorized into aporphines, monobenzylisoquinolines, and bisbenzylisoquinolines have been identified in lotus (Deng et al., 2016; Yang et al., 2017). In this study, bis-BIAs, such as liensinine, isoliensinine, and neferine, were identified as the predominant alkaloids in lotus plumule, which is consistent with the results of previous studies (Deng et al., 2016; Menendez-Perdomo and Facchini, 2020). Notably, isoliensinine was not detected in CNP, which suggests the effects of genotype on alkaloid composition in lotus plumule (**Figure 1F** and **Supplementary Figure 1D**). In contrast, aporphine-type BIAs, including *N*-nornuciferine, *O*-nornuciferine, anonaine, nuciferine, and romaine, were predominantly accumulated in the lotus leaf (Chen et al., 2013; Deng et al., 2016). This observed abundant accumulation of different alkaloid types underscores the numerous pharmacological potentials of the lotus plant.

The biosynthesis pathway of aporphine-type BIAs in lotus leaf has previously been reported (Yang et al., 2017; Deng et al., 2018). However, the bis-BIAs biosynthesis pathway in plumule is yet to be characterized. Here, 16 structural genes potentially involved in bis-BIAs biosynthesis, including *NnTyDC*, *NnNCS*, *Nn6OMT,* and *NnCYP80A*, were identified in the lotus plumule (**Figure 3**). The accumulation of bis-BIAs was detected from 15 DAP, while most related biosynthesis genes were significantly upregulated from 12 DAP, and this suggests a delayed initiation of bis-BIAs biosynthesis after structural gene activation in lotus plumule (**Figures 3B,D**). Decarboxylation of tyrosine to yield *N*-methylcoclaurine is the most common pathway of alkaloids biosynthesis (Hagel and Facchini, 2013). However, our results identified genes with contrasting expression patterns in both bis-BIAs and aporphine-type BIAs biosynthesis. For example, the expression of *Nn6OMT1* (NNU_19035) was downregulated

in aporphine-type BIAs biosynthesis in leaf (Yang et al., 2017) but was significantly upregulated during bis-BIAs biosynthesis in plumule (**Figure 3D**). Similarly, *NnCYP80G*, a homolog of *NnCYP80A* and a potential structural gene in the aporphine-type BIAs biosynthesis (Deng et al., 2018), was highly expressed in lotus leaf but with extremely low expression levels in the plumule (**Figure 3E**). Taken together, these results suggest flexibility in the lotus alkaloids biosynthesis.

NnCNMT is a single gene copy in the lotus genome that encodes the CNMT enzyme, which catalyzes the conversion of (*S*)-Coclaurine to (*S*)-*N*-Methylcoclaurine. The expression of *NnCNMT* was upregulated in the lotus leaf (Yang et al., 2017) but was barely detectable in the plumule (**Figure 3B**). This result contradicts the independent production of alkaloids in lotus plumule, thus additional proteomic and enzyme activity studies on CNMT are warranted. Interestingly, previous identification of BIAs in leaf bleeding sap led to the speculation that bis-BIAs are mainly synthesized in the leaf and then transported to the plumule (Deng et al., 2016); thus, our results provide additional evidence supporting this bis-BIAs accumulation pattern in lotus plumule.

Characterization of Chl Biosynthesis in Lotus Plumule

The Chl biosynthesis is an essential cellular process for plant photosynthesis. Unlike most crops, the lotus plumule is green in color due to the presence of Chl in its seeds, which is an adaptive trait for seeds' vitality and longevity (Ji et al., 2001; Shen-Miller, 2007). All structural genes related to the Chl biosynthesis pathway and their homologs were identified in the lotus genome, with the expression levels of most genes showing a positive correlation with Chl content in the plumule (**Figure 4**). This result suggests that the Chl biosynthesis pathway is conserved in lotus. However, a comparison between lotus and *Arabidopsis* Chl biosynthesis pathway-related genes identified some independent evolutionary patterns. For example, a different number of isoforms of HEMA, GSA, POR, and CAO were observed, with *Arabidopsis* having three, two, three, and one while lotus having two, one, two, and two isoforms, respectively. *NnCHLG1* and its homolog *NnCHLG2,* encoding Chl synthase and catalyzing the last step of Chl biosynthesis, showed contrasting expression patterns during lotus plumule development (**Figure 4F**). Similarly, inconsistent expression patterns were also observed in *NnHEMA1* and *NnHEMA2* genes (**Figure 4F**). The contrasting expression patterns between the paralogs of *CHLG* and *HEMA* could suggest that the genes are undergoing functional divergence in the lotus. Furthermore, the expression levels of *NnHEMD* and *NnCHLM*, which are single-copy genes encoding uroporphyrinogen III synthase (UROS) and SAM Mg-protoporphyrin IX methyltransferase (MgPMT), respectively, were very low, and their functions in Chl synthesis need to be further determined (**Figure 4F**).

The green lotus plumule is enclosed in the middle of the seed, and thus could be assumed to undergo light-independent Chl biosynthesis (Yakovlev and Zhukova, 1980). However, previous studies have reported light-dependent Chl biosynthesis in the lotus plumule via specialized chloroplast with giant granum and photosystem structures (Zuo Bao-yu et al., 1992; Ji et al., 2001). In this study, lotus plumule incubated in dark conditions developed yellowish color with severely decreased Chl content, which further confirmed the light-dependent Chl biosynthesis reaction (**Figures 6A,B**). It is reasonable to speculate that lotus seeds utilize the thin semi-transparent integuments around the plumule during the early stages of development to sense light signals for light-dependent Chl biosynthesis reaction (Ji et al., 2001). In addition, previous anatomical studies identified three pores inside lotus seeds and showed that the tissue structures at both ends of seeds are relatively loose, which could allow light penetration (Chen and Zhang, 1988; Huang et al., 2011). Overall, these results provide potential evidence that the lotus plumule is sensitive to light stimuli at the structural level. Moreover, the absence of genes encoding DPOR in chloroplast and nuclear genome of the lotus was consistent with the previous conclusion that members of DPOR genes were completely lost in angiosperms, thus, inhibiting their ability to form Chl under light-independent reactions (Gabruk and Mysliwa-Kurdziel, 2020). Interestingly, two highly homologous genes encoding LPOR were identified in the lotus genome with significant upregulated expression levels in plumule during Chl biosynthesis (**Figures 4E,F**), further providing evidence for light-dependent Chl biosynthesis in lotus plumule.

The Potential Connections Between the Pathways Leading to bis-BIAs and Chl Biosynthesis in Lotus Plumule

The currently available literature has not been able to resolve the connection between BIAs and Chl biosynthesis in plants (Baldwin, 1988; Wei et al., 2012; Setiawati et al., 2018). The correlation between alkaloids and Chl biosynthesis varies among plant species, for example, no significant correlation was observed between purine alkaloids and Chl in green tea cultivars, while a strong correlation existed between alkaloids and Chl biosynthesis in *Ephedra procera* (Parsaeimehr et al., 2010; Wei et al., 2012). In this study, potential connections between these two pathways were detected in the lotus plumule. First, bis-BIAs and Chl showed similar biosynthesis and accumulation patterns in lotus plumule, with both showing continuous accumulation from 15 DAP to 21 DAP, despite a delayed initiation of bis-BIAs biosynthesis relative to Chl biosynthesis (**Figures 1E,F**). Second, some structural genes of the bis-BIAs and Chl biosynthesis pathways showed similar expression patterns during lotus plumule development. Correlating in the expression of structural genes related to these two synthetic pathways revealed 13 highly co-expressed associations ($r \geq 0.8$) between bis-BIAs biosynthesis genes with at least one Chl biosynthesis gene (**Supplementary Figure 15**). In addition, some co-expressed TFs with bis-BIAs and Chl biosynthesis genes, such as *NnMYB16* and *NnGLK1,* were identified (**Figure 7E**). As a key Chl biosynthesis regulator, *NnGLK1* also showed co-expressed with some bis-BIAs biosynthesis genes, including *NnNCS*, *Nn6OMTs*, and *NnCYP80A* (**Figure 7E**). We therefore speculated that a common

transcriptional regulatory mechanism might exist between these two pathways. Third, light is a co-regulator of bis-BIAs and Chl biosynthesis in lotus plumule, and our results showed that bis-BIAs and Chl biosynthesis in lotus plumule were strongly inhibited under dark conditions, with the content of bis-BIAs and Chl decreasing by 18.4 and 75.5% at 18 DAP, respectively, relative under normal light exposure (**Figures 6B,D**). The light-induced co-regulation of alkaloids and Chl biosynthesis has previously been reported (Zhao et al., 2001; Zhu et al., 2015; Yu et al., 2018; Li et al., 2021). For example, the light improved alkaloids and Chl biosynthesis in the *Catharanthus roseus* callus and enhanced the content of vindoline and Chl in illuminated callus by ∼ 3–4 folds and 10–20 folds, respectively (Zhao et al., 2001).

This study provides unprecedented information and resources which could potentially be applied in lotus seed preservation and bis-BIAs detection in the lotus plumule. For example, the flavor quality of fresh lotus seeds deteriorates rapidly from 15 DAP due to increased alkaloid accumulation, leading to bitterness (Tu et al., 2020; Sun et al., 2021). Decreased alkaloid accumulation in lotus plumule under dark treatment was observed in this study. Thus, storing harvested seedpods in the dark could be a practical way to extend the shelf-life of fresh lotus seeds during postharvest storage. In addition, the correlation between bis-BIAs and Chl contents observed in our study suggests that determining Chl content alone could adequately be used as a potentially cost-effective indicator for predicting bis-BIAs content, which is usually expensive and labor-intensive.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA747903.

AUTHOR CONTRIBUTIONS

MY, YL, and HSu contributed to the conception and design of the study. HSu, HSo, DY, MZ, and YW performed the experiments and data analysis. HSu wrote the manuscript. MY, YL, HSu, XD, JL, JX, and LC revised the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the Biological Resources Program CAS (Grant No. KFJ-BRP-007-009), the National Natural Science Foundation of China (Grant No. 31872136), and the Hubei Provincial Natural Science Foundation of China (Grant No. 2020CFB484).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.885503/full#supplementary-material

REFERENCES

Baldwin, I. T. (1988). Short-term damage-induced increases in tobacco alkaloids protect plants. *Oecologia* 75, 367–370. doi: 10.1007/BF00376939

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 49, W317–W325. doi: 10.1093/nar/gkab447

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, S., Fang, L., Xi, H., Guan, L., Fang, J., Liu, Y., et al. (2012). Simultaneous qualitative assessment and quantitative analysis of flavonoids in various tissues of lotus (*Nelumbo nucifera*) using high performance liquid chromatography coupled with triple quad mass spectrometry. *Anal. Chim. Acta*. 724, 127–135. doi: 10.1016/j.aca.2012.02.051

Chen, S., Guo, W., Qi, X., Zhou, J., Liu, Z., and Cheng, Y. (2019). Natural alkaloids from lotus plumule ameliorate lipopolysaccharide-induced depression-like behavior: integrating network pharmacology and molecular mechanism evaluation. *Food Funct*. 10, 6062–6073. doi: 10.1039/C9FO01092K

Chen, S., Zhang, H., Liu, Y., Fang, J., and Li, S. (2013). Determination of lotus leaf alkaloids by solid phase extraction combined with high performance liquid chromatography with diode array and tandem mass spectrometry detection. *Anal. Lett*. 46, 2846–2859. doi: 10.1080/00032719.2013.816960

Chen, W., and Zhang, S. (1988). A study on ecological anatomy in *Nelumbo Nucifera* Gaertn. *Acta. Ecol. Sin*. 8, 277–282.

Chezem, W. R., Memon, A., Li, F. S., Weng, J. K., and Clay, N. K. (2017). SG2-type R2R3-MYB transcription factor MYB15 controls defense-induced lignification and basal immunity in *Arabidopsis*. *Plant Cell* 29, 1907–1926. doi: 10.1105/tpc.16.00954

Coelho, G. C., Rachwal, M. F. G., Dedecek, R. A., Curcio, G. R., Nietsche, K., and Schenkel, E. P. (2007). Effect of light intensity on methylxanthine contents of *Ilex paraguariensis* A. St. Hil. *Biochem. Syst. Ecol.* 35, 75–80. doi: 10.1016/j.bse.2006.09.001

Cui, L., Huang, F., Zhang, D., Lin, Y., Liao, P., Zong, J., et al. (2015). Transcriptome exploration for further understanding of the tropane alkaloids biosynthesis in *Anisodus acutangulus*. *Mol. Genet. Genomics* 290, 1367–1377. doi: 10.1007/s00438-015-1005-y

Deng, X., Zhao, L., Fang, T., Xiong, Y., Ogutu, C., Yang, D., et al. (2018). Investigation of benzylisoquinoline alkaloid biosynthetic pathway and its transcriptional regulation in lotus. *Hortic. Res.* 5, 29. doi: 10.1038/s41438-018-0035-0

Deng, X., Zhu, L., Fang, T., Vimolmangkang, S., Yang, D., Ogutu, C., et al. (2016). Analysis of isoquinoline alkaloid composition and wound-induced variation in *Nelumbo* using HPLC-MS/MS. *J. Agric. Food. Chem* 64, 1130–1136. doi: 10.1021/acs.jafc.5b06099

Fracasso, A., Trindade, L. M., and Amaducci, S. (2016). Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biol.* 16, 115. doi: 10.1186/s12870-016-0800-x

Gabruk, M., and Mysliwa-Kurdziel, B. (2020). The origin, evolution and diversification of multiple isoforms of light-dependent protochlorophyllide oxidoreductase (LPOR): focus on angiosperms. *Biochem. J.* 477, 2221–2236. doi: 10.1042/BCJ20200323

Goyal, E., Amit, S. K., Singh, R. S., Mahato, A. K., Chand, S., and Kanika, K. (2016). Transcriptome profiling of the salt-stress response in *Triticum aestivum* cv. *Kharchia Local. Sci. Rep*. 6, 27752. doi: 10.1038/srep27752

Guo, X., Li, Y., Li, C., Luo, H., Wang, L., Qian, J., et al. (2013). Analysis of the Dendrobium officinale transcriptome reveals putative alkaloid biosynthetic genes and genetic markers. *Gene* 527, 131–138. doi: 10.1016/j.gene.2013.05.073

Hagel, J. M., and Facchini, P. J. (2013). Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. *Plant Cell Physiol.* 54, 647–672. doi: 10.1093/pcp/pct020

He, S. M., Song, W. L., Cong, K., Wang, X., Dong, Y., Cai, J., et al. (2017). Identification of candidate genes involved in isoquinoline alkaloids biosynthesis in *Dactylicapnos scandens* by transcriptome analysis. *Sci. Rep*. 7, 9119. doi: 10.1038/s41598-017-08672-w

Huang, T., Li, C., Xiao, L. (2011). Discussion on the greening of the Lotus (*Nelumbo Nucifera* Gaertn.) seed embryo bud. *Chin. Hortic. Abstr*. 27, 16–17.

Ji, H. W., Li, L. B., and Kuang, T. Y. (2001). The chlorophyll biosynthesis in lotus embryo is light-dependent. *Acta. Bot. Sin*. 43, 693–698.

Kishi-Kaboshi, M., Seo, S., Takahashi, A., and Hirochika, H. (2018). The MAMP-responsive MYB transcription factors MYB30, MYB55 and MYB110 activate the HCAA synthesis pathway and enhance immunity in rice. *Plant Cell Physiol*. 59, 903–915. doi: 10.1093/pcp/pcy062

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol*. 33, 1870–1874. doi: 10.1093/molbev/msw054

Lanver, D., Muller, A. N., Happel, P., Schweizer, G., Haas, F. B., Franitza, M., et al. (2018). The biotrophic development of ustilago maydis studied by RNA-Seq analysis. *Plant Cell* 30, 300–323. doi: 10.1105/tpc.17.00764

Li, J., Shi, T., Huang, L., He, D., Nyong'A., Yang, P. F., et al. (2018). Systematic transcriptomic analysis provides insights into lotus (*Nelumbo nucifera*) seed development. *Plant Growth Regul*. 86, 339–350. doi: 10.1007/s10725-018-0433-1

Li, Q., Xu, J., Yang, L., Sun, Y., Zhou, X., Zheng, Y., et al. (2021). LED light quality affect growth, alkaloids contents, and expressions of amaryllidaceae alkaloids biosynthetic pathway genes in lycoris longituba. *J. Plant Growth Regul*. 41, 257–270. doi: 10.1007/s00344-021-10298-2

Limwachiranon, J., Huang, H., Shi, Z., Li, L., and Luo, Z. (2018). Lotus flavonoids and phenolic acids: health promotion and safe consumption dosages. *Compr. Rev. Food Sci. Food Saf*. 17, 458–471. doi: 10.1111/1541-4337.12333

Liu, B., Li, J., Yi, R., Mu, J., Zhou, X., and Zhao, X. (2019). Preventive effect of alkaloids from lotus plumule on acute liver injury in mice. *Foods* 8:36. doi: 10.3390/foods8010036

Liu, T., Zhu, M., Zhang, C., and Guo, M. (2017). Quantitative analysis and comparison of flavonoids in lotus plumules of four representative lotus cultivars. *J. Spectroscopy* 2017, 1–9. doi: 10.1155/2017/7124354

Menendez-Perdomo, I. M., and Facchini, P. J. (2020). Isolation and characterization of two O-methyltransferases involved in benzylisoquinoline alkaloid biosynthesis in sacred lotus (*Nelumbo nucifera*). *J. Biol. Chem* 295, 1598–1612. doi: 10.1074/jbc.RA119.011547

Ming, R., Vanburen, R., Liu, Y., Mei, Y., Han, Y., Li, L. T., et al. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol*. 14:R41. doi: 10.1186/gb-2013-14-5-r41

Mitsis, T., Efthimiadou, A., Bacopoulou, F., Vlachakis, D., Chrousos, G., and Eliopoulos, E. (2020). Transcription factors and evolution: an integral part of gene expression (Review). *World Acad. Sci. J*. 2, 3–8. doi: 10.3892/wasj.2020.32

Morley, P. J., Jump, A. S., West, M. D., and Donoghue, D. N. M. (2020). Spectral response of chlorophyll content during leaf senescence in European beech trees. *Environ. Res. Commun*. 2:071002. doi: 10.1088/2515-7620/aba7a0

Morris, J. S., and Facchini, P. J. (2019). Molecular origins of functional diversity in benzylisoquinoline alkaloid methyltransferases. *Front. Plant Sci*. 10, 1058. doi: 10.3389/fpls.2019.01058

Mukherjee, P. K., Mukherjee, D., Maji, A. K., Rai, S., and Heinrich, M. (2009). The sacred lotus (*Nelumbo nucifera*) - phytochemical and therapeutic profile. *J. Pharm. Pharmacol*. 61, 407–422. doi: 10.1211/jpp/61.04.0001

Parsaeimehr, A., Sargsyan, E., and Javidnia, K. (2010). Influence of plant growth regulators on callus induction, growth, chlorophyll, ephedrine and pseudoephedrine contents in Ephedra procera. *J. Med. Plants Res*. 4, 1308–1317. doi: 10.5897/JMPR10.202

Reinbothe, C., El Bakkouri, M., Buhr, F., Muraki, N., Nomata, J., Kurisu, G., et al. (2010). Chlorophyll biosynthesis: spotlight on protochlorophyllide reduction. *Trends Plant Sci*. 15, 614–624. doi: 10.1016/j.tplants.2010.07.002

Sakuraba, Y., Kim, E. Y., Han, S. H., Piao, W., An, G., Todaka, D., et al. (2017). Rice phytochrome-interacting factor-like1 (OsPIL1) is involved in the promotion of chlorophyll biosynthesis through feed-forward regulatory loops. *J. Exp. Bot*. 68, 4103–4114. doi: 10.1093/jxb/erx231

Setiawati, T., Ayalla, A., Nurzaman, M., and Mutaqin, A. Z. (2018). Influence of light intensity on leaf photosynthetic traits and alkaloid content of kiasahan (*Tetracera scandens* L.). *IOP Conf. Ser. Earth Environ. Sci*. 166, 012025. doi: 10.1088/1755-1315/166/1/012025

Shen-Miller, J. (2007). Sacred lotus, the long-living fruits of China Antique. *Seed Sci. Res*. 12, 131–143. doi: 10.1079/SSR2002112

Shi, T., Rahmani, R. S., Gugger, P. F., Wang, M., Li, H., Zhang, Y., et al. (2020). Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plant. *Mol. Biol. Evol*. 37, 2394–2413. doi: 10.1093/molbev/msaa105

Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207–208. doi: 10.1093/bioinformatics/18.1.207

Sun, H., Li, J., Song, H., Yang, D., Deng, X., Liu, J., et al. (2020). Comprehensive analysis of AGPase genes uncovers their potential roles in starch biosynthesis in lotus seed. *BMC Plant Biol*. 20:457. doi: 10.1186/s12870-020-02666-z

Sun, H., Liu, Y., Ma, J., Wang, Y., Song, H., Li, J., et al. (2021). Transcriptome analysis provides strategies for postharvest lotus seeds preservation. *Postharvest. Biol. Tec*. 179:111583. doi: 10.1016/j.postharvbio.2021.111583

Tripathy, B. C., and Pattanayak, G. K. (2012). Chlorophyll biosynthesis in higher plants. *Photosynthesis* 34, 63–94. doi: 10.1007/978-94-007-1579-0_3

Tu, Y., Yan, S., and Li, J. (2020). Impact of harvesting time on the chemical composition and quality of fresh lotus seeds. *Hortic. Environ. Biotechnol*. 61, 735–744. doi: 10.1007/s13580-020-00233-x

Wang, L., Fu, J., Li, M., Fragner, L., Weckwerth, W., and Yang, P. (2016). Metabolomic and proteomic profiles reveal the dynamics of primary metabolism during seed development of lotus (*Nelumbo nucifera*). *Front. Plant Sci*. 7, 750. doi: 10.3389/fpls.2016.00750

Wang, Y., Fan, G., Liu, Y., Sun, F., Shi, C., Liu, X., et al. (2013). The sacred lotus genome provides insights into the evolution of flowering plants. *Plant J*. 76, 557–567. doi: 10.1111/tpj.12313

Waters, M. T., Wang, P., Korkaric, M., Capper, R. G., Saunders, N. J., and Langdale, J. A. (2009). GLK transcription factors coordinate expression of the photosynthetic apparatus in *Arabidopsis*. *Plant Cell* 21, 1109–1128. doi: 10.1105/tpc.108.065250

Wei, K., Wang, L.-Y., Zhou, J., He, W., Zeng, J.-M., Jiang, Y.-W., et al. (2012). Comparison of catechins and purine alkaloids in albino and normal green tea cultivars (*Camellia sinensis* L.) by HPLC. *Food Chem*. 130, 720–724. doi: 10.1016/j.foodchem.2011.07.092

Wu, Z., Gui, S., Quan, Z., Pan, L., Wang, S., Ke, W., et al. (2014). A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol*. 14, 289. doi: 10.1186/s12870-014-0289-0

Xia, H., Zhu, L., Zhao, C., Li, K., Shang, C., Hou, L., et al. (2020). Comparative transcriptome analysis of anthocyanin synthesis in black and pink peanut. *Plant Signal Behav*. 15, 1721044. doi: 10.1080/15592324.2020.1721044

Yakovlev, M. S., and Zhukova, G. Y. (1980). Chlorophyll in embryos of angiosperm seeds, a review. *Botaniska Notiser*. 133, 323–336.

Yamamoto, H., Kusumi, J., Yamakawa, H., and Fujita, Y. (2017). The effect of two amino acid residue substitutions via RNA editing on dark-operative protochlorophyllide oxidoreductase in the black pine chloroplasts. *Sci. Rep*. 7, 2377. doi: 10.1038/s41598-017-02630-2

Yang, M., Zhu, L., Li, L., Li, J., Xu, L., Feng, J., et al. (2017). Digital gene expression analysis provides insight into the transcript profile of the genes involved in aporphine alkaloid biosynthesis in lotus (*Nelumbo nucifera*). *Front. Plant Sci*. 8, 80. doi: 10.3389/fpls.2017.00080

Yang, M., Zhu, L., Pan, C., Xu, L., Liu, Y., Ke, W., et al. (2015). Transcriptomic analysis of the regulation of rhizome formation in temperate and tropical lotus (*Nelumbo nucifera*). *Sci. Rep*. 5, 13059. doi: 10.1038/srep13059

Yu, B., Liu, Y., Pan, Y., Liu, J., Wang, H., and Tang, Z. (2018). Light enhanced the biosynthesis of terpenoid indole alkaloids to meet the opening of cotyledons in process of photomorphogenesis of Catharanthus roseus. *Plant Growth Regul*. 84, 617–626. doi: 10.1007/s10725-017-0366-0

Zhao, J., Zhu, W. H., and Wh, Q. (2001). Effects of light and plant growth regulators on the biosynthesis of vindoline and other indole alkaloids in Catharanthus roseus callus cultures. *Plant Growth Regul*. 33, 43–49. doi: 10.1023/A:1010722925013

Zhu, J., Wang, M., Wen, W., and Yu, R. (2015). Biosynthesis and regulation of terpenoid indole alkaloids in Catharanthus roseus. *Pharmacogn rev* 9, 24–28. doi: 10.4103/0973-7847.156323

Ziegler, J., and Facchini, P. J. (2008). Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* 59, 735–769. doi: 10.1146/annurev.arplant.59.032607.092730

Zuo Bao-yu, L. G., Tang, C. Q., Jiang, G., and Ting-yun, K. (1992). Changes of thylakoid membrane stacks and Chl a/b ratio of chloroplast from sacred lotus (*Nelumbo nucifera*) seeds during their germination under light. *J. Integr. Plant Biol.* 34, 645–650.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 01 July 2022
doi: 10.3389/fpls.2022.928760

Investigation of Enzymes in the Phthalide Biosynthetic Pathway in *Angelica sinensis* Using Integrative Metabolite Profiles and Transcriptome Analysis

Wei-Meng Feng, Pei Liu, Hui Yan*, Guang Yu, Sen Zhang, Shu Jiang, Er-Xin Shang, Da-Wei Qian and Jin-Ao Duan**

Jiangsu Key Laboratory for High Technology Research of TCM Formulae, Jiangsu Collaborative Innovation Center of Chinese Medicinal Resources Industrialization, National and Local Collaborative Engineering Center of Chinese Medicinal Resources Industrialization and Formulae Innovative Medicine, Nanjing University of Chinese Medicine, Nanjing, China

OPEN ACCESS

Edited by:
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Jaime Barros-Rios, University of Missouri, United States
Xiaohui Yan, Tianjin University of Traditional Chinese Medicine, China

***Correspondence:**
Pei Liu
liupei@njucm.edu.cn
Hui Yan
glory-yan@163.com
Jin-Ao Duan
dja@njucm.edu.cn

Specialty section:
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

Received: *26 April 2022*
Accepted: *13 June 2022*
Published: *01 July 2022*

Citation:
Feng W-M, Liu P, Yan H, Yu G, Zhang S, Jiang S, Shang E-X, Qian D-W and Duan J-A (2022) Investigation of Enzymes in the Phthalide Biosynthetic Pathway in Angelica sinensis Using Integrative Metabolite Profiles and Transcriptome Analysis. Front. Plant Sci. 13:928760. doi: 10.3389/fpls.2022.928760

The roots of *Angelica sinensis* (Oliv.) Diels are well known for their efficacy in promoting blood circulation. Although many studies have indicated that phthalides are the main chemical components responsible for the pharmacological properties of *A. sinensis,* the phthalide biosynthetic pathway and enzymes that transform different phthalides are still poorly understood. We identified 108 potential candidate isoforms for phthalide accumulation using transcriptome and metabolite profile analyses. Then, six enzymes, including phospho-2-dehydro-3-deoxyheptonate aldolase 2, shikimate dehydrogenase, primary amine oxidase, polyphenol oxidase, tyrosine decarboxylase, and shikimate O-hydroxycinnamoyl transferase, were identified and proven to be involved in phthalide accumulation by heterologously expressing these proteins in *Escherichia coli*. We proposed a possible mechanism underlying phthalide transformation and biosynthetic pathways in *A. sinensis* based on our findings. The results of our study can provide valuable information for understanding the mechanisms underlying phthalide accumulation and transformation and enable further development of quality control during the cultivation of *A. sinensis*.

Keywords: *Angelica sinensis*, phthalides biosynthetic pathway, transcriptome, regulation mechanism, prokaryotic expression

INTRODUCTION

The *Angelica sinensis* (Oliv.) Diels is a high-altitude plant found in the marginal region of the Qinghai-Tibet Plateau. The roots of *A. sinensis* have a long history of being widely used in Traditional Chinese Medicine for treating various gynecological conditions (Fang et al., 2012). As one of the most frequently used Chinese medicinal materials in clinical practice, the germplasm, cultivation, harvesting, medicinal components, and pharmacological activities of *A. sinensis* have received extensive attention and continuous research. The effectiveness and therapeutic mechanisms of the *A. sinensis* are being explored and investigated. Over 180 phytochemicals have been identified in *A. sinensis* grown under high-altitude conditions, including phthalides, phenylpropanoids,

terpenoids, alkynes, and alkaloids (Zou et al., 2018). Among these, phthalides were selected as marker compounds for quality control and pharmacokinetic studies of *A. sinensis* (Wei and Huang, 2015). The extensive cultivation of the plant in different areas across the country revealed the problem of early flowering in *A. sinensis*, which causes root burn and reduction of plant oils, seriously affecting its quality and yield. Previous studies (Li et al., 2020c) mainly focused on the influencing factors and compound content analysis of the early flowering *A. sinensis*. However, the molecular mechanism underlying the change in phthalides during the early flowering of *A. sinensis* is still unclear. Therefore, identifying the key enzymes that affect the synthesis and accumulation of phthalides in the early flowering process provides important guidance in cultivating *A. sinensis*.

Phthalides are among the most important active ingredients in volatile plant oils and a characteristic component of important natural compounds from Umbelliferae plants, such as *A. sinensis* and *Ligusticum chuanxiong* (Tang et al., 2021). Recent research has indicated that phthalides are the main chemical components related to the bioactivity and pharmacological properties of *A. sinensis,* such as anti-asthma, anti-convulsant, inhibition of platelet aggregation, and enhancement of blood flow (Yi et al., 2009). In particular, n-Butylphthalide was approved by the State Food and Drug Administration of China in 2005 as a modern drug for treating ischemic strokes (Zou et al., 2018). Moreover, butylphthalide and ligustilide that show insecticidal activity against the B- and Q-biotype females of *Bemisia tabaci* (Chae et al., 2011) and *Drosophila melanogaster* (Miyazawa et al., 2004) are potential alternatives to conventional arthropod control products. This has received considerable attention from the public because they are relatively safe and poses fewer risks to the environment (Isman, 2006).

Phthalides have been recognized for their broad-spectrum biological activities. As important index components in *A. sinensis*, it is important to determine the factors that regulate phthalide accumulation (Karmakar et al., 2014). First, the expression of enzymes in the phthalide biosynthetic pathway is considered one of the key factors. Elucidating the biosynthesis of phthalides began with the structural determination of mycophenolic acid, a phthalide fragment derived from the polyketide pathway (Birch et al., 1958). Thereafter, researchers identified the biogenetic origin of butylphthalide by conducting feeding experiments to explain the formation of ligustilide in *Levisticum officinale* and determined that the alkylphthalide has polyketide precursors (Mitsuhashi and Nomura, 1966). Although it has been explored, phthalide biosynthesis, especially the interconversion mechanism between different phthalides, remains unclear and needs further investigation.

The compound content of medicinal plants varies at different developmental stages. The normal growth cycle of *A. sinensis* is 3 years, with seedlings raised in the first year, drug-forming in second year, and bolting and flowering in the third year. However, 20–30% of plants bolt and flower in the second year (Yu et al., 2019). The early bolting and flowering have a significant effect on the accumulation of secondary metabolites of *A. sinensis*, especially the reduction of volatile oil components mainly composed of phthalides. The early bolting and flowering significantly reduce the yield and quality of the roots, which seriously affects the medicinal material available on the market and the economic benefits by farmers of medical crops (Li et al., 2021). There are various reasons for early flowering, such as seedling size, environmental temperature, hormones, and microorganisms, all of which may cause early bolting in *A. sinensis*, and hence, the problem of early flowering cannot be solved by fixing a single factor (Li et al., 2020c). Thus, it may be a feasible strategy to increase the phthalide content during the flowering of *A. sinensis* to reduce waste, improve the market supply of medicinal raw materials, and alleviate the economic losses of pharmaceutical farmers. Although early flowering plants cannot be directly introduced into the market as medicinal materials, plants with high medicinal ingredient contents may become new sources of raw active ingredients.

In this study, the ultra-high performance liquid chromatography–tandem mass spectrometry (UHPLC–MS/MS) method was utilized to identify six phthalides, including ligustilide, butylphthalide, butylidenephalide, senkyunolide H, senkyunolide I and senkyunolide A, in the roots of normal flowering and early flowering *A. sinensis* plants. The absolute levels of the six phthalides and the changes in their proportions were analyzed. Illumina MiSeq high-throughput sequencing technology was used to investigate the root transcriptome. We aimed to explore candidate enzymes that positively correlated with phthalide accumulation, followed by an analysis of the content of the six phthalides and their transcriptional expression. The function of the candidate isoforms and the potential phthalide biosynthetic pathways were further determined using quantitative real-time polymerase chain reaction (qRT-PCR) and prokaryotic expression. The results from our study expand the understanding of the changes in the phthalide content between early flowering and normal flowering in *A. sinensis*. This provides insights for developing a new plant variety with a high phthalide level or a characteristic phthalide content.

MATERIALS AND METHODS

Plant Materials, Chemicals, and Reagents

In this study, root samples from the normal flowering (ZC-1 to ZC-6) and early flowering (ZT-1 to ZT-6) fresh *A. sinensis* plants of the plant strain "Mingui No. 1," were selected as the experimental materials. Samples were collected on August 22, 2018, from Tanchang, Gansu Province (104.14780 E, 34.12113 N, Height- 2,260 m). Dr. Hui Yan from the Nanjing University of Chinese Medicine authenticated the roots. Fresh *A. sinensis* roots were collected, flash-frozen in liquid nitrogen, and transported in dry ice. The samples were stored at −80°C at the Jiangsu Collaborative Innovation Center of Chinese Medicinal Resources Industrialization. The high-throughput sequencing

Abbreviations: UHPLC-MS/MS, ultra-high performance liquid chromatography-tandem mass spectrometry; qRT-PCR, quantitative real-time polymerase chain reaction; MeJA, methyl jasmonate; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MRM, multiple reaction monitoring; FPKM, fragments per kilobase of exon per million fragments mapped; COG, Clusters of Orthologous Genes; KOG, EuKaryotic Orthologous Groups.

analysis was performed at Entrusted Frasergen Bioinformatics Co., Ltd. (Wuhan, China).

Tissue culture seedlings of *A. sinensis* "Mingui No. 1" were grown in a culture room at 23 ± 1°C under a 14-h photoperiod and 2000 lx. Methyl jasmonate (MeJA; 100 μM; Ho et al., 2020; Lv et al., 2021; Zhou et al., 2021) was added to each flask prior to the solidification of the medium. When cultivation under static conditions for 30 to 40 days and the cotyledons were flattened and the true leaf sprouted, tissue culture seedlings were transferred to MeJA-containing medium for further culture. To verify the candidate isoforms, tissue culture seedlings of *A. sinensis* were collected at the following time points after treatment with MeJA: 0, 24, 48, 72, and 96 h.

Reference Standards, Including Ligustilide, Butylphthalide, Butylidenephalide, Senkyunolide H, Senkyunolide I, and Senkyunolide A (**Supplementary Figure S1**), all 98% Purified, Were Obtained From Liangwei Biochemical Reagent Ltd. (Nanjing, China).

Transcriptome Sequencing, Assembly, and Analysis

Two randomly selected plant roots were combined for transcriptome analysis, and each group (ZT and ZC groups) contained three samples for analysis. Libraries were constructed from root mRNA and sequenced using the PacBio Sequel and Illumina HiSeq X Ten PE150 platforms (Illumina, San Diego, CA, United States) by the Frasergen Biotechnology Company (Wuhan, China). Sequencing libraries were generated using the NEB Next® Ultra TM RNA Library Prep Kit (NEB, Ipswich, MA, United States) based on Illumina® manufacturer's protocols, and index codes were added to attribute sequences to each sample.

Due to the limited genomic information of *A. sinensis*, the full-length transcript sets from the roots of normal and early flowering *A. sinensis* plants by PacBio SMRT three-generation high-throughput sequencing technology were used as the reference isoforms for both subsequent bioinformatics analysis and comparative transcriptomics analysis. Sequencing reads were aligned to the reference isoforms using Tophat2 (v2.1.1) and Bowtie2 (v2.2.2) with default parameters (Li et al., 2020b). The expression of genes and isoforms was quantified using the RSEM software package (RNASeq by Expectation–Maximization v1.3.0). Gene expression differentiation was screened using the following criteria: fold change ≥2 and false discovery rate < 0.05. The differentially expressed genes were subjected to enrichment analysis of Gene Ontology (GO) functions and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, with $p \leq 0.01$ and false discovery rate ≤ 0.05 as the thresholds for both analyses by KOBAS (v3.0; Li et al., 2020a). All transcripts were annotated against the Non-Redundant (NR), GO, and KEGG databases using Diamond (v0.8.33). qRT-PCR analysis was performed to validate the transcriptome data. Correlation analysis was performed by selecting transcripts that were differentially expressed between ZT and ZC groups and were consistent with the trend of phthalide content. The datasets generated during the current study were deposited and are available at the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA749925 (PRJNA749925).[1]

Quantification of Phthalides in Root Samples and Meja-Treated Samples

Standard and sample solutions were prepared using an established method, as described previously (Feng et al., 2021). Chromatographic analyses were performed using a Waters Acquity UPLC system (Waters Corp., MA, United States), whereas mass spectromet was conducted using an AB SCIEX Triple Quad 6,500 plus (AB SCIEX Corp., Framingham, MA, United States) with electrospray ionization. The dwell time was automatically set using the MultiQuant software. Raw data were processed using MultiQuant v3.0.2 (AB SCIEX Corp.). A detailed description of the standard solution, chromatographic conditions, and methodology validation were presented in our previous study (Feng et al., 2021).

RNA Extraction and qRT-PCR

For qRT-PCR analysis, the total RNA of the tissue culture seedlings of *A. sinensis* was extracted at each time point (0, 24, 48, 72, and 96 h) after MeJA treatment. The RNA prep pure plant kit (polysaccharides polyphenolics-rich; Tiangen, Beijing, China) was used for RNA extraction with on-column DNA digestion according to the manufacturer's protocol. Total RNA (1.5 μg) was reverse transcribed using random primers and conditions described in the EasyScript All in-one First-Strand cDNA Synthesis SuperMix for qPCR (One-Step gDNA Removal; Trans gene, Beijing, China). RNA and cDNA concentrations and purities were estimated using a DS-11 spectrophotometer (DeNovix, Wilmington, DE, United States). qRT-PCR was performed on an ABI 7500 real-time PCR system (Applied Biosystems, Waltham, MA, United States; Plasencia et al., 2016). Relative gene expression was estimated using the housekeeping gene 18S rRNA as a reference according to the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). Following the cycling stage, product melting curves were generated to ensure the specificity of product formation. All procedures were performed according to the manufacturer's instructions. Primers used for qPCR are listed in **Supplementary Table S1**. Two technical replicates were used for each sample, and three samples were analyzed for each group.

Prokaryotic Expression Function Verification of Key Candidate Isoforms

The full-length cDNA of the seven candidate isoforms was cloned into the pET-28a (+) vector. The recombinant plasmids were transformed into *Escherichia coli* BL21 (DE3) cells to express recombinant proteins. The positive clones were incubated in Luria-Bertani medium in the presence of kanamycin. *E. coli* BL21 (DE3) were grown in 250 ml Erlenmeyer baffle flasks containing 100 ml of Luria-Bertani medium in a rotary shaker at 160 rpm and 37°C. When the optical density of the cultures at 600 nm reached 0.5–0.6, recombinant proteins were expressed

[1]https://www.ncbi.nlm.nih.gov/bioproject/

in *E. coli* cells following induction by the addition of 0.5 mM isopropyl-β-D-thiogalactoside and incubation overnight at 16°C. After centrifugation at 32000 × *g* for 10 min, the cells were resuspended in 1 ml phosphate buffer saline solution (pH 7.2–7.4). The cells were lysed using an ultrasonic disrupter at 15% power for 3 min, centrifuged at 12000 × *g* for 10 min, and then purified using Ni-NTA affinity chromatography under nature conditions, following the manufacturer's instructions (Cytiva, Seattle, WA, United States). After purification with Ni-NTA chromatography, the protein sample was analyzed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (Zhao et al., 2020). The protein thus obtained was used for the subsequent enzymatic reaction.

Centrifugal tubes (1.5 ml) containing 250 μl reaction mixtures, which included 100 μl of crude enzyme solution, 110 μl of 25 mM Tris–HCl (pH 7.5), 10 μl of 100 mM magnesium chloride hexahydrate ($6H_2O$ $MgCl_2$), 10 μl of 50 mM DL-dithiothreitol, and 20 μl of various phthalide (5 mg/ml)-methanol extracts, were incubated for 1 h at 30°C in the dark. Thereafter, the mixtures were centrifuged at 12000 × *g* for 10 min prior to UHPLC–MS/MS analysis (Xiong et al., 2020).

Statistical Analysis

A Student's *t*-test for the phthalide content and relative expression analysis was performed using SPSS v21.0. The results, presented as the mean ± standard deviation, were processed and optimized using GraphPad Prism 7.0.

RESULTS

Targeted Metabolite Profile Analysis of Phthalide Contents in Roots of Normal and Early Flowering *Angelica sinensis*

UHPLC–MS/MS was used to simultaneously determine six phthalide markers in 12 samples of *A. sinensis* (**Supplementary Table S2**). The most selective and specific transition was chosen for multiple reaction monitoring (MRM) determination, and all the MRM parameters are provided in **Supplementary Table S3**. The UHPLC method was validated by assessing the linearity, precision, stability, limit of detection, limit of quantification, and recovery (Feng et al., 2021).

Principal component analysis was performed to determine variations in the metabolites. Overall, the metabolite profiles of the two groups of samples differed substantially. The first principal component accounted for 68.4% of the metabolic variance between the ZT and ZC groups (**Supplementary Figure S2**). The level of ligustilide was found to be the highest among the six phthalides in the ZC (6.191–6.713 mg/g) and ZT samples (4.186–6.586 mg/g). The total amount of the six phthalides was 5.421 and 7.258 mg/g in the ZT and ZC samples, respectively. The average levels of ligustilide, butylphthalide, senkyunolide H, and senkyunolide A were found to be significantly higher in the ZC samples (6.439 ± 0.22 mg/g, 181.6 ± 37.08, 2.795 ± 0.66, and 535.5 ± 81.09 μg/g, respectively) compared to those in the ZT samples (5.327 ± 0.95 mg/g, 3.088 ± 1.30, 1.516 ± 0.52, and 7.495 ± 2.49 μg/g, respectively; $p < 0.05$). However, the levels of senkyunolide I and butylidenephalide were not significantly different between the ZC (56.20 ± 8.45 and 42.47 ± 6.82 μg/g, respectively) and ZT samples (43.51 ± 17.60 and 37.07 ± 13.10 μg/g, respectively; **Figure 1A**; **Supplementary Table S2**).

Through the absolute quantification of a single component, the levels of all six phthalides were lower in the ZT samples than those in the ZC samples (**Figure 1**). Moreover, the ratio of each phthalide level to the total phthalide level were distorted in the ZT and ZC samples (**Figure 1B**). The ratios of senkyunolide H, senkyunolide I, and butylidenephalide showed no obvious changes between the ZC (0.04, 0.78, and 0.58%, respectively) and ZT samples (0.03, 0.78, and 0.67%, respectively). However, the ratio of butylphthalide to senkyunolide A was significantly higher in the ZC samples (2.51 and 7.39%, respectively) than that in the ZT samples (0.06 and 0.15%, respectively; $p < 0.05$). The ratio of ligustilide was significantly higher in the ZT samples (98.33%) than that in the ZC samples (88.70%; $p < 0.05$). We then selected three characteristic phthalides with large differences for the diagnostic ratio analysis (**Figure 1C**). The diagnostic ratios of ligustilide/butylphthalide, ligustilide/senkyunolide A, and senkyunolide A/butylphthalide were 1884.01, 813.63, and 3.07, respectively, in the ZT samples and 36.94, 12.31, 2.97, respectively, in the ZC samples. These results reflect the change in the composition ratio and the difference in the amounts of the phthalides.

Differential Transcriptomic Analysis in Roots of Normal and Early Flowering *Angelica sinensis*

The RNA-seq yielded 41.36 Gb of clean data, with an average of 6.89 Gb for each sample, with 90.74% of bases scoring > Q30 (**Supplementary Table S4**). A total of 91,519 isoforms were obtained after assembly. The N50 length obtained was approximately 1,631 bp for the ZC samples and 1904 bp for the ZT samples. The transcriptome data results were validated by qRT-PCR, including five highly expressed genes in ZT (**Supplementary Figures S3A,B**) and five highly expressed genes in ZC (**Supplementary Figures S3C,D**). The results showed that the expression levels of the transcriptome were generally consistent with the gene expression trends detected by qRT-PCR, which proved that the transcriptome sequencing results were reliable. We performed functional annotation of the isoforms using various databases, including NR, Swiss-Prot, KEGG, Clusters of Orthologous Genes (COG), EuKaryotic Orthologous Groups (KOG), and GO. Gene expression was estimated using fragments per kilobase of exon per million fragments mapped (FPKM). To identify the differentially expressed genes (isoforms) relevant to phthalide components, we compared the FPKM values of each isoform in ZC to those in ZT samples and retained the isoforms with fold change >2 and a false discovery rate correction set at $p < 0.05$ (Livak and Schmittgen, 2001). The volcano plot was showed that there were 8,824 different isoforms, including 4,455 upregulated and 4,369 downregulated isoforms, found using ZT as the control (**Figure 2A**). The cluster plot was

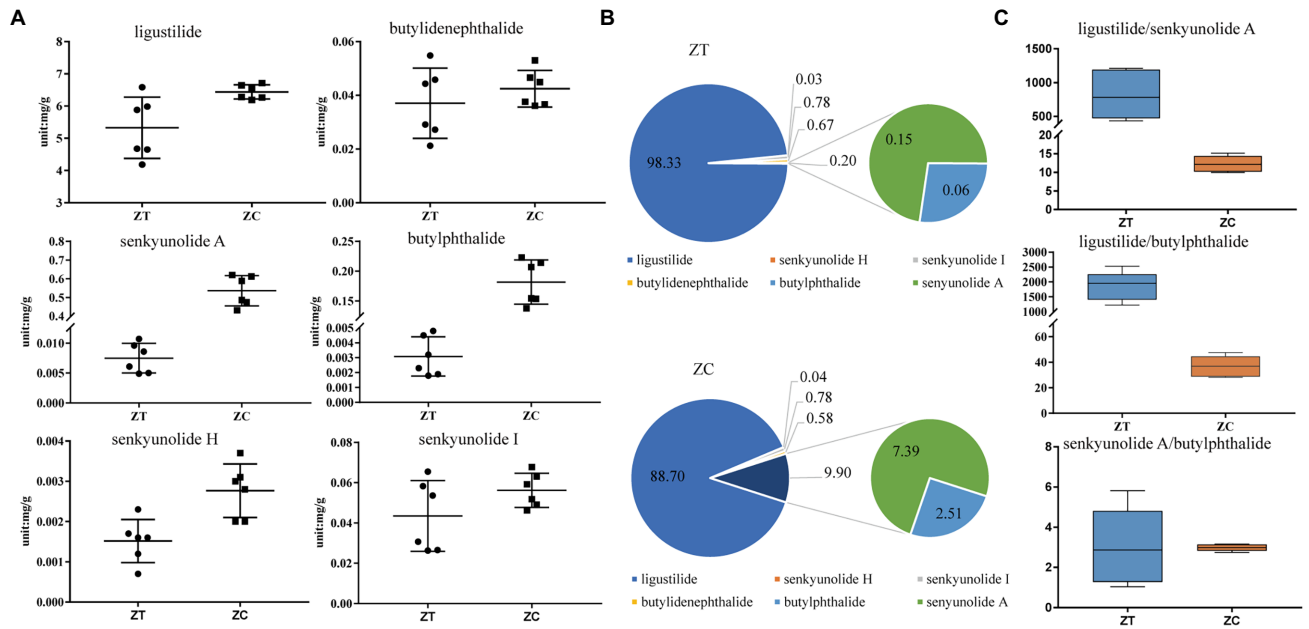


FIGURE 1 | Levels of the tested components across samples. **(A)** Content levels of the six phthalides in ZT and ZC samples; **(B)** The percentage of each phthalide level in the total phthalide of ZT and ZC samples; **(C)** The diagnostic ratios of three characteristic phthalides with large differences, ligustilide/butylphthalide, ligustilide/senkyunolide A, and senkyunolide A/butylphthalide, in ZT and ZC samples.

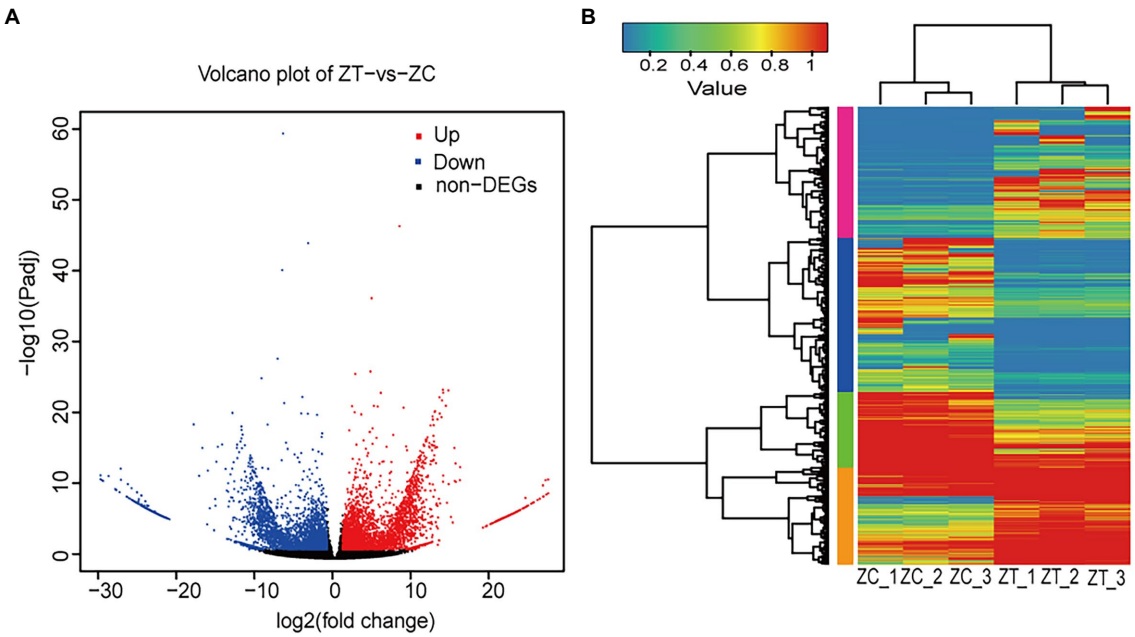


FIGURE 2 | The Volcano plot and Cluster plot of the differential expression isoforms in ZC and ZT samples [**(A)** Volcano plot; **(B)** Cluster plot. ZC: root samples from normal flowering plant; ZT: root samples from early flowering plant].

showed that the expression of isoforms in the two group has a difference in four subclusters (**Figure 2B**). There was a high expression of ZC sample in the subcluster_1 and subcluster_4, and a low expression of ZC sample in the subcluster_2 and subcluster_3 (**Supplementary Figure S4**). Then, the GO and KEGG enrichment analysis were performed on the differential isoforms.

A GO enrichment analysis was conducted to identify the biological functions of the upregulated and downregulated isoforms obtained from the different combinations

(**Supplementary Figure S5**). The significantly different isoforms were enriched in terms that are divided into three categories: biological process, cellular component, and molecular function. We observed that the isoforms were enriched in GO terms such as metabolic processes, cellular processes, cells, cell parts, binding, and catalytic activity. The significant enrichment of cell and metabolic processes may be related to the early bolting phenomenon, which is mainly manifested in the rapid growth and development of plants and significant changes in the components of medicinally active ingredients. A KEGG enrichment analysis was also performed (**Supplementary Figure S6**). Significant enrichment was obtained with functions such as carbon metabolism, starch and sucrose metabolism, biosynthesis of amino acids, plant hormone signal transduction, and the mitogen activated protein kinase signaling pathway. KEGG enrichment analysis also involved primary and secondary metabolism and growth and development hormone regulation.

A relationship analysis was conducted to reveal a Spearman correlation between isoform expression and phthalide content (**Supplementary Table S5**). Due to the higher content of phthalide in ZC, transcripts with higher expression levels in ZC than ZT were selected for further analysis. Correlation analysis indicated that 57 enzymes had significant positive correlations with phthalide accumulation (**Supplementary Table S6**). Based on the results of the correlation analysis and isoform function annotation, 108 isoforms were used as the key candidate isoforms for further verification experiments (**Figure 3**).

Validation of the Potential Regulatory Role of Key Candidate Isoforms in the Differential Phthalides Content

UHPLC–MS/MS was applied to simultaneously determine six markers in tissue culture seedlings of *A. sinensis* after MeJA treatment (**Supplementary Table S7**). There was an upward trend in the average butylphthalide, and senkyunolide A levels after MeJA treatment. There was an upward trend in the average ligustilide level during 0-48 h MeJA treatment, but a downward trend during 48-96 h MeJA treatment. In contrast, the average levels of butylidenephthalide, senkyunolide I, and senkyunolide H had a downward trend during 0-48 h MeJA treatment, but an upward trend during 48-96 h MeJA treatment. The expression of 108 candidate isoforms was determined using qPCR. This trend in the expression of the seven isoforms was consistent with that of the corresponding phthalide level (**Figure 4**; **Supplementary Table S8**). To ensure the reliability of the results, actin was also used as the housekeeping gene in addition to the 18S rRNA gene to verify the selected isoforms again (**Supplementary Figure S7**). The housekeeping genes 18S rRNA and actin were proved to be stably expressed in the seeds of *A. sinensis* treated with MeJA at different times, based on analysis using BestKeeper software (Pfaffl et al., 2004). The candidate isoforms of phospho-2-dehydro-3-deoxyheptonate aldolase 2 (17) and primary amine oxidase-like (23, 24) may lead to changes in the ligustilide levels, whereas tyrosine decarboxylase (38) may be the reason for the changes in the senkyunolide A levels. The candidate isoforms of shikimate dehydrogenase (21) and polyphenol oxidase (36) are associated with changes in senkyunolide I, senkyunolide H, and shikimate O-hydroxycinnamoyl transferase (43), which showed a trend toward changes in the levels of butylidenephthalide.

Functional Characterization of Key Candidate Isoforms

To examine these enzymes *in vitro*, we expressed the seven isoforms in *E. coli,* and the proteins were extracted and purified (**Supplementary Figure S8**). The extracts of *A. sinensis* were incubated with enzymes and subjected to UHPLC–MS/MS analysis. In accordance with *in vitro* experimental results (**Supplementary Figure S9**; **Figure 5**), 17 increased butylphthalide levels, whereas 21 increased enkyunolide H levels but decreased butylidenephthalide levels. In addition, 21 and 23 can increase ligustilide levels, 24 can increase the levels of ligustilide, butylidenephthalide, and senkyunolide A, 36 can increase senkyunolide I levels, 38 can increase the levels of senkyunolide A and butylphthalide, and 43 can increase butylidenephthalide levels. However, whether these enzymes are indispensable in the phthalide biosynthetic pathway is not certain and would require knockout experiments to verify this. The data presented here show that overexpression of these seven enzymes had an important influence on the accumulation of phthalides.

DISCUSSION

Angelica sinensis has a long history of use as a traditional herbal medicine and spice in food processing. Hence, analyzing the characteristic components of *A. sinensis* can contribute to its better application in real life. As the plant develops and grows, the effective components continue to be synthesized and are accumulated, and the ratio of the different components is maintained in a balance. However, early flowering results in a decrease in the content of these components, and the ratio becomes distorted. The analysis of the metabolite components of the ZT and ZC samples reveals that the levels of ligustilide, senkyunolide H, senkyunolide I, butylidenephthalide, butylphthalide, and senkyunolide A in the ZT samples were reduced compared with those of the ZC samples. Moreover, we should not only pay attention to the absolute amount of the individual biologically active component but also monitor the changes in the proportions of the different components to ensure better quality and yield of medicinal materials and thus better clinical treatment. The diagnostic ratio refers to the ratio between the specific components in a sample. It can characterize the respective chemical compositions of different samples and is used to determine whether the sources of the two samples are the same (Liu et al., 2021). Because it is less affected by the outside world and is a more diversified evaluation criterion, it is widely used in the traceability and identification of pollutants in the environment (Biache et al., 2014), medical disease diagnosis (Nauck and Meier, 2012), and identification of Chinese medicinal materials in food items (Wu et al., 2022). It is also used to

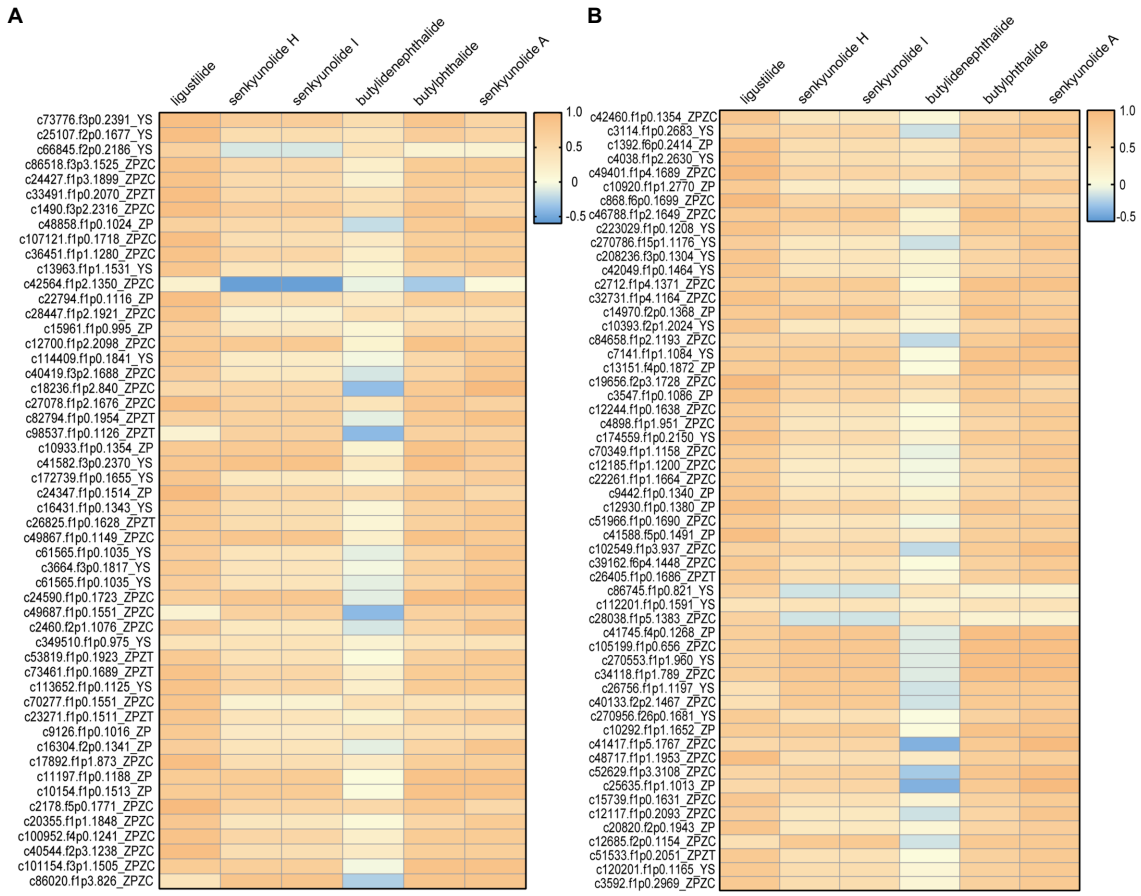


FIGURE 3 | The 108 isoforms that exhibited a significant positive/negative correlation with phthalide accumulation. **(A)** 1-54; **(B)** 55-108.

monitor the quality related to the production process of Chinese medicinal materials, about which there are relatively few reports. For example, in Traditional Chinese Medicine, the head, body, and tail of the *A. sinensis* roots are used to treat different diseases owing to their different pharmaceutical functions (Yang et al., 2021). A large amount of research evidence suggests that the difference in the amount of the components in the head, body, and tail of the roots may be the main reason for its different pharmaceutical efficacies (Xue et al., 2012). Thus, it may be a more comprehensive strategy to evaluate the quality of medicinal materials using multiple methods of multi-component monitoring and characteristic-component diagnostic ratios. During the planting process, the early flowering rate of *A. sinensis* reaches 20–30%. Once early flowering occurs, the roots of *A. sinensis* are lignified and cannot be used as medicine, which has resulted in a huge waste of resources and economic loss. The ratios of ligustilide/butylphthalide (1885 in ZT, 37 in ZC) and ligustilide/senkyunolide A (814 in ZT, 12 in ZC) also revealed a steady-state imbalance when plants are flowering early. Therefore, cultivating early flowering-resistant plants has become more popular, resulting in more effective plant ingredients available for medicinal use.

This study aimed to determine the key enzymes in phthalide accumulation and the molecular mechanism underlying the synthesis and accumulation of phthalide components to provide a feasible strategy for increasing the phthalide levels in the flowering *A. sinensis* to reduce waste, improve the market supply of medicinal raw materials, and alleviate the economic losses of farmers. Studies on the mechanism underlying phthalide biosynthesis have been reported previously, but the enzymes related to the accumulation of phthalides, especially the enzymes involved in different phthalide transformations, have not yet been identified, requiring further investigation (León et al., 2017). In this study, we combined transcriptome and targeted metabolite profile analyses to explore potential enzymes or pathways involved in the differential regulation of phthalide accumulation. Six enzymes, including phospho-2-dehydro-3-deoxyheptonate aldolase 2 (17), shikimate dehydrogenase (21), primary amine oxidase-like (23, 24), polyphenol oxidase (36), tyrosine decarboxylase (38), and shikimate *O*-hydroxycinnamoyl transferase (43), have shown potential for the regulation of phthalide accumulation.

Phospho-2-dehydro-3-deoxyheptonate aldolase 2 (17) and shikimate dehydrogenase (21) catalyze the first and fourth committed steps of the shikimate pathway, respectively, both of which are required for the synthesis of aromatic amino acids and other aromatic metabolites in bacteria, microbial

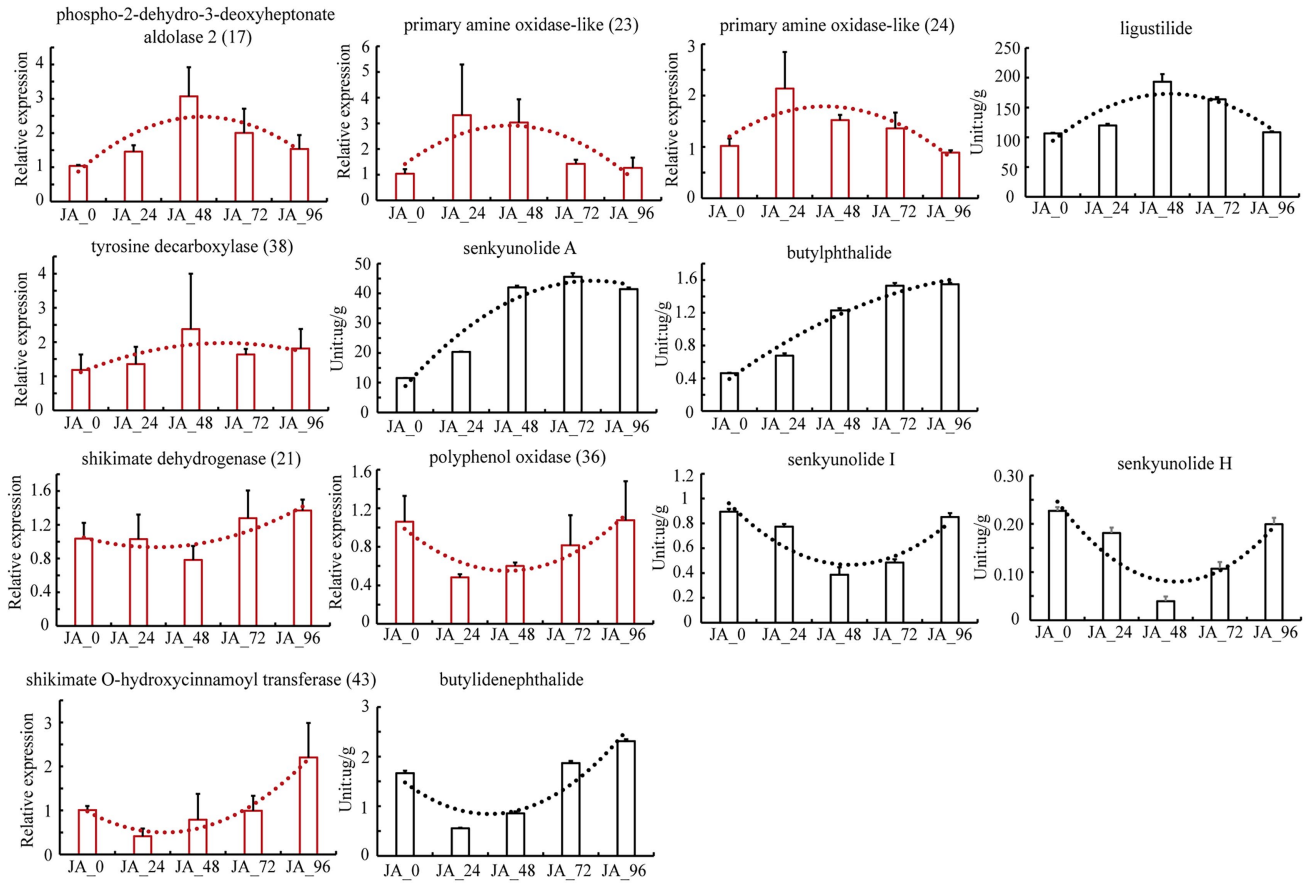


FIGURE 4 | The levels and relative expression of the tested components and isoforms across samples after treatment with methyl jasmonate.

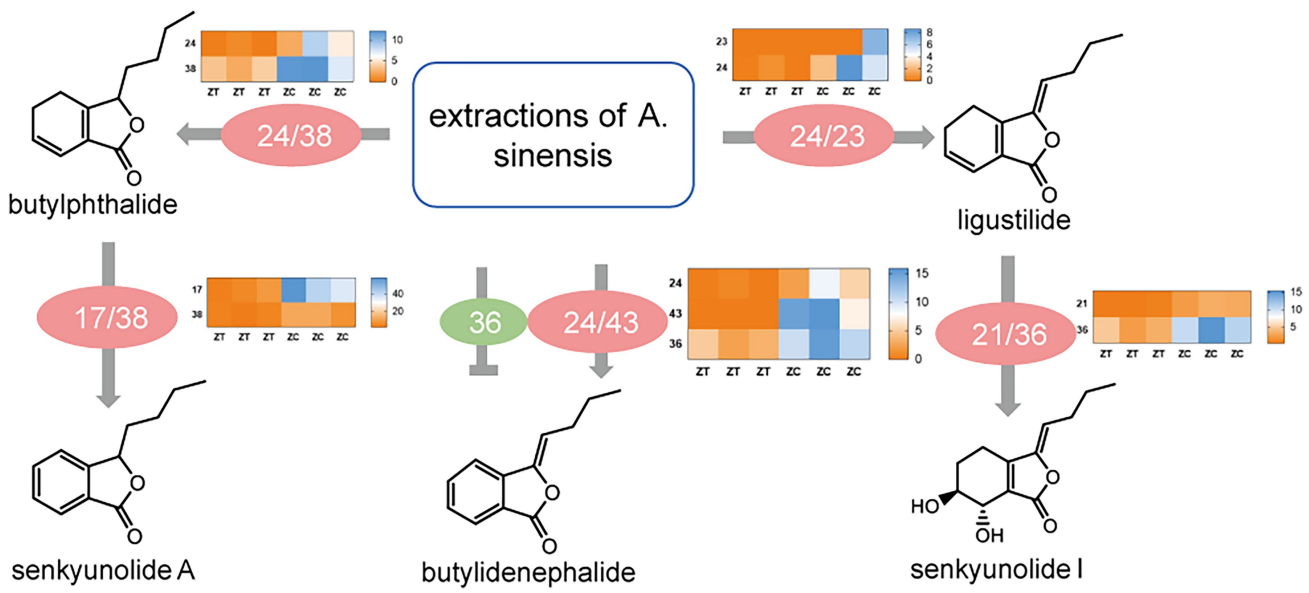


FIGURE 5 | Schematic model of the function of key candidate isoforms in phthalide accumulation and transformation and heatmap of the differential expression of genes involved in phthalide accumulation in ZT and ZC samples. ZC, root samples from normal flowering plant; ZT, root samples from early flowering plant.

eukaryotes, and plants (Bagautdinov and Kunishima, 2007; Heyes et al., 2014). Polyphenol oxidase (36) is a group of Cu-containing enzymes that catalyzes the oxidation of several phenols to *o*-quinones (Prexler et al., 2019). Polyphenol oxidases participate in two oxidation reactions. The first is hydroxylation of the ortho-position adjacent to an existing hydroxyl group. The second mechanism is the oxidation of o-dihydroxybenzenes to o-benzoquinones (Taranto et al., 2017). Tyrosine decarboxylase (38), a pyridoxal phosphate-dependent amino acid decarboxylase, is a key enzyme in dopamine synthesis, and its catalytic products are implicated in the defense response (Gayathri and Manoj, 2020). In the tyrosine metabolic pathway, l-tyrosine is used as a substrate to catalyze its decarboxylation to form tyramine. Moreover, tyrosine decarboxylase can also decarboxylate phenylalanine to produce phenylethylamine, another biogenic amine (Marcobal et al., 2012). Shikimate *O*-hydroxycinnamoyl transferase (43) catalyzes the synthesis of shikimate and quinate esters. It appears to control the biosynthesis and turnover of major plant phenolic compounds, such as lignin and chlorogenic acid (Hoffmann et al., 2003). Although the specific reactions that participate in the synthesis and transformation of phthalides are unknown, the catalytic reaction of heterologously expressed proteins in *E. coli* has proved that they promote the accumulation of certain phthalides.

Through the analysis of these enzyme reactions, it is speculated that they may regulate the conversion of different phthalide components through various pathways, such as oxidation, isomerization, and hydroxylation. As the most abundant phthalide in *A. sinensis*, ligustilide is a volatile and unstable compound with an α, β-unsaturated lactone in its structure. Senkyunolide I and 6, 7-epoxyligustilide were the major degradation products when ligustilide was stored at room temperature under direct sunlight. Ligustilide is likely to degrade into 6, 7-epoxyligustilide through oxidation and then transform into senkyunolide I by further hydrolysis (Zou et al., 2013). In addition, senkyunolide I and 6, 7-epoxyligustilide are also the *in vivo* metabolites of ligustilide (Ding et al., 2008; Yan et al., 2008). These studies suggest that they may have the same synthesis and transformation pathways in plants. In our study, polyphenol oxidase (36) and shikimate dehydrogenase (21) contributed to the transformation of ligustilide to senkyunolide I. Polyphenol oxidase (36) can react with o-benzoquinone and water to produce catechol and

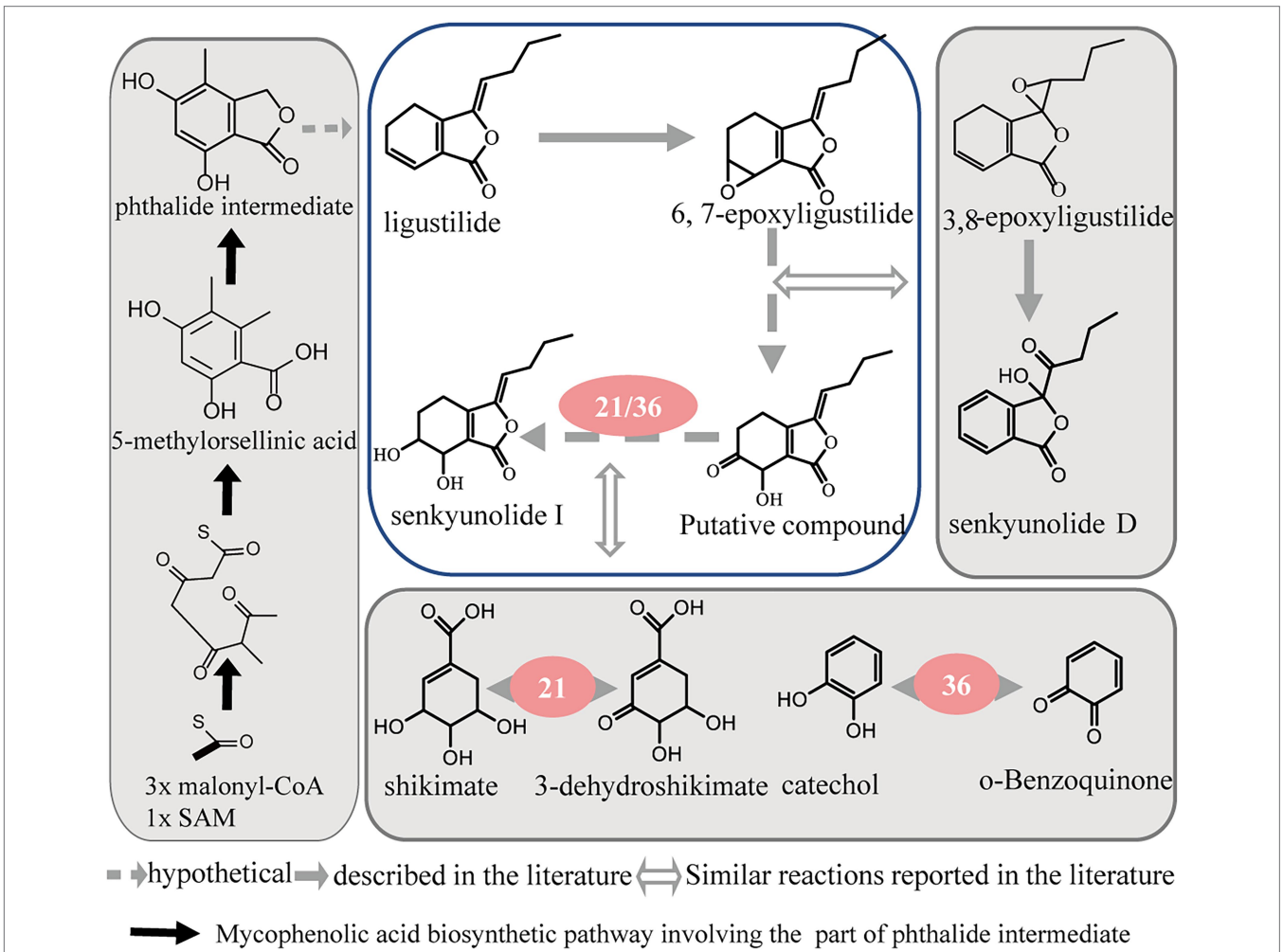


FIGURE 6 | Proposed reaction pathways from ligustilide to senkyunolide I.

oxygen. 3-dehydroshikimate, NADPH, and H^+ generate shikimate and $NADP^+$ under the action of shikimate dehydrogenase (21). In the above reaction, polyphenol oxidase (36) and shikimate dehydrogenase (21) mainly act on the conversion of the compound from a carbonyl to a hydroxyl group. The proposed reaction pathways from ligustilide to senkyunolide I are shown in **Figure 6** by summarizing the enzyme function and the present study on phthalide transformation (Schinkovitz et al., 2008).

Overall, our study explored candidate enzymes corresponding to different phthalide compounds that attempt to build phthalide biosynthetic and transformed pathways. Seven candidate isoforms involved in phthalide accumulation and transformation have been identified. Due to the lack of relevant research, we could not fully clarify the mechanism of action of each enzyme. Only the proposed reaction pathway from ligustilide to senkyunolide I is shown. Further investigations into the enzymatic properties and their regulatory mechanisms are required to completely understand phthalide accumulation and transformation in *A. sinensis*. These findings may also provide insights into the genes that can make for potential genetic engineering targets for enriching phthalides in *A. sinensis*.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA749925.

AUTHOR CONTRIBUTIONS

W-MF, PL, HY, and J-AD conceived and designed the experiments. W-MF performed the experiments. W-MF, PL, E-XS, and SZ analyzed the data. J-AD, D-WQ, SZ, GY, HY, and SJ contributed reagents, materials, and analysis tools. W-MF and PL wrote the manuscript. All authors have read and approved the final manuscript.

FUNDING

This research was supported financially by Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202005). China Agriculture Research System of MOF and MARA (No. CARS-21), Ministry of Finance Central Level of the Special (No. 2060302), National Natural Science Foundation of China (81773848). This work was also partly sponsored by Six talents peaks project in Jiangsu Province (JNHB-066), Jiangsu Province 333 High-level Talents Training Project, Qing Lan Project, the Major Projects of Natural Science Foundation of Universities in Jiangsu Province (19KJA320002) and Natural Science Foundation of Jiangsu Province, China (BK20201403).

ACKNOWLEDGMENTS

We thank Shanghai Majorbio Bio-Pharm Technology Co. Ltd. for providing a sequencing platform and Frasergen Biotechnology Co. Ltd. for conducting high-throughput sequencing detection and analysis. We also thank Editage[2] for its linguistic assistance during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.928760/full#supplementary-material

[2]https://www.editage.com/

REFERENCES

Bagautdinov, B., and Kunishima, N. (2007). Crystal structures of shikimate dehydrogenase AroE from *Thermus thermophilus* HB8 and its cofactor and substrate complexes: insights into the enzymatic mechanism. *J. Mol. Biol.* 373, 424–438. doi: 10.1016/j.jmb.2007.08.017

Biache, C., Mansuy-Huault, L., and Faure, P. (2014). Impact of oxidation and biodegradation on the most commonly used polycyclic aromatic hydrocarbon (PAH) diagnostic ratios: implications for the source identifications. *J. Hazard. Mater.* 267, 31–39. doi: 10.1016/j.jhazmat.2013.12.036

Birch, A. J., English, R. J., Massy-Westropp, R. A., Slaytor, M., and Smith, H. (1958). Studies in relation to biosynthesis. Part XIV. The origin of nuclear methyl groups in mycophenolic acid. *J. Chem. Soc.* 67:365. doi: 10.1039/jr9580000365

Chae, S. H., Kim, S. I., Yeon, S. H., Lee, S. W., and Ahn, Y. J. (2011). Adulticidal activity of phthalides identified in Cnidium officinale rhizome to B- and Q-biotypes of *Bemisia tabaci*. *J. Agric. Food Chem.* 59, 8193–8198. doi: 10.1021/jf201927t

Ding, C., Sheng, Y., Zhang, Y., Zhang, J., and Du, G. (2008). Identification and comparison of metabolites after oral administration of essential oil of Ligusticum chuanxiong or its major constituent ligustilide in rats. *Planta Med.* 74, 1684–1692. doi: 10.1055/s-0028-1088309

Fang, L., Xiao, X. F., Liu, C. X., and He, X. (2012). Recent advance in studies on *Angelica sinensis*. *Chin. Herbal Med.* 1, 12–25. doi: 10.3969/j.issn.1674-6384.2012.01.004

Feng, W. M., Liu, P., Yan, H., Zhang, S., Shang, E. X., Yu, G., et al. (2021). Impact ho, of *Bacillus velezensis* on Phthalides accumulation in *Angelica sinensis* (Oliv.) by stoichiometry and microbial diversity analysis. *Front. Microbiol.* 11:611143. doi: 10.3389/fmicb.2020.611143

Gayathri, S. C., and Manoj, N. (2020). Crystallographic snapshots of the Dunathan and Quinonoid intermediates provide insights into the reaction mechanism of group II decarboxylases. *J. Mol. Biol.* 24:166692. doi: 10.1016/j.jmb.2020.10.026

Heyes, L. C., Reichau, S., Cross, P. J., Jameson, G. B., and Parker, E. J. (2014). Structural analysis of substrate-mimicking inhibitors in complex with *Neisseria meningitidis* 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase - The importance of accommodating the active site water. *Bioorg. Chem.* 57, 242–250. doi: 10.1016/j.bioorg.2014.08.003

Ho, T. T., Murthy, H. N., and Park, S. Y. (2020). Methyl Jasmonate induced oxidative stress and accumulation of secondary metabolites in plant cell and organ cultures. *Int. J. Mol. Sci.* 3:716.

Hoffmann, L., Maury, S., Martz, F., Geoffroy, P., and Legrand, M. (2003). Purification, cloning, and properties of an acyltransferase controlling shikimate

and quinate ester intermediates in phenylpropanoid metabolism. *J. Biol. Chem.* 278, 95–103. doi: 10.1074/jbc.M209362200

Isman, M. B. (2006). Botanical insecticides, deterrents, and repellents in modern agriculture and an increasingly regulated world. *Annu. Rev. Entomol.* 51, 45–66. doi: 10.1146/annurev.ento.51.110104.151146

Karmakar, R., Pahari, P., and Mal, D. (2014). Phthalides and phthalans: synthetic methodologies and their applications in the total synthesis. *Chem. Rev.* 114, 6213–6284. doi: 10.1021/cr400524q

León, A., Del-Ángel, M., Ávila, J. L., and Delgado, G. (2017). Phthalides: distribution in nature, chemical reactivity, synthesis, and biological activity. *Prog. Chem. Org. Nat. Prod.* 104, 127–246. doi: 10.1007/978-3-319-45618-8_2

Li, D., Chen, G., Ma, B., Zhong, C., and He, N. (2020a). Metabolic profiling and transcriptome analysis of mulberry leaves provide insights into flavonoid biosynthetic. *J. Agric. Food Chem.* 68, 1494–1504. doi: 10.1021/acs.jafc.9b06931

Li, M. F., Kang, T. L., Jin, L., and Wei, J. H. (2020c). Research progress on bolting and flowering of Angelica Sinensis and regulation pathways. *Chin. Tradit. Herb. Drug* 51, 5894–5899. doi: 10.7501/j.issn.0253-2670.2020.22.029

Li, M., Li, J., Wei, J., and Paré, P. W. (2021). Transcriptional controls for early bolting and flowering in *Angelica sinensis*. *Plants* 10:1931. doi: 10.3390/plants10091931

Li, L., Wu, H. X., Ma, X. W., Xu, W. T., Liang, Q. Z., Zhan, R. L., et al. (2020b). Transcriptional mechanism of differential sugar accumulation in pulp of two contrasting mango (*Mangifera indica* L.) cultivars. *Genomics* 112, 4505–4515. doi: 10.1016/j.ygeno.2020.07.038

Liu, G. J., Wang, B., Jiang, F., Wu, W. Q., Gao, F., and Fan, X. L. (2021). Identification of *Salvia miltrorrhiza* lllegally added in food based on the diagnostic ratios of characteristic componennts. *Food Sci. Technol.* 46, 277–283.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C (T)) method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lv, Z. Y., Sun, W. J., Jiang, R., Chen, J. F., Ying, X., Zhang, L., et al. (2021). Phytohormones jasmonic acid, salicylic acid, gibberellins, and abscisic acid are key mediators of plant secondary metabolites. *World J. Tradit. Chin. Med.* 7, 307–325. doi: 10.4103/wjtcm.wjtcm_20_21

Marcobal, A., De Las Rivas, B., Landete, J. M., Tabera, L., and Muñoz, R. (2012). Tyramine and phenylethylamine biosynthetic by food bacteria. *Crit. Rev. Food Sci. Nutr.* 52, 448–467. doi: 10.1080/10408398.2010.500545

Mitsuhashi, H., and Nomura, M. (1966). Studies on the constituents of umbelliferae plants. XII. Biogenesis of 3-butylphthalide. *Chem. Pharm. Bull.* 14, 777–778. doi: 10.1248/cpb.14.777

Miyazawa, M., Tsukamoto, T., Anzai, J., and Ishikawa, Y. (2004). Insecticidal effect of phthalides and furanocoumarins from Angelica acutiloba against *Drosophila melanogaster*. *J. Agric. Food Chem.* 52, 4401–4405. doi: 10.1021/jf0497049

Nauck, M. A., and Meier, J. J. (2012). Diagnostic accuracy of an "amended" insulin-glucose ratio for the biochemical diagnosis of insulinomas. *Ann. Intern. Med.* 157, 767–775. doi: 10.7326/0003-4819-157-11-201212040-00004

Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: Best Keeper-excel-based tool using pair-wise correlations. *Biotechnol. Lett.* 26, 509–515. doi: 10.1023/B:BILE.0000019559.84305.47

Plasencia, A., Soler, M., Dupas, A., Ladouce, N., Silva-Martins, G., Martinez, Y., et al. (2016). Eucalyptus hairy roots, a fast, efficient and versatile tool to explore function and expression of genes involved in wood formation[J]. *Plant Biotechnol. J.* 14, 1381–1393. doi: 10.1111/pbi.12502

Prexler, S. M., Frassek, M., Moerschbacher, B. M., and Dirks-Hofmeister, M. E. (2019). Catechol oxidase versus Tyrosinase classification revisited by site-directed mutagenesis studies. *Angew. Chem. Int. Ed. Engl.* 58, 8757–8761. doi: 10.1002/anie.201902846

Schinkovitz, A., Pro, S. M., Main, M., Chen, S. N., Jaki, B. U., Lankin, D. C., et al. (2008). Dynamic nature of the ligustilide complex. *J. Nat. Prod.* 71, 1604–1611. doi: 10.1021/np800137n

Tang, F., Yan, Y. M., Yan, H. L., Wang, L. X., Hu, C. J., Wang, H. L., et al. (2021). Chuanxiongdiolides R4 and R5, phthalide dimers with a complex polycyclic skeleton from the aerial parts of Ligusticum chuanxiong and their vasodilator activity. *Bioorg. Chem.* 107:104523. doi: 10.1016/j.bioorg.2020.104523

Taranto, F., Pasqualone, A., Mangini, G., Tripodi, P., Miazzi, M. M., Pavan, S., et al. (2017). Polyphenol oxidases in crops: biochemical, physiological and genetic aspects. *Int. J. Mol. Sci.* 18:377. doi: 10.3390/ijms18020377

Wei, W. L., and Huang, L. F. (2015). Simultaneous determination of ferulic acid and phthalides of Angelica sinensis based on UPLC-Q-TOF/MS. *Molecules* 3, 4681–4694.

Wu, W. Q., Jiang, F., Gan, G. P., Fan, X. L., Liu, G. J., Wang, B., et al. (2022). Identification of illegally added Epimedium in food based on high resolution mass spectrometry library and diagnostic ratio of characteristic components. *Modern Food Sci. Technol.* 38, 1–10.

Xiong, R., Chen, Z., Wang, W., Jiang, L., Xiang, Y., and Fan, J. (2020). Combined transcriptome sequencing and prokaryotic expression to investigate the key enzyme in the 2-C-methylerythritol-4-phosphate pathway of *Osmanthus fragrans*. *Funct. Plant Biol.* 47, 945–958. doi: 10.1071/FP19365

Xue, W. X., Hua, Y. L., Guo, Y. S., Ji, P., Wu, H. Y., and Wei, Y. M. (2012). Change of composition in different parts of Angelicae sinensis based on geoherbs. *J. Gansu Agric. Univ.* 2, 149–154.

Yan, R., Ko, N. L., Li, S. L., Tam, Y. K., and Lin, G. (2008). Pharmacokinetics and metabolism of ligustilide, a major bioactive component in Rhizoma Chuanxiong, in the rat. *Drug Metab. Dispos.* 36, 400–408. doi: 10.1124/dmd.107.017707

Yang, J., Zhang, C., Li, W. H., Zhang, T. E., Fan, G. Z., Guo, B. F., et al. (2021). Comprehensive analysis of Transcriptomics and metabolomics between the heads and tails of Angelica Sinensis: genes related to Phenylpropanoid biosynthesis pathway. *Comb. Chem. High Throughput Screen.* 24, 1417–1427. doi: 10.2174/1386207323999201103221952

Yi, L., Liang, Y., Wu, H., and Yuan, D. (2009). The analysis of Radix *Angelicae sinensis* (Danggui). *J. Chromatogr. A* 1216, 1991–2001. doi: 10.1016/j.chroma.2008.07.033

Yu, G., Zhou, Y., Yu, J. J., Hu, X. Q., Tang, Y., Yan, H., et al. (2019). Transcriptome and digital gene expression analysis unravels the novel mechanism of early flowering in *Angelica sinensis*. *Sci. Rep.* 9:10035. doi: 10.1038/s41598-019-46414-2

Zhao, Q., Qiu, B., Li, S., Zhang, Y., Cui, X., and Liu, D. (2020). Osmotin-Like protein gene from Panax notoginseng is regulated by Jasmonic acid and involved in defense responses to *Fusarium solani*. *Phytopathology* 110, 1419–1427. doi: 10.1094/PHYTO-11-19-0410-R

Zhou, W., Shi, M., Deng, C., Lu, S., Huang, F., Wang, Y., et al. (2021). The methyl jasmonate-responsive transcription factor SmMYB1 promotes phenolic acid biosynthetic in *Salvia miltiorrhiza*. *Hortic Res.* 8:10. doi: 10.1038/s41438-020-00443-5

Zou, J., Chen, G. D., Zhao, H., Huang, Y., Luo, X., Xu, W., et al. (2018). Triligustilides A and B: two pairs of phthalide trimers from *Angelica sinensis* with a complex polycyclic skeleton and their activities. *Org. Lett.* 20, 884–887. doi: 10.1021/acs.orglett.8b00017

Zou, A. H., Cheng, M. C., Zhuo, R. J., Wang, L., and Xiao, H. B. (2013). Structure elucidation of degradation products of Z-ligustilide by UPLC-QTOF-MS and NMR spectroscopy. *Acta. Pharma. Sinica.* 48, 911–916.

Conflict of Interest: The 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fpls.2022.989036

OPEN ACCESS

EDITED BY
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Pengda Ma,
Northwest A&F University, China
Junbo Gou,
Agricultural Genomics Institute at Shenzhen (CAAS), China

*CORRESPONDENCE
Jianjun Li
043081@htu.edu.cn
Juanjuan Yu
yujuan8186@163.com

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 08 July 2022
ACCEPTED 22 August 2022
PUBLISHED 12 September 2022

CITATION
Li J, Yu X, Shan Q, Shi Z, Li J, Zhao X, Chang C and Yu J (2022) Integrated volatile metabolomic and transcriptomic analysis provides insights into the regulation of floral scents between two contrasting varieties of *Lonicera japonica*. *Front. Plant Sci.* 13:989036. doi: 10.3389/fpls.2022.989036

Integrated volatile metabolomic and transcriptomic analysis provides insights into the regulation of floral scents between two contrasting varieties of *Lonicera japonica*

Jianjun Li[1]*, Xinjie Yu[1], Qianru Shan[2], Zhaobin Shi[2], Junhua Li[1], Xiting Zhao[1], Cuifang Chang[3] and Juanjuan Yu[2]*

[1]Green Medicine Biotechnology Henan Engineering Laboratory, Engineering Technology Research Center of Nursing and Utilization of Genuine Chinese Crude Drugs in Henan Province, College of Life Sciences, Henan Normal University, Xinxiang, China, [2]Henan International Joint Laboratory of Agricultural Microbial Ecology and Technology, College of Life Sciences, Henan Normal University, Xinxiang, China, [3]State Key Laboratory Cell Differentiation and Regulation, College of Life Sciences, Henan Normal University, Xinxiang, Henan, China

Lonicera japonica Thunb., belonging to the Caprifoliaceae family, is an important traditional Chinese medicinal plant. The *L. japonica* flower (LJF) is widely used in medicine, cosmetics, drinks, and food due to its medicinal and sweet-smelling properties. Considerable efforts have been devoted to investigating the pharmacological activities of LJF; however, the regulatory mechanism of the floral scents remains unknown. We previously selected and bred an elite variety of *L. japonica* var. *chinensis* Thunb. called 'Yujin2', which has a strong aroma and is used in functional drinks and cosmetics. In order to reveal the regulatory mechanism of the floral scents of LJF, volatile metabolomic and transcriptomic analyses of the LJF at the silver flowering stage of 'Yujin2' (strong aroma) and 'Fengjin1' (bland odor) were performed. Our results revealed that a total of 153 metabolites and 9,523 genes were differentially regulated in LJF between 'Yujin2' and 'Fengjin1'. The integrated analysis of omics data indicated that the biosynthetic pathways of terpenoids (i.e., monoterpenoids, including geraniol and alpha-terpineol; sesquiterpenoids, including farnesol, farnesal, and alpha-farnesene; triterpenoid squalene), tryptophan and its derivatives (methyl anthranilate), and fatty acid derivatives, were major contributors to the stronger aroma of 'Yujin2' compared to 'Fengjin1'. Moreover, several genes involved in the terpenoid biosynthetic pathway were characterized using quantitative real-time PCR. These results provide insights into the metabolic mechanisms and molecular basis of floral scents in LJF, enabling future screening of genes related to the floral scent regulation, such as alpha-terpineol synthase, geranylgeranyl diphosphate synthase, farnesyl pyrophosphate synthase, anthranilate synthase, as well as transcription

factors such as MYB, WRKY, and LFY. The knowledge from this study will facilitate the breeding of quality-improved and more fragrant variety of *L. japonica* for ornamental purpose and functional beverages and cosmetics.

KEYWORDS

***Lonicera japonica* Thunb., floral scent, metabolomics, transcriptomics, volatile**

Introduction

In plants, floral scents play vital roles in attracting pollinators, deterring pathogens and parasites, and serving as signals in response to biotic and abiotic stresses (Schiestl, 2015). In addition, floral scents are important commercial traits of ornamental plants and are also economically important for quality in the food, drink, perfume, cosmetic, and pharmaceutical industries (Mostafa et al., 2022). Floral scents consist of a mixture of volatile organic compounds (VOCs), which are lipophilic and are characterized by low molecular weights and high melting points. To date, more than 1,700 floral VOCs have been identified, which are categorized as terpenoids, phenylpropanoids/benzenoids, fatty acid derivatives, and amino acid derivatives. Over the past decade, a large number of studies on floral VOCs have improved our understanding of their functions, biosynthesis, and regulation. Currently, the research on floral scents is mainly focused on common ornamental plants, such as rose, orchid, and tulip (Mostafa et al., 2022). However, the constituents and abundances of the VOCs in floral scents vary widely among plants. The mechanisms underlying the floral scent regulation of certain important medicinal and edible plant species remain to be elucidated (Zhu et al., 2019; Mostafa et al., 2022).

Lonicera japonica Thunb., an important traditional Chinese medicinal plant, is a perennial semi-evergreen twining species of the Caprifoliaceae family member that is cultivated worldwide, particularly in Asian countries such as China, Japan, and Korea. The *L. japonica* flower (LJF) has been prescribed in traditional Chinese medicine for the treatment of infections, fever, sores, swelling, and influenza for thousands of years (Shang et al., 2011). Recently, LJF has also been demonstrated to inhibit influenza A viruses and COVID-19 (Zhou et al., 2015, 2020). Moreover, LJF is also widely used in cosmetics, food, beverages, and ornamental groundcovers due to its medicinal properties and sweet-smelling or attractively colored flowers (Wang et al., 2016).

Owing to its vital role, a considerable amount of research has been devoted to analyzing the chemical constituents and pharmacological activities of LJF (Yoo et al., 2008; Wang et al., 2016; Li et al., 2020). Among these, some progress has been made in analyzing the volatile components of LJF from different origins (Du et al., 2015), parts (Yan et al., 2020), treatments (Cai et al., 2013), and flowering stages (Wang et al., 2009). However, comparative analysis of VOC profiles among different *L. japonica* varieties is very scarce so far. The correlation between characteristic aroma pattern and VOC profile of LJF remains elusive.

'Yujin2', an elite variety of *L. japonica* var. *chinensis* Thunb. that was selected and bred by our group, has strong cold and drought resistances, high yield, as well as early and long flowering, with large flower buds. The LJF of 'Yujin2' has strong aroma and no bitterness and is purple-red at the budding stage and has thus become a good choice for flower tea and herbal drinks (Li et al., 2019). In contrast, the LJF of 'Fengjin1', the predominant variety of *L. japonica* Thunb. grown in Henan, has a bland odor, bitter taste, is green-to-white at the budding stage, and is mainly used in medicines. Therefore, 'Yujin2' and 'Fengjin1', two contrasting varieties with distinct aroma concentration, are excellent materials for researching the correlation between floral scent and VOC profile of LJF.

Omics technologies, especially transcriptomics, are effective technologies for studying the mechanisms underlying plant growth and development. Considerable numbers of transcriptomic studies have been conducted on LJF. However, these previous studies mainly focused on the mechanism of bioactive constituent biosynthesis during flower development (Li et al., 2019; Yang et al., 2019; Wang et al., 2020; Xia et al., 2021) and in response to salt stress (Cai et al., 2021, 2022) and light stress (Fang et al., 2020). Knowledge of the mechanism underlying the accumulation of VOCs associated with floral scents remains limited. The decoded genome of *L. japonica* has provided valuable information for research into gene function and transcriptomic analysis (Pu et al., 2020; Xiao et al., 2021).

In this study, headspace solid-phase microextraction (HS-SPME) coupled with GC–MS was performed for the identification of the VOCs of LJF at the silver stage (the stage with the strongest aroma) of these two contrasting varieties of *L. japonica* Thunb. Furthermore, correlation analysis of the transcriptomic profiles with metabolomic data was performed to elucidate the regulatory mechanism of the floral scents of LJF (Figure 1). Candidate genes were selected for validation *via* quantitative real-time PCR (qRT-PCR). These findings provide new information for the metabolic mechanisms and molecular basis of floral scents of LJF.

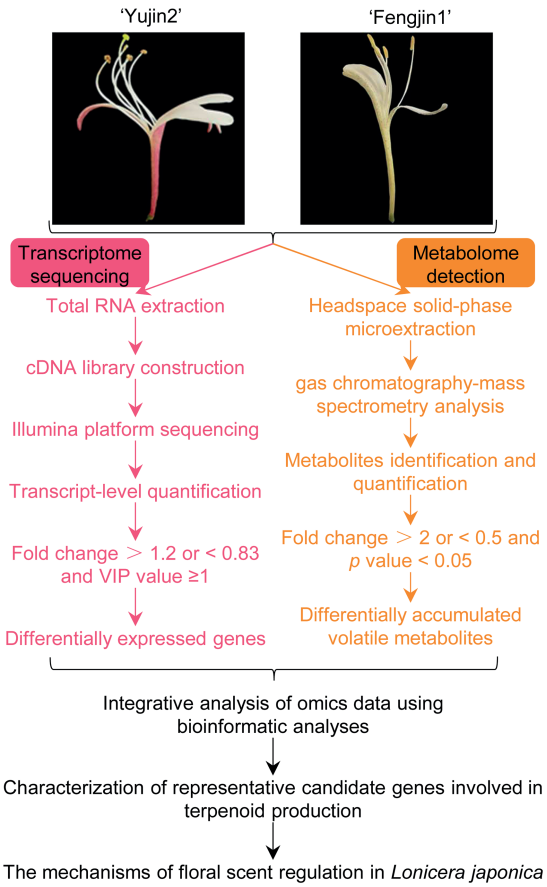


FIGURE 1
Overview of the metabolomics and transcriptomics workflow. *Lonicera japonica* flowers (LJFs) at the silver flowering stage of two contrasting varieties of *L. japonica*, 'Yujin 2' and 'Fengjin 1' were collected. Headspace solid-phase microextraction (HS-SPME) coupled with gas chromatography–mass spectrometry (GC–MS) and transcriptomics analysis were performed for volatile organic compound (VOC) profiling and global gene expression patterns, respectively. Integrated metabolomics and transcriptomics analysis serves to elucidate the regulatory mechanism of floral scents of LJFs by bioinformatic analyses.

Materials and methods

Plant materials

LJFs at the silver flowering stage of two contrasting varieties of *L. japonica*, 'Yujin 2' and 'Fengjin 1', were collected in May 2019 in the Fengqiu honeysuckle germplasm resource nursery (Henan, China, 114°47′ N, 35°20′ E). Three biological replicates were performed, and each replicate was collected from at least three separate plants.

HS-SPME and GC–MS analysis

The LJF samples were ground into powder in liquid nitrogen; then, 1.0 g of the powder was immediately placed in a 20 ml head-space vial containing 10 µl of internal standard (−)-carvone (50 µg/ml) and 2 ml of saturated NaCl solution. Then, at 100°C, the samples were shaken for 5 min, and a 120 µm divinylbenzene/carboxen/polydimethylsiloxane fiber was exposed to the headspace of the samples to absorb the volatiles for 15 min for GC–MS analysis.

The absorbed volatiles were analyzed using an Agilent Model 8,890 GC and a 5977B MS (Agilent Technologies, Stockport, United Kingdom) equipped with a DB-5MS capillary column (30 m × 0.25 mm × 0.25 µm). High-purity helium gas was used as a carrier gas with a velocity of 1.2 ml/min. The injector temperature was 250°C, and the detector temperature was 280°C. The oven temperature was programmed as follows: 40°C (3.5 min), increasing at 10°C/min to 100°C, at 7°C/min to 180°C, at 25°C/min to 280°C, and 280°C was maintained for 5 min. The MS was operated in full scan mode (50–500 amu at 1 scan/s) with electron ionization mode at 70 eV. The repeatability of the analysis process was monitored using a quality control (QC) sample after every 10 samples. Overlapping analysis of the total ion current in different QC samples was used to indicate the instrumental stability for GC–MS analysis.

Identification of VOCs was achieved by comparing the mass spectra with the data system library (MWGC) and linear retention index. Peak determination and peak area integration were carried out with MassHunter quantitative analyses (Agilent Technologies, Santa Clara, CA, United States). The relative content of each compound was calculated using the internal standard normalization method. Differentially accumulated volatiles (DAVs) were obtained using the threshold of quantitative fold change >1.2 or <0.83 and VIP value ≥1 between 'Yujin2' and 'Fengjin1'. The odor qualities of the metabolites were obtained from the PubChem database,[1] the Good Scent Company website,[2] and the literature.

RNA extraction, cDNA library construction and sequencing, transcript-level quantification, and DEG screening

Total RNA was isolated from 1 g of frozen sample using a mirVanaTM miRNA Isolation Kit (Ambion, Xian, China) according to the manufacturer's protocol. The integrity and purity of the RNA were assessed using an Agilent 2,100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States) and a NanoDrop microspectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA), respectively. The construction and sequencing of the cDNA library were conducted by OE Biotech Co., Ltd. (Shanghai, China). Three biological replicates per variety were conducted.

1 https://pubchem.ncbi.nlm.nih.gov/
2 http://www.thegoodscentscompany.com/index.html

The clean reads were obtained by removing reads containing poly-N and low-quality reads from the raw data. Then, the clean reads were mapped to the *L. japonica* reference genome[3] using HISAT. The read counts of each gene were obtained by htseq counting, and the fragments per kilobase million (FPKM) value of each gene was calculated using cufflinks. The differentially expressed genes (DEGs) were identified using the DESeq package in R software with a threshold of the fold change >2 or < 0.5 and $p < 0.05$.

Bioinformatic analysis of omics data

Pearson's correlation analysis and principal component analysis (PCA) were used to evaluate the data quality of the metabolome dataset and transcriptome dataset. Pearson's correlation analyses were carried out in R software using the 'corrplot' package. PCA analyses were performed using the online OECloud tools.[4] The hierarchical cluster analyses were performed using the R package 'pheatmap' with the "scale = row" parameter. Volcano plots were constructed using the online OmicShare platform.[5]

Identified metabolites were annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Compound database[6] and mapped to the KEGG Pathway database.[7] KEGG pathway enrichment analysis of DAVs was performed using Metabolite Set Enrichment Analysis (MSEA). The Gene Ontology (GO) and KEGG pathway enrichment analyses of DEGs were performed using R packages with hypergeometric distribution tests (Kanehisa et al., 2007). The DEGs and DAVs with Pearson's correlation coefficient >0.8 and $p < 0.05$ were selected to construct the transcript-metabolite network, which was visualized by Cytoscape (version 3.7.1).

Transcription factors (TFs) were identified by searching against the Plant Transcription Factor Database (PlantTFDB 4.0),[8] and the best hits with an E value less than $1e^{-5}$ were labeled as TFs. The specific target genes in Arabidopsis regulated by the TFs were predicted using the Gene Transcription Regulation Database (GTRD).[9] The homologs of the predicted Arabidopsis target genes in *L. japonica* were acquired by sequence BLAST. To construct the TF regulatory network, the protein–protein interactions between TF target genes and the floral scent-related DEGs were analyzed using the STRING database (confidence score > 0.7).[10] The TF regulatory network was visualized by Cytoscape (version 3.7.1).

Genome-wide identifications and phylogenetic analyses of the genes involved in pivotal VOCs production were performed. The Hidden Markov Models (HMMs) corresponding to the gene conserved domains were downloaded from the Pfam database[11] and used as the query to identify the gene family from the *L. japonica* protein database[12] using HMMER software. All the putative sequences were further confirmed to have complete conserved domains by the NCBI-CDD database.[13] Phylogenetic analyses of the identified gene family members were performed using MEGAX 11.0 software with the maximum likelihood method.

qRT-PCR analysis

The sequences of candidate genes were retrieved from the *L. japonica* genome database. Specific primer pairs used in qRT-PCR were designed using NCBI Primer-Blast[14] (Supplementary Table 1). Total RNA of LJFs at the silver stage of 'Yujin2' and 'Fengjin1' was isolated using the OminiPlant RNA Kit (Cwbiotech, Beijing, China) and then reverse-transcribed using a HiScript® II Q RT SuperMix for qPCR kit (+ gDNA wiper; Vazyme Biotech, Nanjing, China). qRT-PCR was performed on a LightCycler® 96 real-time PCR system (Roche, Hong Kong, China) using SYBR Green Master Mix (Vazyme Biotech, Nanjing, China; Yu et al., 2021). Three biological and three technical replicates were performed. Relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method (Liu et al., 2019). The *Lonja.ACT2/7* and *Lonja.27738* genes were used as internal references (Liu, 2017).

Results

Metabolome analysis of the LJFs of the two varieties of *Lonicera japonica*

To understand what might contribute to the differences in the aroma of the LJFs of these two varieties, HS-SPME coupled with GC–MS was used to identify the VOCs of LJF at the silver stage (the stage with the strongest aroma) of 'Yujin2' and 'Fengjin1' (Figure 1). Overlapping analysis of the total ion current in different QC samples revealed the instrumental stability was good for GC–MS analysis (Figure 2A). Pearson's correlation analysis showed weak correlations between 'Yujin2' and 'Fengjin1' samples, and strong correlations within the replicate samples (Figure 2B). PCA also showed that the VOC profiles of the two varieties were separate, and the biological replicates of the same

3 https://bigd.big.ac.cn/search/?dbId=gwh&q=honeysuckle
4 https://cloud.oebiotech.cn/task/detail/pca/
5 https://www.omicshare.com/tools/Home/Soft/volcano
6 http://www.kegg.jp/kegg/compound/
7 http://www.kegg.jp/kegg/pathway.html
8 http://planttfdb.cbi.pku.edu.cn/index.php
9 http://gtrd.biouml.org
10 https://string-db.org/
11 http://pfam.xfam.org
12 https://bigd.big.ac.cn/search/?dbId=gwh&q=honeysuckle
13 https://www.ncbi.nlm.nih.gov/cdd
14 https://www.ncbi.nlm.nih.gov/tools/primer-blast/

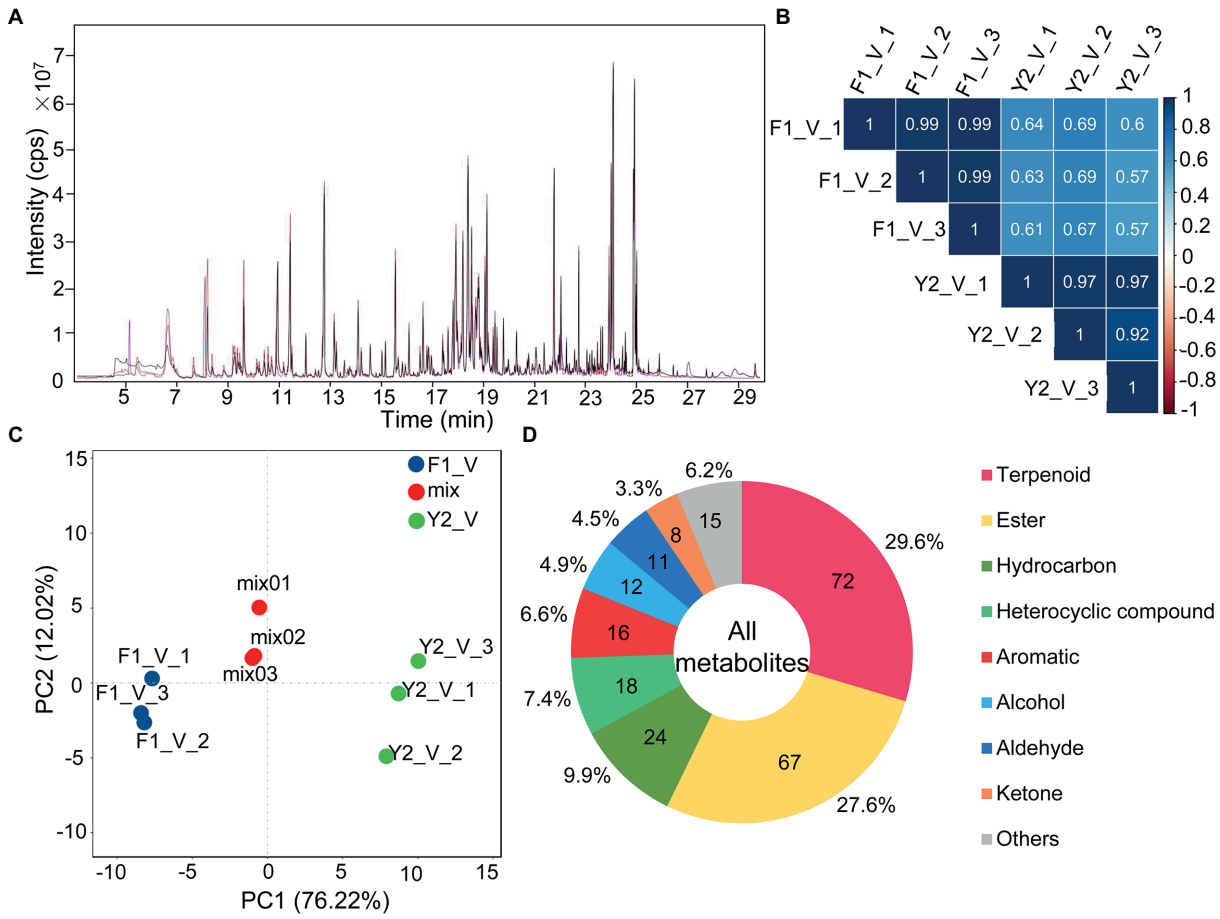


FIGURE 2
Overview of the metabolomics analysis. **(A)** Overlapping analysis of the total ion current in different quality control samples. The abscissa and the ordinate represent the retention time (min) and the intensity of the ion current (cps: count per second), respectively. **(B)** Pearson's correlation analysis of LJF VOC profiles. **(C)** Principal component analysis (PCA) of LJF VOC profiles. **(D)** Component analysis of the identified VOCs of LJFs.

variety were closely grouped (Figure 2C). These results suggest that the metabolome data were highly reproducible and reliable for further analyses, and the VOC profiles were different in the LJFs of the two varieties.

Despite distinct aroma characteristics between 'Yujin2' and 'Fengjin1', the identified metabolite species did not differ between the two varieties. In each of the variety, 243 VOCs were identified from LJFs (Supplementary Table 2). The detected VOCs were classified into different groups: 72 terpenoids (29.6%), 67 esters (27.6%), 24 hydrocarbons (9.9%), 18 heterocyclic compounds (7.4%), 16 aromatics (6.6%), 12 alcohols (4.9%), 11 aldehydes (4.5%), 8 ketones (3.3%), and 15 other compounds. Among these, terpenoids and esters accounted for 57.2% of the total VOCs (Figure 2D).

Interestingly, the levels of the VOCs from LJFs between the two contrasting varieties were quite different. A total of 153 DAVs were identified, of which 41 were increased and 112 were decreased in 'Yujin2' compared to 'Fengjin1' (Figure 3A; Supplementary Table 3). Most of the DAVs induced in 'Yujin2' were terpenoids and esters. However, the DAVs such as hydrocarbons, aldehydes, and halogenated hydrocarbons were decreased in 'Yujin2' (Figure 3B).

Notably, among these DAVs, 22 odorants were increased in 'Yujin2' compared to 'Fengjin1' (Table 1). These odorants have sweet, aromatic, floral, fruity, and other odor qualities. Most noteworthy, methyl anthranilate with a grape-like and orange blossom odor was increased 154.1-fold in 'Yujin2'. Methyl anthranilate was ranked 15th among the identified VOCs in 'Yujin2' (Supplementary Table 2). The particularly high contents and extremely large increases of methyl anthranilate imply its crucial role for the strong aroma of 'Yujin2'. Additionally, the accumulations of farnesol (delicate flowery odor), farnesal (floral minty aroma), trans-alpha-bergamotene (warm tea odor), squalene (faint agreeable odor), and ethyl cinnamate (sweet fruit odor) had more than 6-fold increases (Table 1). Among the increased odorants in 'Yujin2', in addition to methyl anthranilate, farnesol, alpha-farnesene (floral balsamic aroma), trans-alpha-bergamotene, and geraniol (sweet rose odor) were detected to be the top 20 ranking VOCs in 'Yujin2'. The increased and high accumulation of these odorants may be the likely cause of the stronger aroma of LJFs in 'Yujin2' compared to that of 'Fengjin1'.

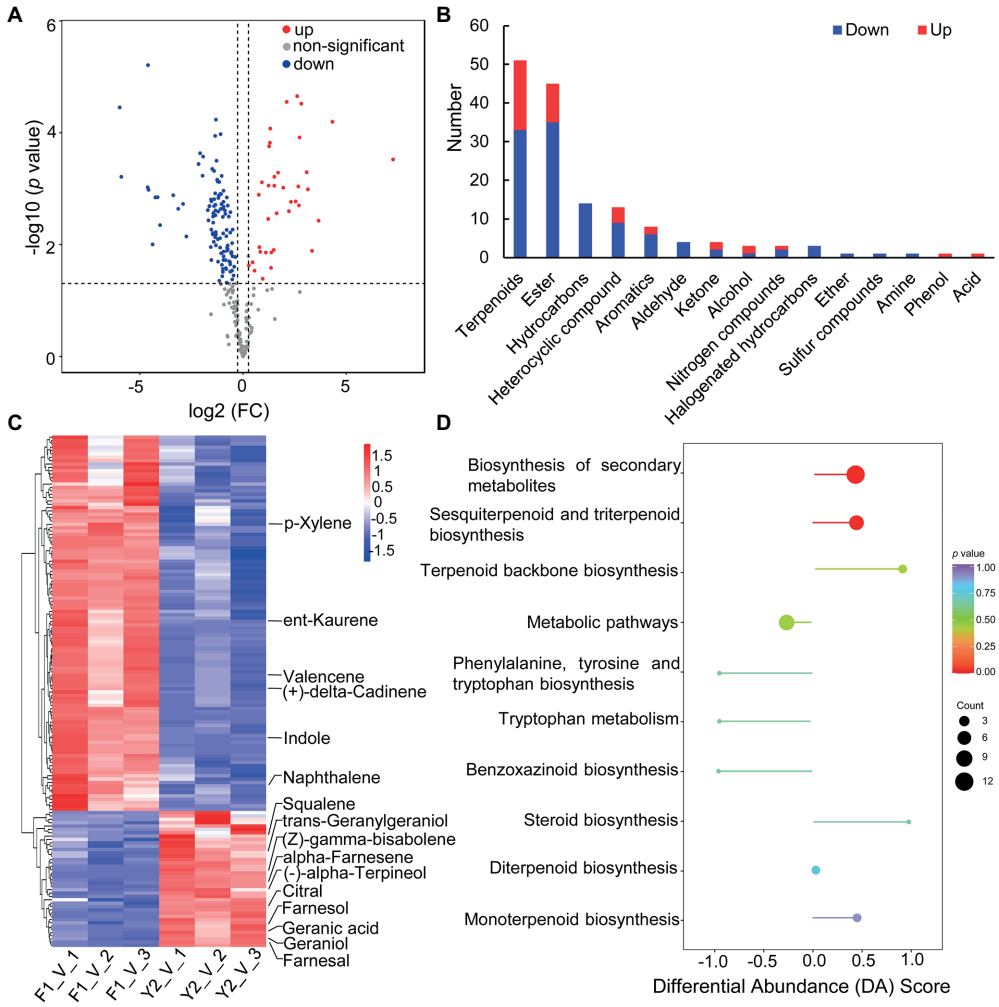


FIGURE 3
Characterization of differentially accumulated volatiles (DAVs) between 'Yujin2' and 'Fengjin1' LJF samples. **(A)** Volcano plot of the identified VOCs. The increased and decreased DAVs are shown in red and blue, respectively. **(B)** Abundance patterns of DAVs in each class. **(C)** Hierarchical clustering analysis of DAVs. The columns represent different samples, and the rows represent individual DAVs. The colors in the heat map represent the normalized values of DAVs, reflecting the relative contents. DAVs with high or low levels are indicated in red or blue, respectively. Detailed information is provided in Supplementary Table 4. The key DAVs that were functionally annotated through the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis are highlighted on the right. **(D)** KEGG enrichment analysis of the DAVs. The ordinate represents the enriched pathway name. The abscissa represents the differential abundance score (DA score), which reflects the overall change in DAVs in the pathway. The length of the line segment represents the absolute value of the DA score, and the size of the dot at the endpoint of the line segment represents the number of DAVs involved in the pathway. For dots shown on the right panel, the longer the line segment, the more inclined the overall expression of the pathway to be upregulated. In contrast, for dots shown on the left panel, the longer the line segment, the more inclined the overall expression of the pathway to be downregulated. The color of the line segment and the dot reflects the *p* value.

To evaluate the accumulation patterns of the DAVs, hierarchical clustering analysis was performed, which revealed that the DAVs were grouped into two clusters (Figure 3C; Supplementary Table 4). Moreover, the DAVs were functionally annotated using the KEGG database to gain further insights into the mechanisms of the biosynthesis of DAVs (Figure 3D, Supplementary Table 5). The 'Yujin2'-induced DAVs were predominantly enriched in the biosynthesis of secondary metabolites, especially terpenoid biosynthesis that included terpenoid backbone biosynthesis, monoterpenoid biosynthesis (alpha-terpineol and geraniol), diterpenoid biosynthesis (geranylgeraniol), and sesquiterpenoid and triterpenoid biosynthesis (farnesol, farnesal, alpha-farnesene, (Z)-gamma-bisabolene, and squalene). In contrast, indoles that were involved in phenylalanine, tyrosine, and tryptophan biosynthesis, as well as benzoxazinoid biosynthesis, were decreased in 'Yujin2' (Figure 3D).

Transcriptome analysis of the LJFs of the two varieties of *Lonicera japonica*

In an attempt to obtain insights into the biosynthesis of the floral VOCs of the two varieties, transcriptomic analysis of the LJFs at the silver stage was performed (Figure 1). An average of 42

TABLE 1 Odorants that were differentially accumulated in 'Yujin2' compared to 'Fengjin1'.

Index	Common name	Class I	CAS	Fold change (Y2 vs. F1)	Odor quality
KMW0503	Methyl anthranilate	Ester	134-20-3	154.05	Grape-like odor; Orange blossom odor ★
NMW0310	Farnesol	Terpenoids	4602-84-0	12.64	Delicate flowery odor; Mild, oily; Weak citrus-lime odor ★
w05	Farnesal	Terpenoids	19317-114	8.48	Floral minty
KMW0555	trans-alpha-Bergamotene	Terpenoids	13474-59-4	7.09	Woody, warm tea
NMW0652	Squalene	Terpenoids	111-02-4	6.74	Faint agreeable odor ★
KMW0574	Ethyl cinnamate	Ester	103-36-6	6.67	Sweet, balsam, fruity, spicy, powdery, berry, plum
NMW0149	Geranic acid	Terpenoids	459-80-3	6.54	Dry, weedy, acidic, green, moldy feet, woody.
KMW0630	Nerolidol	Terpenoids	40716-66-3	5.86	A floral odor ★
KMW0514	Ethyl 3-phenylpropionate	Ester	2021-28-5	4.33	Hyacinth, rose, honey, fruity, rum
KMW0461	Nonanoic acid	Acid	112-05-0	2.86	Fatty odor; Coconut aroma; Slight odor ★
KMW0460	Geraniol	Terpenoids	106-24-1	2.85	A sweet rose odor; Pleasant geranium-like odor; Pleasant, floral odor ★
KMW0431	5-Methyl-6,7-dihydro-5H-cyclopenta[b]pyrazine	Heterocyclic compound	23747-48-0	2.56	Earthy, baked, potato, peanut, roasted
KMW0459	Citral	Terpenoids	141-27-5	2.49	Strong lemon odor ★
KMW0613	alpha-Farnesene	Terpenoids	502-61-4	2.46	A floral, green, and balsamic aroma ★
XMW0561	Farnesol, acetate	Ester	1000352-67-2	2.33	Delicate flowery odor; Mild, oily; Weak citrus-lime odor ★
KMW0716	Isopropyl palmitate	Ester	142-91-6	2.15	Almost odorless ★
XMW1084	cis-3-Hexenyl crotonate	Ester	65405-80-3	1.93	Green vegetable
XMW1486	guaiyl acetate	Ester	134-28-1	1.88	Tea, rose, woody, spicy, green fatty
NMW0071	(−)-alpha-Terpineol	Terpenoids	10482-56-1	1.73	Lilac floral terpenic
KMW0304	p-Mentha-1,3,8-triene	Terpenoids	18368-95-1	1.70	Turpentine camphor, herbal, woody
WMW0127	cis-3-Octen-1-ol	Alcohol	20125-84-2	1.44	Fresh fatty greasy, melon green cortex, herbal earthy fusel spicy
WMW0049	Nerylacetone	Ketone	3879-26-3	1.21	Fatty metallic

The information of the odor quality was obtained preferentially from the PubChem database (marked with pentagrams), and then from the Good Scent Company website.

million clean reads per sample were aligned uniquely to the *L. japonica* genome (Supplementary Tables 6). PCA and Pearson's correlation analysis revealed that the transcriptome data of LJF were reproducible of the same variety and different between the two varieties, which indicate the transcriptome data were reliable for further analyses (Figures 4A, B).

A total of 25,649 and 25,943 expressed genes were quantified in the LJF samples from 'Yujin2' and 'Fengjin1', respectively (Supplementary Table 7). A total of 9,523 DEGs were identified, of which 4,715 genes were upregulated and 4,808 genes were downregulated in 'Yujin2' compared to 'Fengjin1' (Figure 4C; Supplementary Table 8). Hierarchical clustering analysis of the DEGs was performed, and the DEGs were grouped into two clusters (Figure 4D; Supplementary Table 9).

Based on this clustering, the DEGs were functionally annotated using the GO and KEGG enrichment analyses. According to the biological process, the upregulated DEGs were mainly involved in cell wall organization, pectin catabolic process, regulation of pollen tube growth, regulation of pH, carbohydrate metabolic process, etc. With regard to downregulated DEGs, most of them were enriched in the regulation of defense response (including the response to wounding, chitin, and karrikin), polar nucleus fusion, photosynthesis, glutathione metabolic process, and so on (Figure 4E; Supplementary Table 10). Moreover, KEGG pathway enrichment analysis revealed that the DEGs were assigned to 127 KEGG pathways (Supplementary Table 11). The most significantly enriched KEGG pathways are shown in Figure 4F. Notably, brassinosteroid biosynthesis, monoterpenoid biosynthesis, and carotenoid biosynthesis, which are related to the metabolism of terpenoids and polyketides, were significantly enriched (Figure 4F). In addition to these, other pathways related to terpenoids and polyketides were also enriched, including terpenoid backbone biosynthesis, sesquiterpenoid and triterpenoid biosynthesis, as well as diterpenoid biosynthesis. Furthermore, additional pathways associated with VOC biosynthesis were also enriched, including steroid biosynthesis, phenylalanine, tyrosine, and tryptophan biosynthesis, phenylpropanoid biosynthesis, as well as alpha-linolenic acid metabolism (Supplementary Table 12).

Notably, a total of 520 TFs distributed among 56 TF families were differentially expressed in 'Yujin2' compared to 'Fengjin1'. Among these, almost half of the differentially expressed TFs are classified into six families, i.e., APETALA2/ethylene-responsive factor (AP2/ERF-ERF), v-myb avian myeloblastosis viral oncogene homolog (MYB), NAC, WRKY, Basic helix–loop–helix (bHLH), and C2H2. Interestingly, most of the TFs in the GARP-G2-like family and the lateral organ boundary domain (LOB) family were upregulated in 'Yujin2', while the TFs in the WRKY family were mostly downregulated (Figure 4G; Supplementary Table 13). To

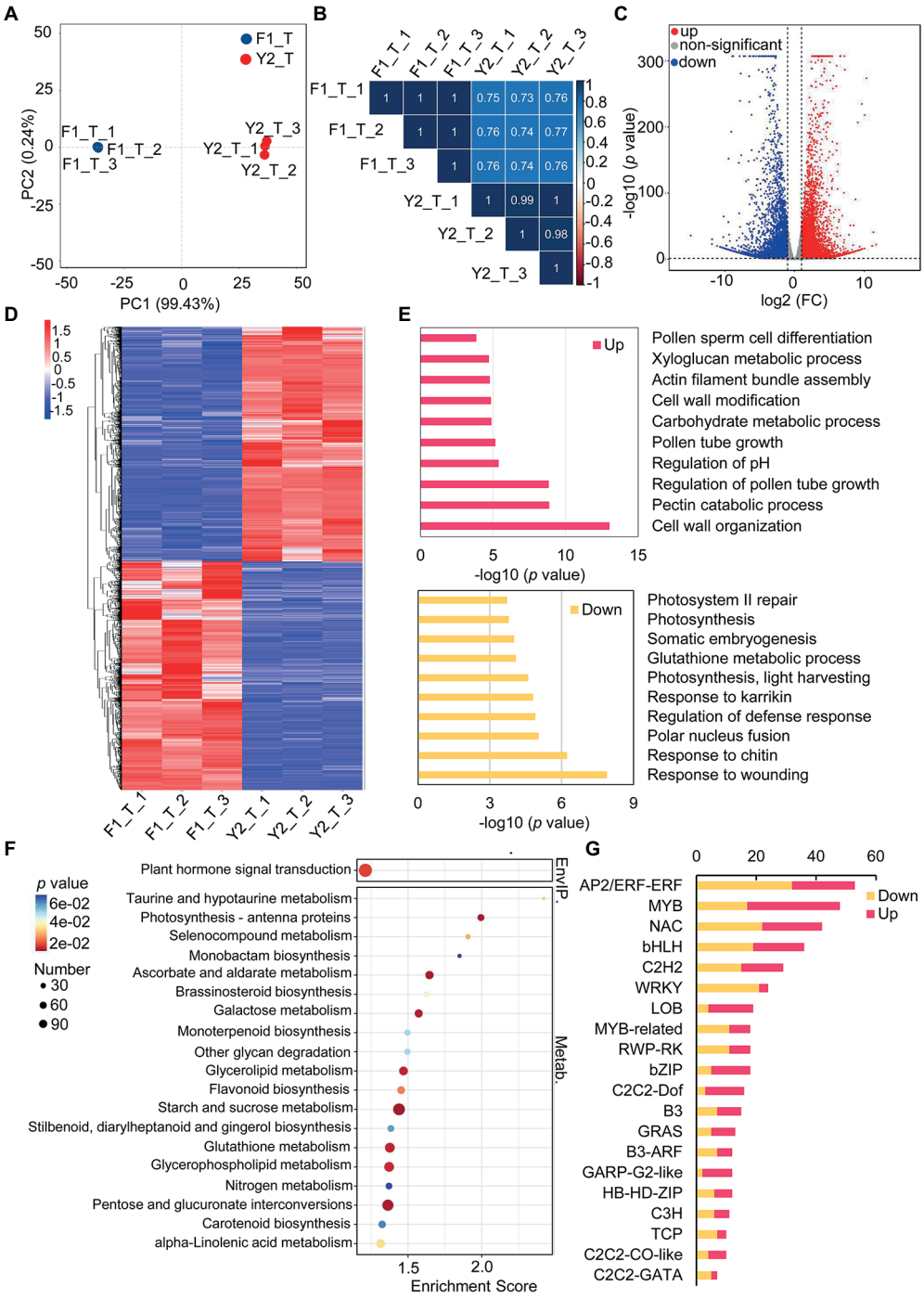


FIGURE 4
Characterization of differentially expressed genes (DEGs) between 'Yujin2' and 'Fengjin1' LJF samples. **(A)** PCA of the LJF transcriptome data. **(B)** Pearson's correlation analysis of LJF transcriptome data. **(C)** Volcano plot of the identified DEGs. The increased and decreased DEGs are shown in red and blue, respectively. **(D)** Hierarchical clustering analysis of DEGs. The columns represent different samples, and the rows represent individual DEGs. The colors in the heat map represent the normalized values of DEGs. DEGs with high or low levels are indicated in red or blue, respectively. Detailed information is provided in Supplementary Table 9. **(E)** Gene Ontology biological process enrichment of DEGs. Detailed information is presented in Supplementary Table 10. **(F)** KEGG enrichment of DEGs. Detailed information is provided in Supplementary Table 11. **(G)** Top 20 differentially expressed transcription factor genes. Detailed information is provided in Supplementary Table 13.

further investigate the TFs related to VOC production in floral scent of LJF, these differentially expressed TFs were used to construct a regulatory network of TFs and the floral scent-related genes (Figure 5, Supplementary Table 12 and 13). Six TF families, including MYB3, AP2, LFY, ARF6, WRKY33 and bZIP TFs, were predicted to regulate 40 target

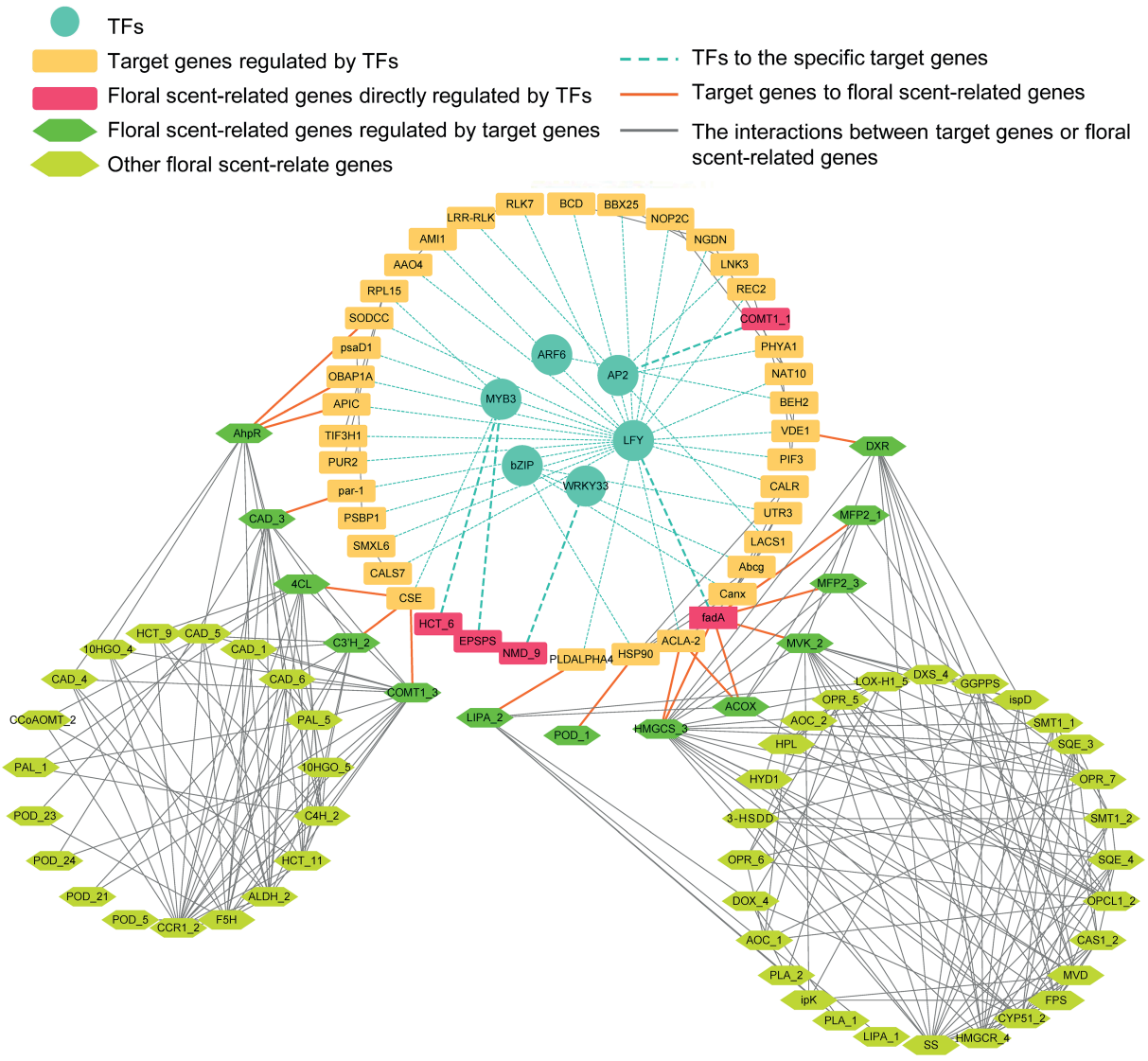


FIGURE 5
A regulatory network of transcription factors (TFs) and the floral scent-related genes. The circular, rectangular, hexagonal nodes indicate the TFs, target genes, and floral scent-related genes, respectively. The turquoise, orange, and gray edges indicate the TFs to the specific target genes predicted by GTRD, target genes to floral scent-related genes and the interactions between floral scent-related genes predicted by STRING. The raw data and the full names of the abbreviations can be found in Supplementary Table 15.

genes that were directly or indirectly associated with the floral scent-related genes. Among the 40 target genes, 25 were predicted to be regulated by LFY, which was increased 5.81-fold in 'Yujin2'. This indicates that LFY was an important TF in the regulatory network of floral scent in LJF. Among the LFY-regulated target genes, acetyl-CoA acyltransferase (fadA) that involved in alpha-linolenic acid metabolism was increased 3.3-fold in 'Yujin2' and interacted with several floral scent-related genes, such as mevalonate kinase (MVK) and hydroxymethylglutaryl-CoA synthase (HMGCS) that involved in terpenoid biosynthesis, as well as acyl-CoA oxidase (ACOX) and fatty acid beta-oxidation multifunctional protein (MFP2) that involved in alpha-linolenic acid metabolism. This suggests the important role of fadA in the regulatory network. In addition, a caffeate O-methyl transferase (COMT1) and a hydroxycinnamoyl transferase (HCT), which were upregulated in 'Yujin2', were predicted to be regulated by AP2 and MYB3, respectively. A (+)-neomenthol dehydrogenase (NMD) and a 5-enolpyruvylshikimate 3-phosphate synthase (EPSPS), which were downregulated in 'Yujin2', were predicted to be regulated by WRKY33 and MYB3, respectively. Moreover, there are a number of genes related to floral scent that were predicted to be directly or indirectly regulated by the TF-target genes, suggesting they are likely to be indirectly regulated by the TFs.

10.3389/fpls.2022.989036

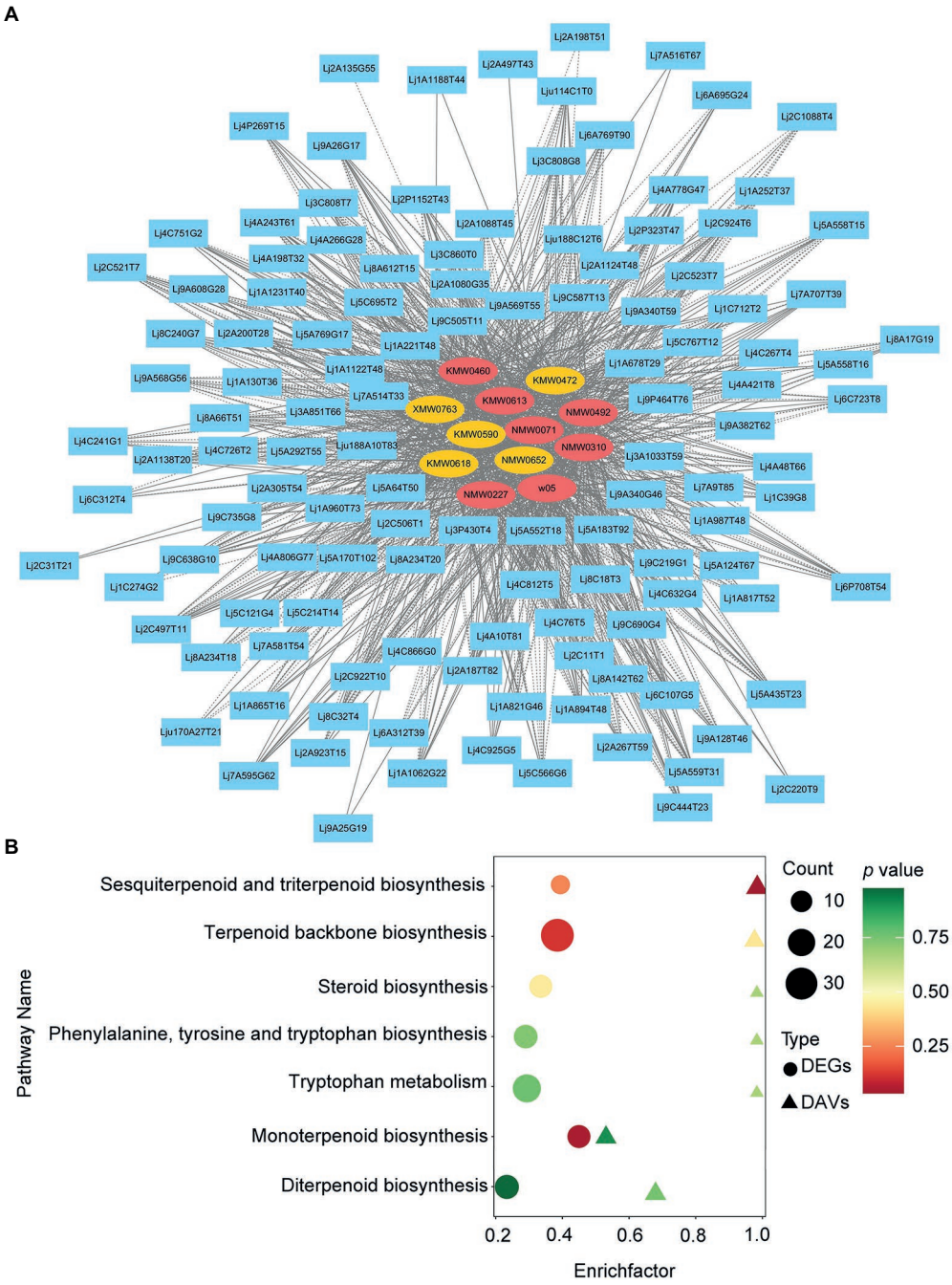


FIGURE 6
Integrated volatile metabolome and transcriptome analysis of VOC accumulation in LJFs. **(A)** Gene-metabolite correlation network representing DEGs and DAVs. The gene-metabolite pairs are connected by edges. Blue rectangular nodes represent genes. Red and yellow elliptical nodes represent increased and decreased metabolites, respectively. Solid and dashed edges represent positive and negative correlations, respectively. **(B)** Bubble map of KEGG pathways co-enriched from metabolome and transcriptome data. The ordinate represents KEGG terms, and the abscissa represents the enrichment factor of each term.

Integrated volatile metabolome and transcriptome analysis of VOC accumulation in LJFs

To generate the candidate genes likely to be involved in LJF VOC biosynthesis, we integrated the transcriptome data and the VOC profiling data by correlation analysis (Figure 6A; Supplementary Table 14). In total, 12 DAVs showed higher corrections with DEGs, which included geraniol, L-alpha-terpineol, alpha-farnesene, and trans-geranylgeraniol, etc. Based on KEGG analysis, the DEGs and DAVs between 'Yujin2' and 'Fengjin1' were integrated into several pathways related to VOC

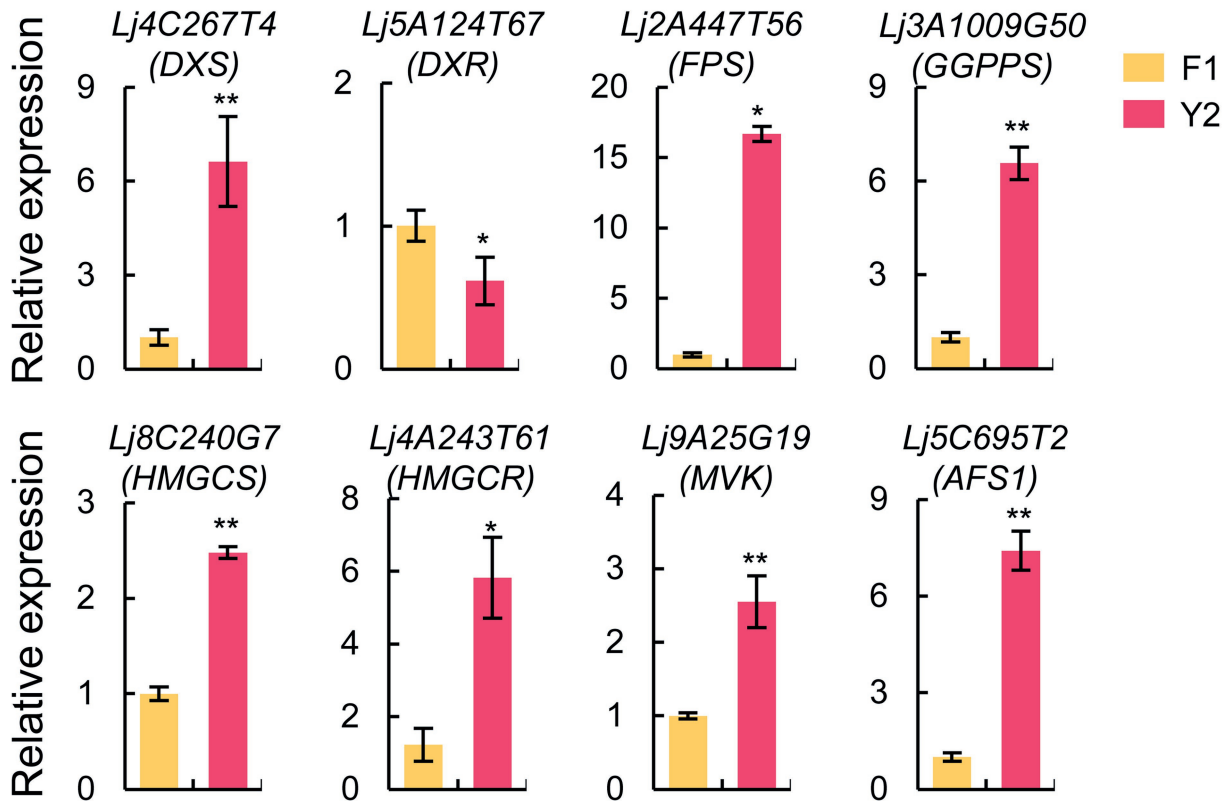


FIGURE 7
Expression analysis of candidate genes involved in the terpenoid biosynthetic pathway based on quantitative real-time PCR (qRT-PCR). The actin gene was used as an internal reference. The values are presented as the mean±standard deviation (n=3). Significant differences between 'Yujin2' and 'Fengjin1' are marked with asterisks (**p<0.01, *p<0.05, Student's t-test). AFS1, alpha-farnesene synthase; DXR, 1-deoxy-D-xylulose-5-phosphate reductoisomerase; DXS, 1-deoxy-D-xylulose-5-phosphate synthase; FPS, farnesyl diphosphate synthase; GGPPS, geranylgeranyl diphosphate synthase; HMGCR, hydroxymethylglutaryl-CoA reductase; HMGCS, hydroxymethylglutaryl-CoA synthase; MVK, mevalonate kinase.

biosynthesis, including sesquiterpenoid and triterpenoid biosynthesis, terpenoid backbone biosynthesis, steroid biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, tryptophan metabolism, monoterpenoid biosynthesis, and diterpenoid biosynthesis (Figure 6B).

Genome-wide identification and phylogenetic analyses of the genes involved in pivotal VOCs production in *Lonicera japonica*

To gain further insight into the evolutionary relationships and expression patterns among the subfamilies of the genes involved in pivotal VOCs production, phylogenetic trees of the gene families were constructed (Supplementary Figure 1). The phylogenetic analyses of multiple gene family members that were differentially regulated in 'Yujin2' revealed the gene expression of the members associated with the evolution of the gene family, including TPS, TIDS, 1-deoxy-D-xylulose-5-phosphate synthase (DXS), HMGCS, hydroxymethylglutaryl-CoA reductase (HMGCR), MVK, squalene epoxidase 1 (SQE1), and LUS. For instance, three DXS members of the clade I were upregulated, while another DXS member of the clade II was downregulated. Similar change patterns in different clades of the subfamilies of the gene family were also observed in HMGCS, HMGCR, MVK, LUS, and SQE1. The TPS gene family was divided into seven subfamilies a-g as previously reported (Supplementary Figure 2; Jiang et al., 2019). The differentially regulated family members of alpha-farnesene synthase 1 (AFS1) and TES both belong to TPS-b subfamilies but in different clades. These results suggest the members of these gene families play different roles in terpenoid biosynthesis.

Characterization of terpenoid biosynthesis-related genes in LJFs

The expression patterns of the candidate genes involved in terpenoid biosynthesis were verified using qRT-PCR (Figure 7; Supplementary Figure 3). The results showed that seven genes were upregulated, while one gene was downregulated in 'Yujin2' compared to 'Fengjin1', similar to the trends revealed by transcriptomic analysis. These verified genes included DXS,

1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), and geranylgeranyl diphosphate synthase (GGPPS), involved in the 2-c-methylerythritol 4-phosphate (MEP) pathway, and four genes involved in the mevalonic-acid (MVA) pathway, including HMGCS, HMGCR, MVK, and AFS1.

Discussion

The regulatory mechanisms of floral scents are sophisticated and complex. Integrated metabolomic and transcriptomic analysis allows for representation of gene-to-metabolite networks to decipher the mechanisms involved in the regulation of LJF floral scents. In the current study, the metabolomic and transcriptomic analyses of LJFs at the silver flowering stage of 'Yujin2' and 'Fengjin1' indicated that the biosynthesis of terpenoids and amino acid and fatty acid derivatives was pivotal for the stronger aroma in 'Yujin2' compared to 'Fengjin1'.

Differential accumulation of key odorants contributes to the stronger aroma of LJFs in 'Yujin2'

Previous studies have found that the primary VOCs of LJFs are monoterpenes, sesquiterpenes, fatty acids, and their esters (Wang et al., 2009; Ilie et al., 2014). With very few exceptions, the VOCs of LJFs reported in previous studies were also detected in this study (Supplementary Table 2; Wang et al., 2009; Ilie et al., 2014; Du et al., 2015; Lin et al., 2020; Yan et al., 2020). Floral scents tend to be mixtures of many compounds, but there are always major compounds that contribute the most significantly. In line with previous studies, the characteristic aromatic compounds of LJFs were also detected at high levels in the two varieties in this study, such as terpenoids that included linalool, farnesol, alpha-farnesene, delta-cadinene, trans-alpha-bergamotene, and geraniol, as well as esters that included methyl anthranilate and geranyl isovalerate (Supplementary Table 2; Wang et al., 2009; Ilie et al., 2014; Du et al., 2015; Lin et al., 2020; Yan et al., 2020). Among these abundant VOCs, linalool and geranyl isovalerate exhibited no significant difference between 'Yujin2' and 'Fengjin1'. Interestingly, the abundant VOCs that included methyl anthranilate, farnesol, alpha-farnesene, trans-alpha-bergamotene, and geraniol, as well as other VOCs that included squalene, nerolidol, hedycaryol, nonanoic acid, citral, isopropyl palmitate, and alpha-terpineol, accumulated more in 'Yujin2' compared to 'Fengjin1' (Table 1). These increased odorants, which have sweet, pleasant, floral, and fruity odor qualities, are likely to be the reasons for the stronger aroma of LJFs in 'Yujin2' compared to 'Fengjin1'. Notably, methyl anthranilate, a grape scent and flavor compound, can impart a pleasant aroma (Luo et al., 2019). In consideration of its most significant increase and high accumulation, methyl anthranilate would likely be the greatest cause of the stronger aroma in 'Yujin2' compared to 'Fengjin1'.

It is important for plants to emit floral scents in order to attract pollinators, deter pathogens, and respond to biotic and abiotic stressors (Schiestl, 2015). Linalool and E-β-ocimene are reported as being attractive to pollinators including bees and moths to enhance propagation (Parachnowitsch et al., 2012; Gerard et al., 2017). No significant difference in the accumulation of linalool and E-β-ocimene between 'Yujin2' and 'Fengjin1' suggested that the ability of these two varieties to attract pollinators was not significantly different, despite differences in the LJF aroma (Supplementary Table 2). (E)-α-bergamotene and particular homoterpenes or terpene derivatives help to protect plants from microbial pathogens and abiotic stresses by recruiting the enemies of pests, and (E, E)-α-farnesene has been reported to influence multiple interactions between plants and other organisms when exposed to abiotic stresses (Zhou et al., 2017). The significantly increased α-farnesene and α-bergamotene in 'Yujin2' may contribute to its high resistance to pathogens, as reported previously (Li et al., 2016).

Terpenoid biosynthetic pathway was promoted in 'Yujin2' leading to the accumulation of terpenoids

Terpenoids are the most dominant and diverse group of floral VOCs (Muhlemann et al., 2014). In the current study, terpenoids were the largest group identified in JLFs (Figure 2D). In higher plants, terpenoids are derived from the C5 carbon precursors isopentenyl diphosphate (IPP) or dimethylallyl diphosphate (DMAPP) through two alternative pathways, the plastid-localized MEP pathway and the cytosol-localized MVA pathway in flowers and other plant organs (Figure 8; Dudareva et al., 2013).

Monoterpenes and diterpenes represent important classes of aromatic compounds responsible for floral scents (Qiao et al., 2021). Increased accumulations of geraniol (sweet rose odor) and alpha-terpineol (lilac floral terpenic odor), i.e., monoterpenes, and geranylgeraniol, i.e., a diterpene, in 'Yujin2', may contribute to the stronger aroma (Figure 8). Monoterpenes and diterpenes are generated by the plastid-localized MEP pathway. Beginning with pyruvate and glyceraldehyde-3-phosphate, DXS, DXR, and 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (ispD) catalyze the first three reactions of the MEP pathway (Qiao et al., 2021). In this study, we found three DXS genes that were upregulated, while a DXS gene, a DXR, and an ispD were downregulated in 'Yujin2' (Figure 8; Supplementary Table 12). These results suggested that the biosynthesis of the C5 carbon precursors IPP and DMAPP *via* the MEP pathway may be inhibited in 'Yujin2'. However, GGPPS, which primarily utilizes the products of the MEP pathway to condense three IPP molecules and one DMAPP molecule into geranylgeranyl diphosphate (GGPP), was upregulated in 'Yujin2' (Figure 8; Supplementary Table 12). GGPP serves as the entry point leading to the biosynthesis of a diverse group of primary and secondary

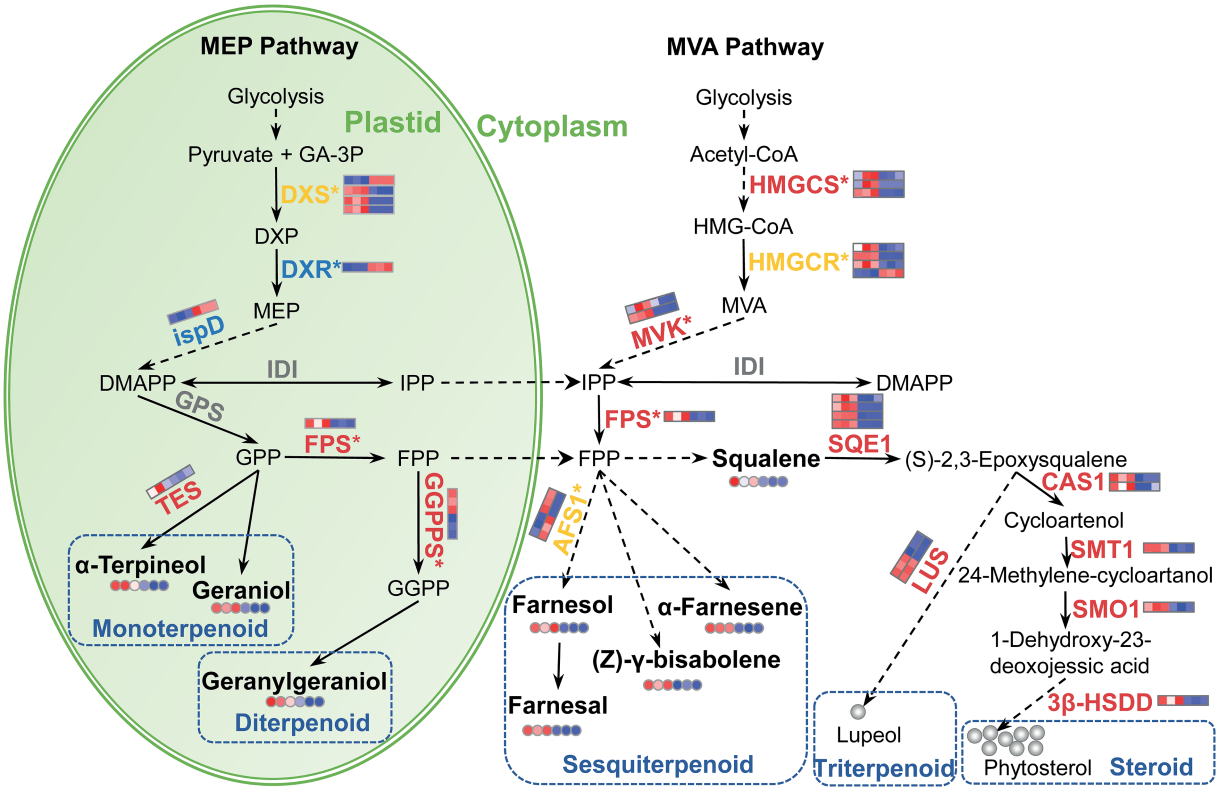


FIGURE 8
Schematic presentation of terpenoid biosynthetic pathway in LJFs of 'Yujin2' compared to 'Fengjin1'. The upregulated, downregulated, mix-regulated, and unchanged genes (bold fonts) and metabolites (small circles) in 'Yujin2' are represented in red, blue, yellow, and gray, respectively. The scale bars adjacent to the name of genes or metabolites are picked from the heatmaps for the transcriptome (Figure 4D) and metabolome (Figure 3C). The first three and the latter three squares of a bar represent the normalized abundance values of DEGs or DAVs from 'Yujin2' and 'Fengjin1', respectively. Red and blue colors reflect high and low expression levels, respectively. The asterisk followed by the gene name indicate that the gene expression was validated by qRT-PCR. The solid line, dashed line, and dotted line indicate a single-step reaction, a multi-step reaction, and the movement of substances, respectively. 3β-HSDD, plant 3β-hydroxysteroid-4α-carboxylate 3-dehydrogenase; Acetyl-CoA, acetoacetyl-CoA; AFS1, alpha-farnesene synthase; CAS1, cycloartenol synthase; DMAPP, dimethylallyl pyrophosphate; DXP, 1-deoxy-d-xylulose 5-phosphate; DXR, DXP reductoisomerase; DXS, DXP synthase; FPP, farnesyl diphosphate; FPS, farnesyl diphosphate synthase; GA-3P, D-glyceraldehyde 3-phosphate; GGPP, geranylgeranyl diphosphate (C20); GGPPS, GGPP synthase; GPP, geranyl diphosphate; GPS, geranyl diphosphate synthase; HMG-CoA, 3-hydroxy-3-methylglutaryl-CoA; HMGR, HMG-CoA reductase; HMGS, HMG-CoA synthase; IDI, IPP isomerase; IPP, isopentenyl diphosphate; LUS, lupeol synthase; MEP, methylerythritol phosphate; MVA, mevalonic acid; MVK, mevalonate kinase; SMO1, plant 4,4-dimethylsterol C-4alpha-methyl-monooxygenase; SMT1, sterol 24-C-methyltransferase; SQE1, squalene monooxygenase; TPS, alpha-terpineol synthase.

terpenoid compounds, including monoterpenes, diterpenoids, and plastid-derived carotenoids and their derivatives (Nagegowda and Gupta, 2020). It has been widely reported that GGPPS contributes to the production of monoterpenes in flowering plants, such as *Arabidopsis thaliana* (Chen et al., 2015), *Chimonanthus praecox* (Shang et al., 2020), *Thymus caespititius* (Adal and Mahmoud, 2020), *Antirrhinum majus*, and *Clarkia breweri* (Tholl et al., 2004). Additionally, we found that several genes involved in monoterpenoid biosynthesis were differentially regulated in 'Yujin2', including increased alpha-terpineol synthase (TES) and mix-regulated 8-hydroxygeraniol dehydrogenase (10HGO) and NMD (Figure 8; Supplementary Table 12). TES catalyzes the conversion of geranyl diphosphate (GPP) to alpha-terpineol (Lima et al., 2013). Our results indicated that the upregulation of GGPPS and TES plays crucial roles in increasing the accumulation of geraniol, alpha-terpineol, and geranylgeraniol in 'Yujin2'.

Sesquiterpenoids have also been identified as important components of floral scents. In this study, three odorants classified as sesquiterpenoids increased significantly in 'Yujin2', including farnesol (weak citrus-lime odor), farnesal (white-lemon-like aroma), and alpha-farnesene (floral, green, and balsamic aroma; Table 1; Supplementary Table 3). Sesquiterpenoids are synthesized *via* the MVA pathway in the cytosol. In the MVA pathway, three genes of HMGCS, three genes of HMGCR, as well as two genes of MVK, which catalyze reactions to produce IPP and DMAPP from the condensation of acetyl-CoA, were upregulated significantly in 'Yujin2' (Figure 8; Supplementary Table 12). This implied that the MVA pathway producing the C5 carbon precursors

IPP and DMAPP in the cytosol was promoted in 'Yujin2'. Consistently, farnesyl pyrophosphate synthase (FPS) was also upregulated. FPS condenses two IPP molecules and one DMAPP molecule to produce farnesyl diphosphate, which can be converted to sesquiterpenoids catalyzed by cytosolic sesquiterpene synthases (Chen et al., 2011). The significantly upregulated HMGCS, HMGCR, MVK, and FPS in 'Yujin2' were closely related to the increases in sesquiterpenoids responsible for the strong aroma.

Triterpenoids are often reported as components of the floral scent in LJFs (Wang et al., 2016; Li et al., 2020). Triterpenoids are biosynthesized *via* different cyclization reactions of squalene. In the current study, squalene with faint agreeable odor was increased significantly in 'Yujin2' (Table 1). In agreement with this, we also found four genes of SQE1, which oxidizes squalene to squalene 2,3-epoxide, and two genes of lupeol synthase (LUS), which converts oxidosqualene to other triterpene alcohols, that were significantly upregulated in 'Yujin2' (Figure 8; Supplementary Table 12). In addition, squalene is a precursor for the synthesis of plant sterols. We also found that four genes involved in catalyzing squalene 2,3-epoxide to form phytosterols were significantly upregulated, including two genes of cycloartenol synthase (CAS1), sterol 24-C-methyltransferase (SMT1), 4,4-dimethylsterol C-4alpha-methyl-monooxygenase (SMO1), and plant 3-beta-hydroxysteroid-4-alpha-carboxylate 3-dehydrogenase (3β-HSDD; Supplementary Table 12). Our results revealed that the biosynthesis of triterpenoids and phytosterols was induced in 'Yujin2'.

Phenylpropanoids and benzenoids did not contribute to the stronger aroma of 'Yujin2'

Phenylpropanoids and benzenoids are the second most ubiquitous class of plant VOCs (Knudsen and Gershenzon, 2006). Most VOCs from this class are derived from L-phenylalanine *via* the shikimate pathway. The first step in the biosynthesis of most phenylpropanoids and benzenoids is the deamination of L-Phe to trans-cinnamic acid (CA) by the enzyme L-Phe ammonia-lyase (PAL; Dudareva and Pichersky, 2006). Among the identified PAL genes, three were upregulated and two were downregulated in 'Yujin2' (Figure 9; Supplementary Table 12). Several genes involved in phenylpropanoid biosynthesis were upregulated in 'Yujin2', including 4-coumarate-CoA ligase-like 1 (4CL), p-coumaroyl quinate/shikimate 3′-hydroxylase (C3′H), and caffeoyl-CoA O-methyltransferase (CCoAOM). There were also a large number of genes related to phenylpropanoid biosynthesis that were downregulated or mix-regulated in 'Yujin2'. However, a series of enzymes required for biosynthesis of the volatiles methyleugenol and methyl isoeugenol were not differentially expressed between 'Yujin2' and 'Fengjin1', including coniferyl alcohol acyltransferase that produces coniferyl acetate, eugenol synthase and isoeugenol synthase that convert coniferyl acetate to eugenol and isoeugenol, respectively, as well as isoeugenol O-methyltransferase required for methylation to produce the volatiles methyleugenol and methyl isoeugenol (Dudareva et al., 2013; Mostafa et al., 2022). Consistent with the lack of significant expression differences of methyleugenol and methyl isoeugenol-related genes, differences in volatile phenylpropanoid accumulation between 'Yujin2' and 'Fengjin1' were not detected in the volatile profiling of LJFs (Figure 9; Supplementary Table 3). These results suggest that the differentially regulated genes involved in the phenylpropanoid biosynthesis pathway mainly contribute to the biosynthesis of non-volatile metabolites such as lignin in LJFs, rather than the biosynthesis of volatile phenylpropanoids (Peled-Zehavi et al., 2015). Our results also indicate that the differences in the floral scents between 'Yujin2' and 'Fengjin1' have a weak relationship with volatile phenylpropanoid accumulation.

Notably, all of the 11 detected volatile benzenoids were decreased in 'Yujin2', i.e., benzaldehyde, eight ester derivatives of benzoic acid (benzyl salicylate, benzyl tiglate, hexyl benzoate, methyl benzoate, ethyl salicylate, benzyl benzoate, 3-hexen-1-ol benzoate, and 3-hexen-1-ol, benzoate, (z)-), as well as two other benzene-substituted derivatives—benzyl chloride and ethyl phenylacetate (Supplementary Table 3). Benzenoids are synthesized from phenylalanine *via* the benzenoid branch of the phenylpropanoid pathway. Consistent with the decreased volatile benzenoids, two genes of BAHD acyltransferase and two genes of SABATH methyltransferase, involved in the final steps of benzenoid volatile formation, were significantly downregulated in 'Yujin2' (Figure 9; Supplementary Table 12; D'Auria et al., 2003; D'Auria, 2006). These results indicated that the biosynthesis and accumulation of benzenoids were suppressed, which were assumed to not contribute to the stronger aroma of 'Yujin2'.

Tryptophan biosynthetic pathway was elevated in 'Yujin2'

In addition to phenylalanine acting as a precursor for phenylpropanoids and benzenoids, two other aromatic amino acids (tyrosine and tryptophan) are precursors for various plant aromatic secondary metabolites (Tzin and Galili, 2010). In the biosynthesis of these three aromatic amino acids, chorismate is the common intermediate. Specifically, a series of genes involved in producing tryptophan from chorismate were upregulated in 'Yujin2', including the anthranilate synthase alpha subunit (ASα) and beta subunit (ASβ), anthranilate phosphoribosyltransferase (AnPRT), N-(5′-phosphoribosyl) anthranilate isomerase 1 (PAI1), and tryptophan synthase alpha chain (TSα; Figure 9; Supplementary Table 12; Tzin and Galili, 2010; Mostafa et al., 2022). However, the volatile indole, an important intermediate during tryptophan biosynthesis, was decreased significantly in

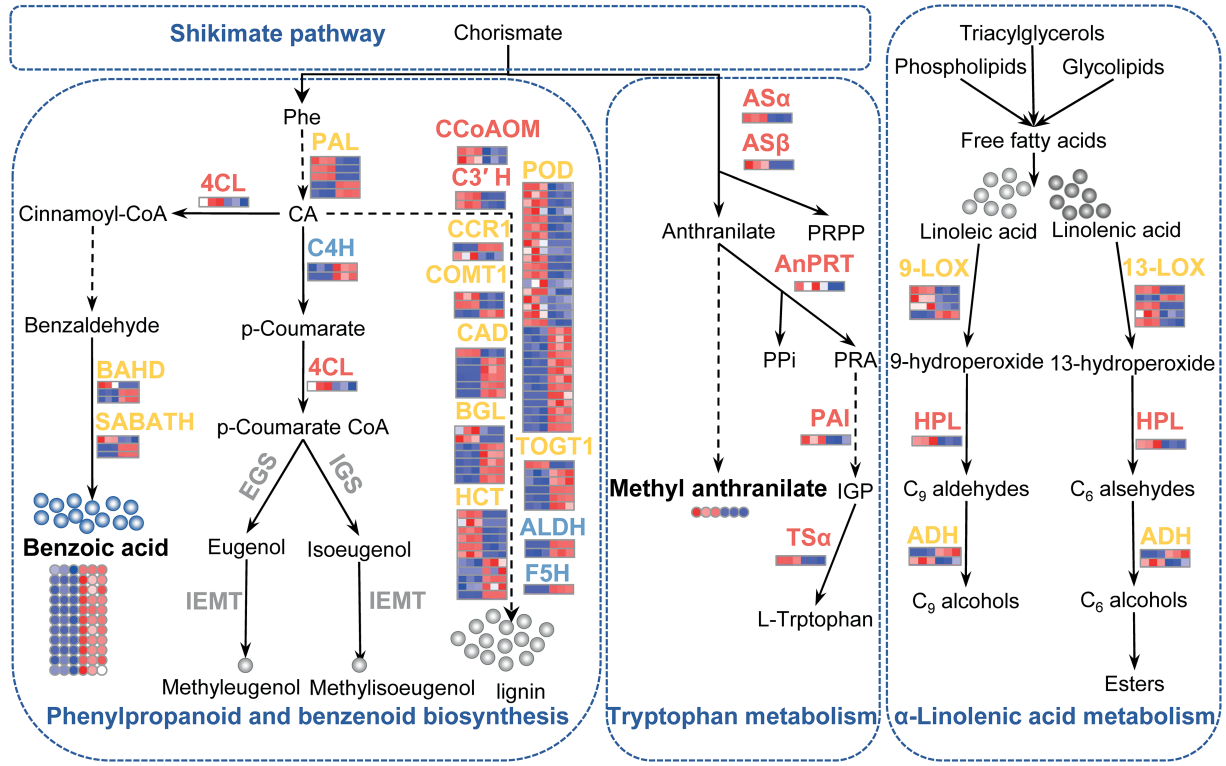


FIGURE 9
Schematic presentation of the biosynthesis of phenylpropanoids, benzenoids, and tryptophan, as well as alpha-linolenic acid metabolism in LJFs of 'Yujin2' compared to 'Fengjin1'. The upregulated, downregulated, mix-regulated, and unchanged genes (bold fonts) and metabolites (small circles) in 'Yujin2' are represented in red, blue, yellow, and gray, respectively. The scale bars adjacent to the name of genes o are picked up from the heatmaps for the transcriptome (Figure 4D). The first three and the latter three squares of a bar represent the normalized abundance values of DEGs from 'Yujin2' and 'Fengjin1', respectively. Red and blue colors reflect high and low expression levels, respectively. The asterisk followed by the gene names indicate that gene expression was validated by qRT-PCR. The solid line and dashed line indicate single-step reactions and multi-step reactions, respectively. 4CL, 4-hydroxycinnamoyl CoA ligase; ADH, alcohol dehydrogenase; ALDH, coniferyl-aldehyde dehydrogenase; AnPRT, anthranilate phosphoribosylanthranilate transferase; ASα, ASβ, anthranilate synthase; BGL, beta-glucosidase; C3′ H, 5-O-(4-coumaroyl)-D-quinate 3′-monooxygenase; C4H, cinnamate 4-hydroxylase; CA, trans-cinnamate; CAD, cinnamyl alcohol dehydrogenase; CCoAOM, caffeoyl-CoA O-methyltransferase; CCR1, cinnamoyl-CoA reductase; COMT1, caffeic acid 3-O-methyltransferase; EGS, eugenol synthase; F5H, ferulate-5-hydroxylase; HCT, shikimate O-hydroxycinnamoyl transferase; HPL, hydroperoxide lyase; IEMT, isoeugenol O-methyltransferase; IGS, isoeugenol synthase; LOX, lipoxygenase; PAI, phosphoribosylanthranilate isomerase; PAL, phenylalanine ammonia-lyase; Phe, phenylalanine; POD, peroxidase; PPi, inorganic pyrophosphate; PRPP, phosphoribosyl pyrophosphate; SABATH, S-adenosylmethionine-dependent methyltransferase; TOGT1, scopoletin glucosyltransferase; TSα, tryptophan synthase alpha subunit.

'Yujin2' (Supplementary Table 3). Indole is a volatile compound emitted to a plant's surroundings and functions as a remote signal, while indole-derived metabolites are mainly non-volatile and are embedded in many biological systems. Our results implied that the upregulation of these genes may facilitate the production of tryptophan and non-volatile derivatives, such as the biosynthesis of scent metabolites and the phytohormone auxin, and well as the coloring of yellow petals in 'Yujin2' (Cna'ani et al., 2018).

Noteworthy, methyl anthranilate, a VOC with grape-like and orange blossom odor that was substantially accumulated and considerably increased in 'Yujin2' compared to 'Fengjin1', is derived from the methylation of anthranilate in plants (Luo et al., 2019). Anthranilate, which is derived from chorismate *via* the AS enzyme complex, is an intermediate in tryptophan biosynthesis (Luo et al., 2019). The significantly upregulated Asα and Asβ in 'Yujin2' may be closely related to the increase of methyl anthranilate and the strong aroma of 'Yujin2' (Figure 9; Supplementary Table 12).

Fatty acid derivative biosynthetic pathway was activated in 'Yujin2'

Fatty acid derivatives also constitute a major class of flower VOCs, which are derived from the unsaturated C18 fatty acids linolenic and linoleic acid. The initiation of volatile fatty acid derivative biosynthesis is catalyzed by 9-and 13-lipoxygenase (LOX), which leads to the formation of 9-and 13-hydroperoxide intermediates, respectively. Then, 9-and 13-hydroperoxide lyases (HPLs), respectively, convert 9-and 13-hydroperoxides to volatile C9 and C6 aldehydes,

which are reduced by alcohol dehydrogenases (ADHs) to C9 and C6 alcohols, and C6 alcohols are converted into esters by alcohol acyltransferase. In the current study, seven genes of LOXs, an HPL gene, and an ADH2 gene were significantly upregulated in 'Yujin2' (Figure 9; Supplementary Table 12). The upregulation of these genes involved in the fatty acid derivative biosynthetic pathway plays an important role in the accumulation of volatile C9 and C6 aldehydes and alcohols, as well as various esters, in 'Yujin2' (Muhlemann et al., 2014).

Transcription factors play a crucial role in the regulation of the floral scent biosynthetic network

Orchestrated production of VOCs from these independent pathways requires coordinated transcriptional regulation of the floral scent biosynthetic network. To date, a number of TF families have been reported to regulate the production of VOCs in various plants, including MYB, bHLH, WRKY, ERF/AP2, bZIP, and NAC-type TFs (Qiao et al., 2021; Mostafa et al., 2022). For instance, ODORANT1 (DOD1), a member of the R_2R_3-type MYB family, regulates the synthesis of precursors in the shikimate pathway and the entry points to the phenylpropanoid and benzenoid biosynthetic pathways (Verdonk et al., 2005; Yoshida et al., 2018). In the current study, DOD1 and its predicted target gene EPSPS involved in shikimate pathway were both downregulated, which were consistent with the significantly decreased accumulation of benzenoids in 'Yujin2' (Figure 5; Supplementary Table 12). In *Osmanthus fragrans* flowers, WRKY TFs play important roles in regulating the biosynthesis of volatile monoterpenes (Ding et al., 2020). The WRKY-regulated NMD, a gene involved in monoterpenoid biosynthesis, was significantly upregulated, which may contribute to the monoterpenes production in 'Yujin2' (Figure 5; Supplementary Table 12). The plant-specific TF LFY is a regulator of early flower development and involved in activating the expression of floral organ identity genes (Wellmer and Riechmann, 2010). However, no information is available to show if LFY is involved in the floral scent regulation, which warrants further investigation. The identification of these differentially regulated TFs between 'Yujin2' and 'Fengjin1' will pave the way for investigating the regulatory functions of TFs in VOC biosynthesis in LJFs.

Conclusion

The regulatory mechanisms of the floral scents of *L. japonica* remain unclear. Integrative analyses of volatile metabolomics and transcriptomics of LJFs allowed identification of VOCs and pathways for the differences of floral aroma between 'Yujin2' and 'Fengjin1' (Figure 10). The differentially regulated pathways and VOCs in 'Yujin2' mainly include: (1) regulation of the TFs (MYB, WRKY, and LFY) for mediating the transcription of genes involved in VOC biosynthesis pathways; (2) activations of TES and GGPPS in MEP pathway for the increased odorous monoterpenoid, including alpha-terpineol and geraniol; (3) promoted HMGCS, HMGCR, MVK, FPS, and AFS1 in the MVA pathway for the induced odorous sesquiterpenoids, including farnesol, farnesal, and alpha-farnesene; (4) upregulated SQE1 and LUS in triterpenoid biosynthesis pathway and several genes involved in biosynthesis of phytosterols; (5) downregulation of biosynthesis and accumulation of benzenoids; (6) elevated tryptophan biosynthetic pathway for the increased tryptophan and methyl anthranilate; and (7) upregulation of the genes involved in the fatty acid derivative biosynthetic pathway for the accumulation of volatile C9 and C6 aldehydes and alcohols, and esters. Further functional analyses of the genes involved in VOC biosynthesis using molecular genetics and other approaches will promote advancements in understanding the regulatory mechanisms of floral scents in *L. japonica*.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found at: NCBI, PRJNA861870.

Author contributions

JL and JY: conceptualization. XY, QS, and ZS: experiments and data analysis. XY and JY: writing—original draft preparation. CC, JL, XZ, and JL: writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the Science and Technology Major Project of Xinxiang City (ZD2020002 to JL), the China Postdoctoral Science Foundation and Doctoral Start-up Funding of Henan Normal University (2022M712143 and 5101049170191 to JY), and the National Natural Science Foundation of China (31970380 to JL).

Acknowledgments

We thank the staff of Wuhan Metware Biotechnology Co., Ltd. (Wuhan, China) and OE Biotech (Shanghai, China) for their support in metabolomics and transcriptomics analyses, respectively.

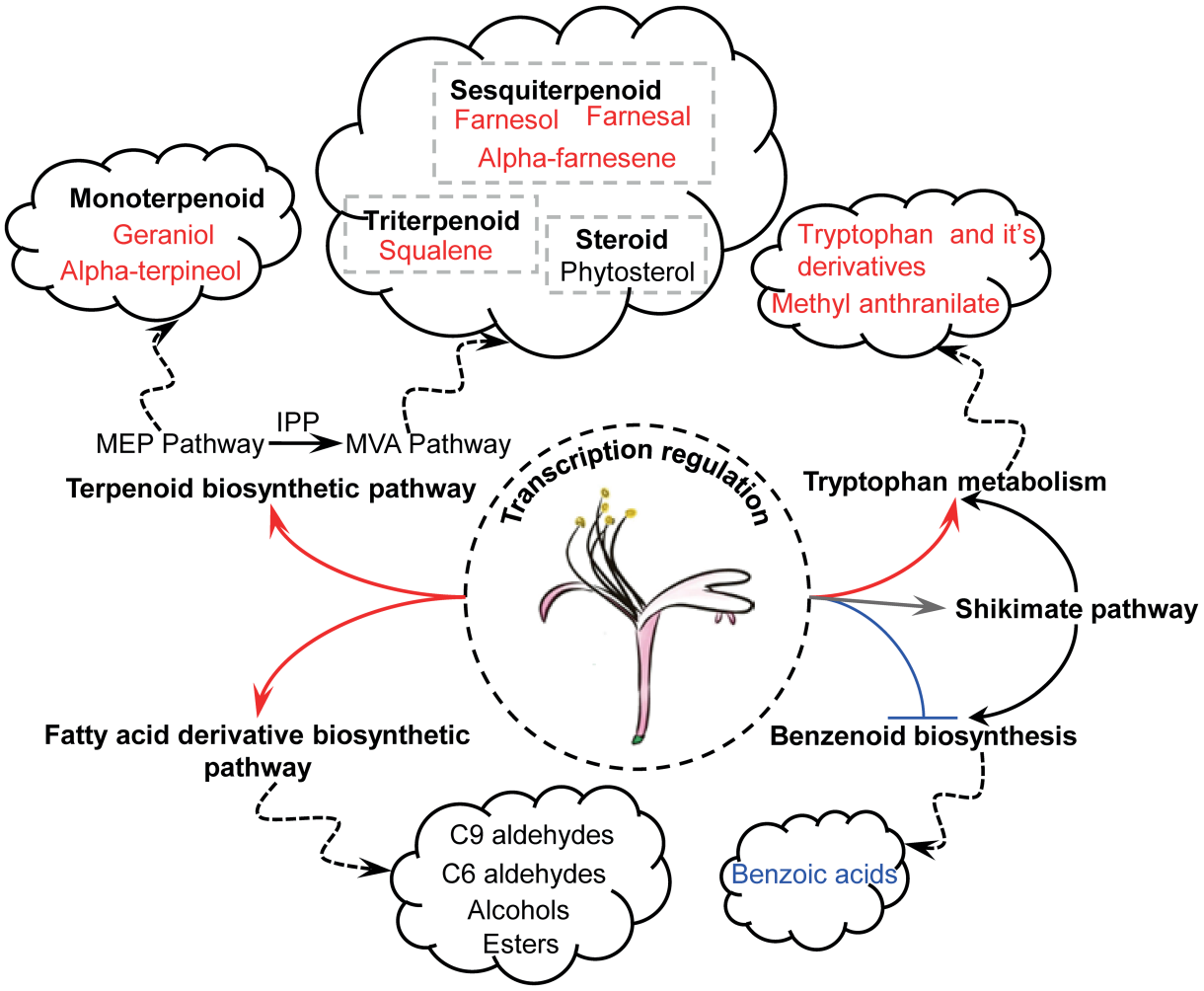


FIGURE 10
Schematic presentation of differentially regulated pathways and VOCs in LJFs of 'Yujin2' compared to 'Fengjin1'. The biosynthetic pathways and accumulations of terpenoids, tryptophan and its derivatives, and fatty acid derivatives were upregulated, while the benzenoid biosynthesis was downregulated in 'Yujin2'.

Conflict of Interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.989036/full#supplementary-material

References

Adal, A. M., and Mahmoud, S. S. (2020). Short-chain isoprenyl diphosphate synthases of lavender (*Lavandula*). *Plant Mol. Biol.* 102, 517–535. doi: 10.1007/s11103-020-00962-8

Cai, H., Cao, G., Li, L., Liu, X., Ma, X., Tu, S., et al. (2013). Profiling and characterization of volatile components from non-fumigated and sulfur-fumigated flos *Lonicerae japonicae* using comprehensive two-dimensional gas chromatography

time-of-flight mass spectrometry coupled with chemical group separation. *Molecules* 18, 1368–1382. doi: 10.3390/molecules18021368

Cai, Z., Wang, C., Chen, C., Chen, H., Yang, R., Chen, J., et al. (2021). Omics map of bioactive constituents in *Lonicera japonica* flowers under salt stress. *Ind. Crop. Prod.* 167:113526. doi: 10.1016/j.plaphy.2022.01.022

Cai, Z., Wang, C., Chen, C., Zou, L., Yin, S., Liu, S., et al. (2022). Comparative transcriptome analysis reveals variations of bioactive constituents in *Lonicera japonica* flowers under salt stress. *Plant Physiol. Bioch.* 173, 87–96. doi: 10.1016/j.plaphy.2022.01.022

Chen, Q., Fan, D., and Wang, G. (2015). Heteromeric geranyl (geranyl) diphosphate synthase is involved in monoterpene biosynthesis in Arabidopsis flowers. *Mol. Plant* 8, 1434–1437. doi: 10.1016/j.molp.2015.05.001

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* 66, 212–229. doi: 10.1111/j.1365-313X.2011.04520.x

Cna'ani, A., Seifan, M., and Tzin, V. (2018). Indole is an essential molecule for plant interactions with herbivores and pollinators. *J. Plant Biol. Crop Res.* 1:1003. doi: 10.33582/2637-7721/1003

D'Auria, J. C. (2006). Acyltransferases in plants: a good time to be BAHD. *Curr. Opin. Plant Biol.* 9, 331–340. doi: 10.1016/j.pbi.2006.03.016

D'Auria, J., Chen, F., and Pichersky, E. (2003). Chapter eleven the SABATH family of MTs in *Arabidopsis thaliana* and other plant species. *Recent Adv. Phytochem.* 37, 253–283. doi: 10.1016/S0079-9920(03)80026-6

Ding, W., Ouyang, Q., Li, Y., Shi, T., Li, L., Yang, X., et al. (2020). Genome-wide investigation of WRKY transcription factors in sweet osmanthus and their potential regulation of aroma synthesis. *Tree Physiol.* 40, 557–572. doi: 10.1093/treephys/tpz129

Du, C., Feng, X., Wang, H., Wu, L., and Li, P. (2015). Analysis of volatile constituents in *Lonicera japonica* Thunb. From different origins by GC-MS. *Agric. Sci. Technol.* 16, 1081–1083. doi: 10.16175/j.cnki.1009-4229.2015.05.051

Dudareva, N., Klempien, A., Muhlemann, J., and Kaplan, I. (2013). Biosynthesis, function and metabolic engineering of plant volatile organic compounds. *New Phytol.* 198, 16–32. doi: 10.1111/nph.12145

Dudareva, N., and Pichersky, E. (2006). "Floral scent metabolic pathways: their regulation and evolution," in *Biology of Floral Scent*, eds. N. Dudareva and E. Pichersky. CRC Press, Boca Raton.

Fang, H., Qi, X., Li, Y., Yu, X., Xu, D., Liang, C., et al. (2020). De novo transcriptomic analysis of light-induced flavonoid pathway, transcription factors in the flower buds of *Lonicera japonica*. *Trees* 34, 267–283. doi: 10.1007/s00468-019-01916-4

Gerard, F., Iolanda, F., Joan, L., and Josep, P. (2017). β-Ocimene, a key floral and foliar volatile involved in multiple nteractions between plants and other organisms. *Molecules* 22:1148. doi: 10.3390/molecules22071148

Ilie, D., Radulescu, V., and Dutu, L. (2014). Volatile constituents from the flowers of two species of honeysuckle (*Lonicera japonica* and *Lonicera caprifolium*). *Farmacia* 44, 206–209. doi: 10.1021/jf950275b

Jiang, S., Jin, J., Sarojam, R., and Ramachandran, S. (2019). A comprehensive survey on the terpene synthase gene family provides new insight into its evolutionary patterns. *Genome Biol. Evol.* 11, 2078–2098. doi: 10.1093/gbe/evz142

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Knudsen, J.T., and Gershenzon, J. (2006). "The chemical diversity of floral scent," in *Biology of Floral Scent*, eds. N. Dudareva and E. Pichersky. CRC Press (Boca Raton).

Li, Y., Li, W., Fu, C., Song, Y., and Fu, Q. (2020). *Lonicerae japonicae* flos and Lonicerae flos: a systematic review of ethnopharmacology, phytochemistry and pharmacology. *Phytochem. Rev.* 19, 1–61. doi: 10.1007/s11101-019-09655-7

Li, J., Lian, X., Ye, C., and Wang, L. (2019). Analysis of flower color variations at different developmental stages in two honeysuckle (*Lonicera japonica* Thunb.) cultivars. *HortScience* 54, 779–782. doi: 10.21273/HORTSCI13819-18

Li, J., Wang, J., Ren, M., Shang, X., Liu, B., and Zhang, G. (2016). Analysis of yield and index components of *Lonicera japonica* new strain and major cultivars. *J. Henan Agri. Univ.* 50, 19–24. doi: 10.16445/j.cnki.1000-2340.2016.01.004

Lima, A. S., Schimmel, J., Lukas, B., Novak, J., Barroso, J. G., Figueiredo, A. C., et al. (2013). Genomic characterization, molecular cloning and expression analysis of two terpene synthases from *thymus caespititius* (Lamiaceae). *Planta* 238, 191–204. doi: 10.1007/s00425-013-1884-2

Lin, D., Chen, Z., Zhang, W., and Lu, Z. (2020). Identification of key aroma components in honeysuckle essential oil by GC-O and GC-MS. *Shandong Chem. Indus.* 49, 106–120. doi: 10.19319/j.cnki.issn.1008-021x.2020.11.039

Liu, X. (2017). Selection of candidate reference genes for gene Expression Studies by Quantitative real-time PCR in *Lonicera japonica* Thunb. Dissertation, Shanxi: Shanxi Agricultural University.

Liu, H., Guo, S., Lu, M., Zhang, Y., Li, J., Wang, W., et al. (2019). Biosynthesis of $DHGA_{12}$ and its roles in Arabidopsis seedling establishment. *Nat. Commun.* 10:1768. doi: 10.1038/s41467-019-09467-5

Luo, Z., Cho, J. S., and Lee, S. Y. (2019). Microbial production of methyl anthranilate, a grape flavor compound. *Proc. Natl. Acad. Sci. U. S. A.* 116, 10749–10756. doi: 10.1073/pnas.1903875116

Mostafa, S., Wang, Y., Zeng, W., and Jin, B. (2022). Floral scents and fruit aromas: functions, compositions, biosynthesis, and regulation. *Front. Plant Sci.* 13:860157. doi: 10.3389/fpls.2022.860157

Muhlemann, J., Klempien, A., and Dudareva, N. (2014). Floral volatiles: from biosynthesis to function. *Plant Cell Environ.* 37, 1936–1949. doi: 10.1111/pce.12314

Nagegowda, D. A., and Gupta, P. (2020). Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant Sci.* 294:110457. doi: 10.1016/j.plantsci.2020.110457

Parachnowitsch, A. L., Raguso, R. A., and Kessler, A. (2012). Phenotypic selection to increase floral scent emission, but not flower size or colour in bee-pollinated penstemon digitalis. *New Phytol.* 195, 667–675. doi: 10.1111/j.1469-8137.2012.04188.x

Peled-Zehavi, H., Oliva, M., Xie, Q., Tzin, V., OrenShamir, M., Aharoni, A., et al. (2015). Metabolic engineering of the phenylpropanoid and its primary, precursor pathway to enhance the flavor of fruits and the aroma of flowers. *Bioengineering* 2, 204–212. doi: 10.3390/bioengineering2040204

Pu, X., Li, Z., Tian, Y., Gao, R., Hao, L., Hu, Y., et al. (2020). The honeysuckle genome provides insight into the molecular mechanism of carotenoid metabolism underlying dynamic flower coloration. *New Phytol.* 227, 930–943. doi: 10.1111/nph.16552

Qiao, Z., Hu, H., Shi, S., Yuan, X., Yan, B., and Chen, L. (2021). An update on the function, biosynthesis and regulation of floral volatile terpenoids. *Horticulturae* 7:451. doi: 10.3390/horticulturae7110451

Schiestl, F. (2015). Ecology and evolution of floral volatile-mediated information transfer in plants. *New Phytol.* 206, 571–577. doi: 10.1111/nph.13243

Shang, X., Pan, H., Li, M., Miao, X., and Ding, H. (2011). *Lonicera japonica* Thunb.: ethnopharmacology, phytochemistry and pharmacology of an important traditional chinese medicine. *J. Ethnopharmacol.* 138, 1–21. doi: 10.1016/j.jep.2011.08.016

Shang, J., Tian, J., Cheng, H., Yan, Q., Li, L., Jamal, A., et al. (2020). The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. *Genome Biol.* 21, 200–228. doi: 10.1186/s13059-020-02088-y

Tholl, D., Kish, C. M., Orlova, I., Sherman, D., Gershenzon, J., Pichersky, E., et al. (2004). Formation of monoterpenes in *Antirrhinum majus* and *Clarkia breweri* flowers involves heterodimeric geranyl diphosphate synthases. *Plant Cell* 16, 977–992. doi: 10.1105/tpc.020156

Tzin, V., and Galili, G. (2010). New insights into the shikimate and aromatic amino acids biosynthesis pathways in plants. *Mol. Plant* 3, 956–972. doi: 10.1093/mp/ssq048

Verdonk, J. C., Haring, M. A., Van Tunen, A. J., and Schuurink, R. C. (2005). ODORANT1 regulates fragrance biosynthesis in *petunia* flowers. *Plant Cell* 17, 1612–1624. doi: 10.1105/tpc.104.028837

Wang, L., Jiang, Q., Hu, J., Zhang, Y., and Li, J. (2016). Research progress on chemical constituents of *Lonicerae japonicae* flos. *Biomed. Res. Int.* 2016, 1–18. doi: 10.1155/2016/8968940

Wang, H., Li, Y., Wang, S., Kong, D., Sahu, S. K., Bai, M., et al. (2020). Comparative transcriptomic analyses of chlorogenic acid and iuteolosides biosynthesis pathways at different flowering stages of diploid and tetraploid *Lonicera japonica*. *PeerJ* 8:e8690. doi: 10.7717/peerj.8690

Wang, L., Li, M., Yan, Y., Ao, M., Wu, G., and Yu, L. (2009). Influence of flowering stage of *Lonicera japonica* Thunb. on variation in volatiles and chlorogenic acid. *J. Sci. Food Agr.* 89, 953–957. doi: 10.1002/jsfa.3537

Wellmer, F., and Riechmann, J. L. (2010). Gene networks controlling the initiation of flower development. *Trends Genet.* 26, 519–527. doi: 10.1016/j.tig.2010.09.001

Xia, Y., Chen, W., Xiang, W., Wang, D., Xue, B., Liu, X., et al. (2021). Integrated metabolic profiling and transcriptome analysis of pigment accumulation in *Lonicera japonica* flower petals during colour-transition. *BMC Plant Biol.* 21, 98–14. doi: 10.1186/s12870-021-02877-y

Xiao, Q., Li, Z., Qu, M., Xu, W., Su, Z., and Yang, J. (2021). LjaFGD: *Lonicera japonica* functional genomics database. *J. Integr. Plant Biol.* 63, 1422–1436. doi: 10.1111/jipb.13112

Yan, L., Xie, Y., Wang, Y., Zhu, J., Li, M., Liu, X., et al. (2020). Variation in contents of active components and antibacterial activity in different parts of *Lonicera japonica* Thunb. *Asian Biomed.* 14, 19–26. doi: 10.1515/abm-2020-0004

Yang, B., Zhong, Z., Wang, T., Ou, Y., Tian, J., Komatsu, S., et al. (2019). Integrative omics of *Lonicera japonica* Thunb. Flower development unravels molecular changes regulating secondary metabolites. *J. Proteome* 208:103470. doi: 10.1016/j.jprot.2019.103470

Yoo, H., Kang, H., Song, Y., Park, E., and Lim, C. (2008). Anti-angiogenic, antinociceptive and anti-inflammatory activities of *Lonicera japonica* extract. *J. Pharm. Pharmacol.* 60, 779–786. doi: 10.1211/jpp.60.6.0014

Yoshida, K., Oyama-Okubo, N., and Yamagishi, M. (2018). An R_2R_3-MYB transcription factor ODORANT1 regulates fragrance biosynthesis in lilies (*Lilium spp.*). *Mol. Breed.* 168, 598–514. doi: 10.1104/pp.114.252908

Yu, J., JM, G., Z, D., Q, S., B, T., J, K., et al. (2021). Integrative proteomic and phosphoproteomic analyses of pattern-and effector-triggered immunity in tomato. *Front. Plant Sci.* 12:768693. doi: 10.3389/fpls.2021.768693

Zhou, W., Kügler, A., McGale, E., Haverkamp, A., Knaden, M., Guo, H., et al. (2017). Tissue-specific emission of (E)-α-bergamotene helps resolve the dilemma when pollinators are also herbivores. *Curr. Biol.* 27, 1336–1341. doi: 10.1016/j.cub.2017.03.017

Zhou, Z., Li, X., Liu, J., Dong, L., Chen, Q., Liu, J., et al. (2015). Honeysuckle-encoded atypical microRNA2911 directly targets influenza A viruses. *Cell Res.* 25, 39–49. doi: 10.1038/cr.2014.130

Zhou, L., Zhou, Z., Jiang, X., Zheng, Y., Chen, X., Fu, Z., et al. (2020). Absorbed plant MIR2911 in honeysuckle decoction Iinhibits SARS-CoV-2 replication and accelerates the negative conversion of infected patients. *Cell Discov.* 6, 1–4. doi: 10.1038/s41421-020-00197-3

Zhu, G., Gou, J., Klee, H., and Huang, S. (2019). Next-gen approaches to flavor-related metabolism. *Annu. Rev. Plant Biol.* 70, 187–212. doi: 10.1146/annurev-arplant-050718-100353

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 14 September 2022
DOI 10.3389/fpls.2022.994058

OPEN ACCESS

EDITED BY
Zhi-Yan Du,
University of Hawai'i at Mānoa,
United States

REVIEWED BY
Zhansheng Li,
Institute of Vegetables and Flowers (CAAS), China
Haihai Wang,
Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences (CAS), China

*CORRESPONDENCE
Yan Bao
yanbao@sjtu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 14 July 2022
ACCEPTED 15 August 2022
PUBLISHED 14 September 2022

CITATION
Niu Y, Zhang Q, Wang J, Li Y, Wang X and Bao Y (2022) Vitamin E synthesis and response in plants. *Front. Plant Sci.* 13:994058. doi: 10.3389/fpls.2022.994058

Vitamin E synthesis and response in plants

Yue Niu[1,2†], Qian Zhang[1,2†], Jiaojiao Wang[1,2], Yanjie Li[2], Xinhua Wang[2] and Yan Bao[1,2*]

[1]Shanghai Collaborative Innovation Center of Agri-Seeds, Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China, [2]School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China

Vitamin E, also known as tocochromanol, is a lipid-soluble antioxidant that can only be produced by photosynthetic organisms in nature. Vitamin E is not only essential in human diets, but also required for plant environment adaptions. To synthesize vitamin E, specific prenyl groups needs to be incorporated with homogentisate as the first step of reaction. After decades of studies, an almost complete roadmap has been revealed for tocochromanol biosynthesis pathway. However, chlorophyll-derived prenyl precursors for synthesizing tocochromanols are still a mystery. In recent years, by employing forward genetic screening and genome-wide-association approaches, significant achievements were acquired in studying vitamin E. In this review, by summarizing the recent progresses in vitamin E, we provide to date the most updated whole view of vitamin E biosynthesis pathway. Also, we discussed about the role of vitamin E in plants stress response and its potential as signaling molecules.

KEYWORDS

vitamin E, VTE, tocopherol, tocochromanol, biosynthesis, pathway, stress, signal

Introduction

Vitamin E (also called tocochromanols) is an essential micronutrient for humans, which is produced by phototrophs such as plants, and some algae (Falk and Munné-Bosch, 2010). As antioxidant, vitamin E can convert free radicals into less reactive compounds, playing a pivotal role in human health (Zingg, 2007). Insufficient consumption of vitamin E could cause many diseases, including cancers, Alzheimer's disease and cardiovascular disease, but not limited (Sen et al., 2007; Falk and Munné-Bosch, 2010; Sozen et al., 2019).

Tocochromanol molecule contains a polar chromanol ring head and a prenyl side chain. According to the various types of side chains, tocochromanols can be defined as tocopherol, tocotrienol, plastochromanol-8 (PC-8) and tocomonoenol (**Figure 1**). Tocopherol contains fully saturated aliphatic side chain, while the side chain of tocotrienol contains three extra trans double bonds. PC-8 has similar unsaturated as

tocotrienol but longer side chain, whereas tocomonoenol has only one double bond on its side chain (Tian et al., 2007; Sadre et al., 2010; Stacey et al., 2016). Based on the differences in numbers and positions of methyl groups on the chromanol ring head, tocochromanol isoforms can be classified as α, β, γ, and δ (**Figure 1**). While for PC-8, only δ-form was found in nature (Kamal-Eldin and Appelqvist, 1996).

Tocopherols are ubiquitously synthesized in all plant species and especially abundant in photosynthetic tissues and seeds. Tocotrienols are mostly concentrated in monocot species like maize (*Zea mays*) embryo (Falk and Munné-Bosch, 2010), and the three extra double bonds in tocotrienols were believed to able to confer greater potential for scavenging peroxyl radicals (Suzuki et al., 1993). PC-8 was first identified in the leaves of the rubber tree and later found to be also enriched in *Brassica napus*, tomato fruit, and tuber of *Dioscorea alata* (Goffman and Mollers, 2000; Cheng et al., 2007; Zbierzak et al., 2010), and the concentrations of leaf PC-8 are differed by species and developmental stages (Whittle et al., 1965; Kruk et al., 2014). Studies of tocomonoenols suggested that they were mainly accumulated in the seed oil from palm, Slovenia pumpkin and sunflower (Matsumoto et al., 1995; Butinar et al., 2011; Hammann et al., 2015), but the exact function of tocomonoenols *in planta* are yet to be thoroughly verified.

In the past few decades, *Arabidopsis thaliana* has been employed as a plant model for dissecting tocopherol function and its biosynthesis pathway, and many key VTE (*Vi*T*amin E*) enzymes were identified. Recently, by applying EMS-based forward genetic screening and genome-wide association study (GWAS), several key chlorophyll metabolic enzymes were identified to be required for tocopherol biosynthesis, including POR, CLD1 and VTE7 etc. (Lin et al., 2016; Diepenbrock et al., 2017; Wang et al., 2018; Hershberger et al., 2022; Wu et al., 2022). In this review, we summarize the recent progresses in tocochromanol biosynthesis, discussing the role of tocochromanol in plant stress response, and its potentials in signal transduction.

Biosynthesis of vitamin E

To produce tocochromanols, homogentisate (HGA) will be condensed with different prenyl chains by homogentisate phytyltransferase [HPT, also named VTE2 (*Vi*T*amin E* 2 loci)], in one-to-one ratio (**Figure 1**). HPT genes can be found in all green plants, and some algae including cyanobacterium *Synechocystis* (Collakova and DellaPenna, 2001; Savidge et al., 2002). For tocopherols, HGA and phytyl diphosphate (PDP) are catalyzed by VTE2 to generate 2-methyl-6-phytyl-1,4-benzoquinol (MPBQ). VTE2 can also use the tetrahydrogeranylgeranyl pyrophosphate (THGGPP) to generate 2-methyl-6-tetrahydrogeranylgeranyl-1,4-benzoquinol (MTHGGBQ) for producing tocomonoenol in Arabidopsis seeds (Pellaud et al., 2018). In monocots, homogentisate geranylgeranyl transferases (HGGTs) are seed-specific and plastid-targeted, which can condense geranylgeranyl pyrophosphate (GGDP) instead of PDP with HGA to generate 2-methyl-6-geranylgeranyl-1,4-benzoquinol (MGGBQ) for tocotrienol biosynthesis (Cahoon et al., 2003; Yang et al., 2011). Although HGGT and VTE2 are close in structure, their enzyme activities toward different substrates vary quite much. For example, the activity of barley HPPT toward GGDP is 6 times higher than that of PDP; the VTE2 enzyme confers 9 times higher activity toward PDP than that of GGDP. Interestingly, barley HPPT can restore the levels of both tocopherols and tocotrienols in Arabidopsis *vte2* mutant (Yang et al., 2011). Moreover, homogentisate solanesyltransferase (HST) can condense HGA and solanesyl pyrophosphate (SPP) to produce the 2-methyl-6-solanesyl-1,4-benzoquinol (MSBQ), the precursor of PC-8 (Sadre et al., 2006; Tian et al., 2007).

The downstream of vitamin E biosynthesis pathway is divided into two branches. In one branch, the MPBQ, MGGBQ, MSBQ, and MTHGGBQ are methylated by a methyltransferase (MT, VTE3) to produce DMPBQ, DMGGBQ, DMTHGGBQ, and DMSBQ (PQ-9), respectively, the precursors of α- and γ-tocochromanols (Cheng et al., 2003; Van Eenennaam et al., 2003). Then the methylated compounds are cyclized by tocopherol cyclase (TC, VTE1) to produce γ-tocochromanols (Cheng et al., 2003). In the other branch of this pathway, the demethylated compounds (MPBQ, MGGBQ, MSBQ, and MTHGGBQ) are cyclized directly by VTE1 to produce δ-tocochromanols (Cheng et al., 2003). In the final step, the γ- and δ-tocochromanols, respectively, are methylated to produce α- and β-tocochromanols by γ-tocopherol methyltransferase (γ-TMT, VTE4) (Shintani and DellaPenna, 1998; Bergmüller et al., 2003).

Biosynthesis of tocochromanols is mainly carried out in the plastid, but one of its main precursors HGA, which provides the chromanol ring head for tocochromanols, is produced during L-tyrosine (Tyr) degradation in the cytoplasm (**Figure 1**). Through the shikimate pathway, tyrosine aminotransferases (TATs) catalyze reversible reaction between Tyr and 4-hydroxyphenylpyruvate (HPP), and at least two homologous genes *TAT1* and *TAT2* were identified in the genome of *Arabidopsis thaliana* (Siehl et al., 2014; Stacey et al., 2016; Wang et al., 2016). Arabidopsis *TAT1* gene loss of function mutant showed reduced tocopherols under normal condition (Riewe et al., 2012), but the *tat2* single mutants have no effect on Tyr and tocopherol levels. In the *tat1 tat2* double mutants, compared with wild-type and *tat* single mutants, more Tyr were accumulated and fewer tocopherols were maintained, and this effect was amplified under high-light stress (Wang et al., 2019). TAT1 and TAT2 thus work redundantly in Tyr degradation and tocopherol biosynthesis, with TAT1 playing a major role. Then, HPP is converted to HGA by the 4-hydroxyphenylpyruvate dioxygenase (HPPD), which is encoded by a single-copy gene

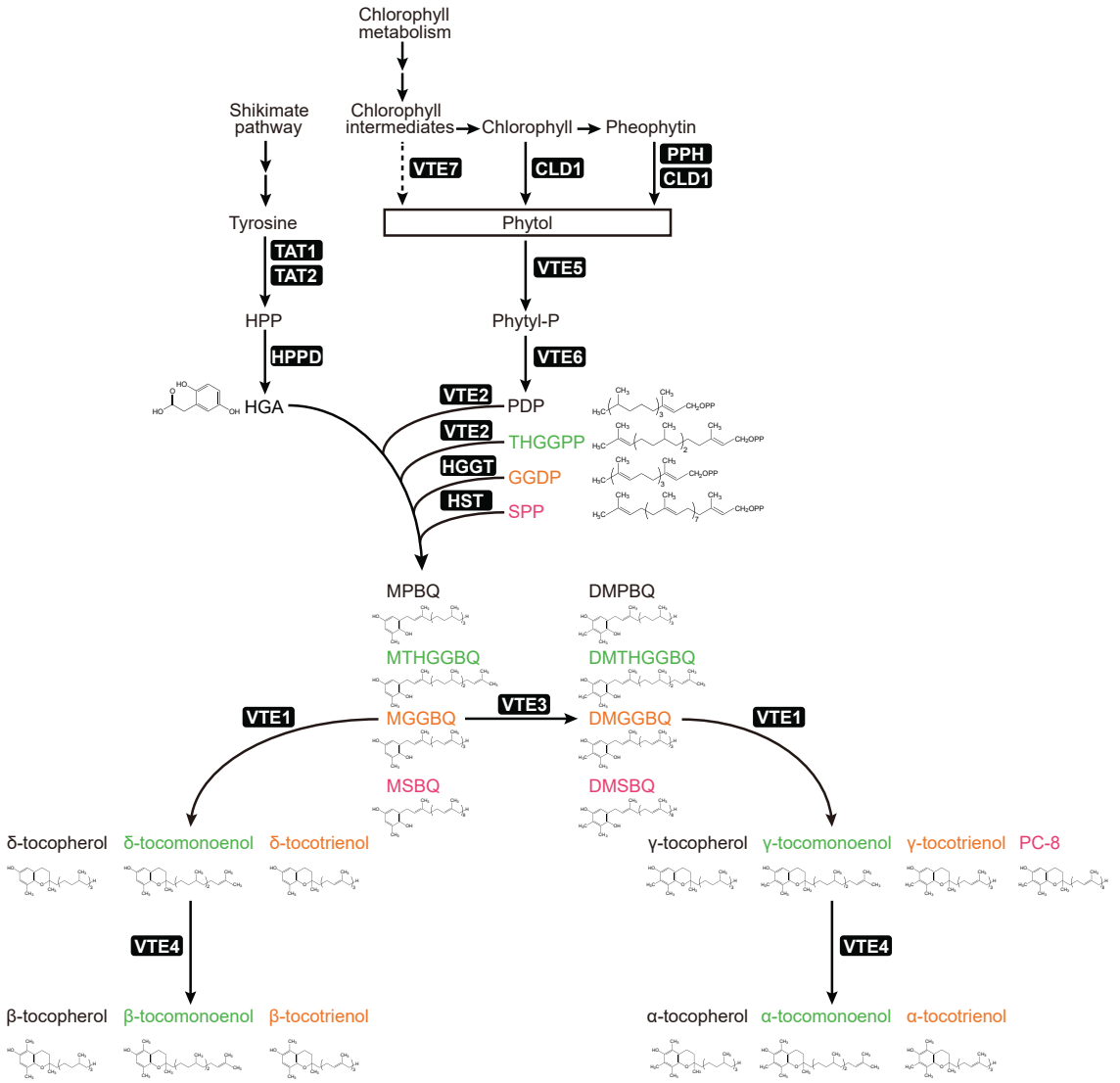


FIGURE 1
Plant tocochromanol biosynthesis pathway and sources of metabolites. Metabolite names were colored for recognizing different tocochromanol synthesis pathways. Abbreviations of metabolites: HPP, 4-hydroxyphenylpyruvate; HGA, homogentisate; Phytyl-P, phytyl-phosphate; PDP, phytyl-diphosphate; THGGPP, tetrahydrogeranylgeranyl pyrophosphate; GGDP, geranylgeranyl pyrophosphate; SPP, solanesyl pyrophosphate; MPBQ, 2-methyl-6-phytyl-1,4-benzoquinol; DMPBQ, 2,3-dimethyl-6-phytyl-1,4-benzoquinol. Abbreviations of enzymes: TAT1, tyrosine aminotransferase 1; TAT2, tyrosine aminotransferase 2; HPPD, 4-hydroxyphenylpyruvate dioxygenase; CLD1, chlorophyll dephytylase 1; PPH, pheophytin pheophorbide hydrolase; VTE, vitamin E biosynthetic enzyme.

in Arabidopsis (Tsegaye et al., 2002). However, HPPDs possess different subcellular localizations according to studies in various plant species (Pellaud and Mène-Saffrané, 2017). For example, HPPD proteins of *Spinacia oleracea*, *Lemna gibba* and maize are targeted to plastids, while in carrot and Arabidopsis they are located in the cytoplasm (Löffelhardt and Kindl, 1979; Fiedler et al., 1982; Garcia et al., 1997; Siehl et al., 2014; Wang et al., 2016).

Another main precursor of tocopherols comes from phytyl diphosphate (PDP), which is generated from two steps of phosphorylation relay using phytol (Gutbrod et al., 2019). The first step is to generate phytyl-phosphate via VTE5 (phytol kinase), which accounts for 80% and 50% total tocopherol biosynthesis in Arabidopsis seeds and leaves, respectively (Valentin et al., 2006). The second step is executed by VTE6 (phytyl phosphate kinase), and due to over accumulation of phytyl-phosphate, *vte6* mutant plant confers severe growth defects. By introducing *vte5* into the *vte6* background, the *vte5 vte6* double mutant resembles wild-type in plant growth, but showing tocopherol deficient and high levels of accumulated chlorophylls (Vom Dorp et al., 2015). GGDP, SPP and THGGPP are the polyprenyl chain precursors of

tocotrienols, PC-8 and tocomonoenols, respectively (Mène-Saffrané, 2018). Of note, all of the four tocochromanol isoprenoid side chains can be produced by GGDP synthases (Pellaud and Mène-Saffrané, 2017).

However, accumulated studies have suggested that the phytol group used for tocopherol biosynthesis mostly comes from the chlorophyll degradation (Valentin et al., 2006; Gutbrod et al., 2019). Thus, finding the relevant hydrolases that can bring down phytol group from chlorophyll and/or chlorophyll derivatives is key for dissecting tocopherol biosynthesis pathway. The first chlorophyllase (CLH) was isolated from *Citrus sinensis*, which shows activity toward chlorophylls, and two *CLH* genes (*AtCLH1* and *AtCLH2*) were found in Arabidopsis (Jacob-Wilk et al., 1999; Tsuchiya et al., 1999). The highest transcripts of *AtCLH1* and *AtCLH2* were found in young leaves and their levels decline gradually during leaf maturation (Chen et al., 2014, Tian et al., 2021). However, previous studies had shown that the two Arabidopsis CLHs were not required for chlorophyll breakdown during leaf senescence. In addition, CLH proteins are targeted to vacuole instead of chloroplasts. Based on functional genomic approach with the features of senescence-related regulation and chloroplast targeting, Schelbert et al. identified a hydrolase called pheophytin pheophorbide hydrolase (PPH) in Arabidopsis. The *pph* mutant showed stay-green phenotype during senescence, compared with its wild-type control. *In vitro*, PPH confers specific activity toward pheophytin but not chlorophylls (Schelbert et al., 2009). Recently, through EMS screening for Arabidopsis heat-sensitive progenies, an Arabidopsis *CHLOROPHYLL DEPHYTYLASE1* (*CLD1*) gene was identified. A G-to-A transition at position 957 of *cld1* gene results in the replacement of Gly-193 by Asp (G193D), which confers much higher activity of *cld1* toward both chlorophylls and pheophytin than that of its native CLD1 (Lin et al., 2016). Following studies found about 15% tocopherol increase in CLD1 and *cld1* overexpression plants, but no significant difference in its miRNA lines (Lin and Charng, 2017). Although CLHs, PPH and CLD1 can cleave phytol directly from chlorophyll and/or pheophytin, none of the single or high order mutants (*clh1/clh2*, *pph,* and *pph/clh1/clh2*) significantly affect tocopherol contents. Moreover, overexpression of the four genes only moderately increased the levels of tocopherol, suggesting the relevant alpha/beta hydrolase is yet to be identified (Zhang et al., 2014; Lin and Charng, 2017). Recently, GWAS of seed tocopherols was applied using 814 Arabidopsis natural variation lines (part of the 1001 Arabidopsis Genome Panel, Alonso-Blanco et al., 2016), a novel seed-specific alpha/beta hydrolase gene *AtVTE7* was identified. AtVTE7 is targeted to the chloroplast envelope and accounts for 55% of total seed tocopherols. Consistent with the results in Arabidopsis, the maize orthologous gene *ZmVTE7* controls 38% and 49% total tocopherols in kernel and leaf, respectively (Albert et al., 2022). Of note, Arabidopsis AtVTE7 is only detected in seed, does not affect tocopherol trait in leaf. Although VTE7 can provide phytol from chlorophyll degradation for tocopherol biosynthesis, this enzyme mainly affects the levels of chlorophyll biosynthetic intermediates, instead of bulk chlorophyll levels (Albert et al., 2022).

In addition to the alpha/beta hydrolases that are directly involved in phytol production, many chlorophyll-metabolism-related genes also contribute to tocopherol homeostasis. For instance, plant NYE [Non-Yellowing, also named SGR (Stay-Green)] are chloroplast-localized Mg-dechelatase proteins, and during chlorophyll degradation NYEs closely cooperate with other chlorophyll catabolic enzymes including PPH (Armstead et al., 2007; Sato et al., 2007; Zhang et al., 2014). Indeed, the seed tocopherol levels of the Arabidopsis double mutant *nye1 nye2* modestly reduced, compared with that in the wild-type (Zhang et al., 2014). Chlorophyll synthesis needs the esterizing of chlorophyllide with either GGDP or PDP. The RNAi lines of *CHLOROPHYLL SYNTHASE* (*CHLSYN*) exhibit significantly reduced chlorophylls but up to 2 times increased tocopherols (Zhang et al., 2015). Moreover, two QTLs that encode homologs of protochlorophyllide reductase (*POR1* and *2*) were revealed in maize, and the two *por* loci had the highest phenotypic variance explained for all four forms of tocopherols calculated (Diepenbrock et al., 2017; Wang et al., 2018). Thus, disturbing the genes involved chlorophyll metabolism can affect tocopherol biosynthesis, suggesting a precise mechanism for balancing chlorophyll metabolism and tocopherol biosynthesis in plants.

Vitamin E in stress response

As important antioxidants, tocochromanols can be boosted to high levels during various biotic and abiotic stresses (**Figures 2**, **3**) (Bao et al., 2020). Meanwhile, plants with low levels of tocochromanols are more susceptible to different stressful treatments, suggesting a crucial role of vitamin E in plant environment adaptions.

Temperature and light intensity are the two key environmental factors that can affect crop yield (Pretty et al., 2010). During high light and heat stress, tocopherols (especially α-tocopherol) are induced to accumulate at high levels, and elevated α-tocopherols were believed to be required for protecting photosystem from savaging singlet oxygen and maintaining the stability of chloroplast (Kruk et al., 2005). Indeed, tocopherol deficient Arabidopsis mutants are more vulnerable to high light (Kobayashi and DellaPenna, 2008). When grown under low temperature, vitamin E deficient mutants are retarded in plant growth (Maeda et al., 2006), which mainly attributes to defects in phloem loading, coincide with the findings in maize and potato (*Solanum tuberosum*) (Russin et al., 1996; Hofius et al., 2004). The combination of high light and low temperature causes strong lipid peroxidation and photooxidative stresses to Arabidopsis, and this effect was exemplified in *vte1* mutant (Havaux et al., 2005). In

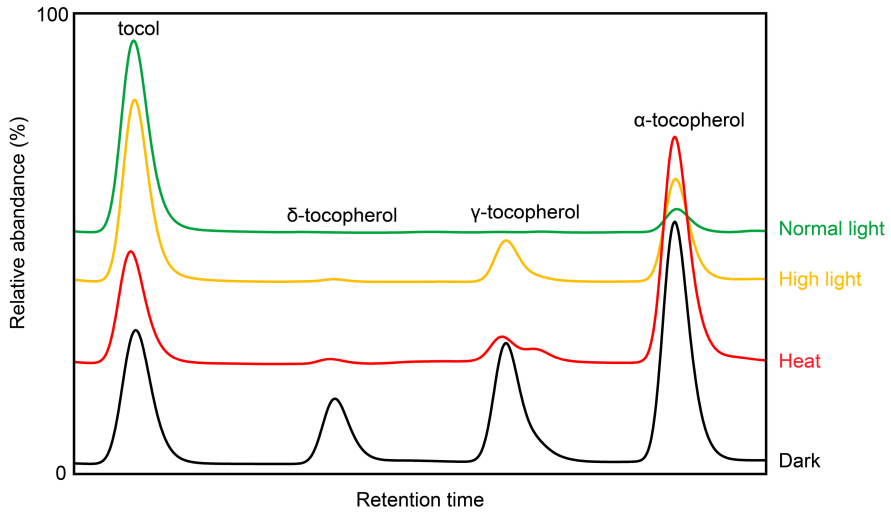


FIGURE 2
Representative HPLC traces of Arabidopsis leaf tocopherols subjected to different growth conditions (high light, heat, and dark). Tocol is used as internal standard [based on Bao et al. (2020), with modification].

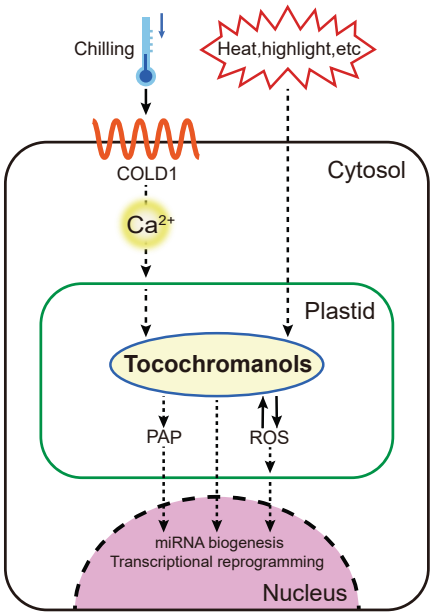


FIGURE 3
A proposed working model of the vitamin E in signal transduction Chilling stress can be sensed by COLD1 to regulate calcium signal and tocochromanol homeostasis for transcriptional reprogramming in the nucleus. Other stresses including high-light and heat can also transduce retrograde signaling between chloroplast and nucleus for miRNA biogenesis, via manipulating PAP (3′-phosphoadenosine-5′-phosphate) and tocochromanols.

maize, compared with high temperature, more tocopherols and tocotrienols were produced under low temperature (Xiang et al., 2019). In addition, tomato *SlVTE5* silenced plants display a strong chlorotic phenotype with low levels of α-tocopherol under the stress combined with high-light and high-temperature (Spicher et al., 2017). Moreover, high-light stress also triggers the expression of *HPPDs* in *Medicago sativa* and *Lactuca sativa* for counteracting and survival strategies (Ren et al., 2011; Jiang et al., 2017).

Drought is one of the most common stresses in limiting farming, which leads to significant yield losses (Godwin and Farrona, 2020). The capacity of HPPDs to resist drought has also been demonstrated in various plant species, including *Lactuca sativa*, Medicago and sweet potato (Ren et al., 2011; Jiang et al., 2017; Kim et al., 2021). In rice, *OsVTE1* is induced to significantly high levels under drought stress (Ouyang et al., 2011), and ectopic overexpression of *AtVTE1* in tobacco can enhance tolerance to drought stress via reducing lipid peroxidation, electrolyte leakage and H_2O_2 content (Liu et al., 2008). Moreover, overexpression of *MsVTE4* increases the levels of both α-tocopherols and total tocopherols in alfalfa, alleviating oxidative damages, leading to higher tolerance to drought stress (Ma et al., 2020).

Soil with unfavorable high level of soluble salts causes salinity stress in plants, limiting crop yield and the area for farming. In tobacco (*Nicotiana tabacum*) *VTE2* silenced plants, total tocopherols decreased 98%, and ion homeostasis was disturbed with sorbitol and methyl viologen treatment (Abbasi et al., 2007). Meanwhile, in tobacco *VTE4* silenced plants, α- and γ-tocopherols were found to play diversified roles in plant stress tolerance (Abbasi et al., 2007). On the other hand, overexpressing *AtVTE4* can reduce superoxide contents, lipid peroxidation and ion leakage under salt stress (Jin and Daniell, 2014). By employing Arabidopsis tocopherol deficient mutants *vte1* (deficient in α- and γ-tocopherols) and *vte4*

(over accumulation of γ-tocopherols), Surówka et al. (2020) revealed that α-tocopherols had stronger regulatory effect than γ-tocopherols through modulating chloroplast biosynthesis pathways and ROS/osmotic-associated compounds under salt stress. Based on the studies above, *AtVTE2* and *AtVTE4* have been engineered in potato breeding for counteracting heavy metal stress (Upadhyaya et al., 2021).

Tocochromanols not only protect plant from abiotic stress, but also contribute to the resistance of biotic stress. *Pseudomonas syringae* can infect a wide range of plant species and leads to serious economic loss (Bohlmann and Keeling, 2008), which is served as a model for dissecting the mechanism of plant-pathogen interactions (Xin et al., 2018). Arabidopsis *vte2* mutants exhibit stronger lipid peroxidation and produce less SA (salicylic acid, a key phytohormone in transducing pathogen signal), resulting in susceptibility to *P. syringae* (Stahl et al., 2019). As a necrotrophic fungus, *Botrytis cinereal* infects many important plant species, including Arabidopsis (Staats et al., 2005). Through detailed analysis and comparison of lipid and hormone changes in *B. cinereal*-infected Arabidopsis *vte1* and *vte4* mutants, Cela et al. found that altered tocopherol compositions could reduce plant tolerance and delay the activation of defense pathway (Cela et al., 2018).

Vitamin E in signal transduction

Vitamin E has long been assessed and studied as an antioxidant, but emerging evidences strongly suggested that vitamin E may also serve as signaling molecules in plants. Reactive oxygen species (ROS) such hydroperoxide and single oxygen that produced in the photosystem have been shown as important signals in the communications between chloroplast and gene expression in the nucleus (**Figure 3**) (Foyer and Noctor, 2003, 2005). When single oxygen accumulated in the chloroplast, tocopherols will be oxidized to produce tocopheryl-radical and hydroperoxide. Reversibly, these two products can be reduced to tocopherols by introducing ascorbate (also known as vitamin C) (Neely et al., 1988). Thus, through eliminating ROS, vitamin E maintains chloroplast redox state and modulates retrograde signaling from the chloroplast to nucleus (**Figure 3**) (Krieger-Liszkay and Trebst, 2006).

Cross-talks between vitamin E and phytohormones were also revealed. For instance, tocopherol deficient *vte1* mutants accumulate more jasmonic acid (JA) and anthocyanin than that of wide-type, causing growth retardation in both high-light and low temperature conditions (Munné-Bosch et al., 2007). During low phosphate (Pi) treatment, both *vte1* and *vte4* mutants accumulate more JA and SA than that in the wide-type. In addition, the expression levels of over 500 transcription factors are significantly affected in *vte1* and *vte4* plants, indicating that tocopherols are involved in phytohormone signaling and transcriptional reprogramming (Allu et al., 2017). In another case, ethylene-responsive *cis* elements were found in the promoter region of Mango (*Mangifera indica*) *MiHPPD* gene. During fruit ripening and leaf senescence, elevated endogenous ethylene can induce *MiHPPD* expression, giving rise to tocopherols contents (Singh et al., 2011). Overall, interactions and communications between vitamin E and phytohormones contribute to plant environment adaptations via fine-tuning downstream gene expressions.

The quantitative trait locus COLD1 was identified recently in japonica rice that can confer tolerance to chilling stress. For sensing low temperature, COLD1 can interact with RGA1 (G-protein α subunit) to activate the Ca^{2+} channel and accelerate GTPase activity of G-protein. Influx of calcium, an important intracellular second messenger, activates gene transcription of vitamin E and vitamin K1 biosynthesis pathways, promoting plant cold tolerance (Luo et al., 2021). Another tocopherol-mediated chloroplast-to-nucleus signaling event was discovered during the study of heat stress-associated microRNA (miRNA) biogenesis. The metabolite 3′-phosphoadenosine-5′-phosphate (PAP) has been broadly racialized as crucial secondary messenger in plant stress responding. Heat stress promotes accumulation of tocopherols, and in turn produces more PAP, which inhibits the nuclear exoribonucleases (XRN), stimulating the biogenesis of microRNAs including miR398 to enhance plant heat tolerance (Fang et al., 2019).

Studies in human and animals suggested that tocopherol-binding protein (TBP) is important for the distribution and transport of α-tocopherol among different tissues. In the latest research, Bermúdez et al. identified the SlTBP (*Solanum lycoperisicum* tocopherol-binding protein) as a homolog of the human α-tocopherol transfer protein (HsTTP). *In vitro* biochemical assay suggested that SlTBP possesses α-tocopherol binding ability. SlTBP is chloroplast-targeted, and knocking down *SlTBP* expression in tomato confers disorders in tocopherol, carotenoid and lipid compositions. Finding of TBP in plants sheds light on understanding vitamin E transport, implying its potential as signaling molecules (Bermúdez et al., 2018).

Future perspectives

Studies in Arabidopsis indicated that 90% and more of the total tocopherols (∼5 ng/mg fresh weight) in leaves are α-tocopherol, while in Arabidopsis seeds, tocopherols (∼370 ng/mg dry seed) are dominated by the γ isoform. Crucial physiological functions of vitamin E for plants were exemplified by the observation that tocopherol defective Arabidopsis mutants were severely affected in seed longevity, germination and seeding growth (Sattler et al., 2004). Differential regulation of the same cassette of genes for tocochromanol biosynthesis in different tissues warrants for future explorations. Chlorophyll and its relevant derivatives are the most abundant pigments

in green plants, and accumulated evidences suggested that chlorophyll-derived phytol groups are the main source for vitamin E biosynthesis (Gutbrod et al., 2019). VTE7 is a novel alpha/beta hydrolase that fits in the missing gap between chlorophyll metabolism and vitamin E production, accounting for more than 50% of total tocopherol biosynthesis in seeds. However, its exact targets still need to be verified. In recent years, tocopherols were found to be involved in both cold response and PAP-mediated retrograde signal transduction. In addition, tocopherol binding protein (TBP) was identified in tomato. Thus, role of tocopherols in acting as signaling transducers are promising and deserved to be investigated in depth. More importantly, as an essential nutrient, engineering balanced vitamin E in crops like soybean, rapeseed will advance plant breeding and benefit human health.

Author contributions

YB concepted the topic of this manuscript and revised the manuscript. YN and QZ drafted the manuscript with YB. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by China Agriculture Research System of MOF and MARA.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abbasi, A. R., Hajirezaei, M., Hofius, D., Sonnewald, U., and Voll, L. M. (2007). Specific roles of alpha- and gamma-tocopherol in abiotic stress responses of transgenic tobacco. *Plant Physiol.* 143, 1720–1738. doi: 10.1104/pp.106.094771

Albert, E., Kim, S., Magallanes-Lundback, M., Bao, Y., Deason, N., Danilo, B., et al. (2022). Genome-wide association identifies a missing hydrolase for tocopherol synthesis in plants. *Proc. Natl. Acad. Sci. U.S.A.* 119:e2113488119. doi: 10.1073/pnas.2113488119

Allu, A. D., Simancas, B., Balazadeh, S., and Munné-Bosch, S. (2017). Defense-related transcriptional reprogramming in Vitamin E-Deficient *Arabidopsis* mutants exposed to contrasting phosphate availability. *Front. Plant Sci.* 8:20. doi: 10.3389/fpls.2017.01396

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. doi: 10.1016/j.cell.2016.05.063

Armstead, I., Donnison, I., Aubry, S., Harper, J., Hortensteiner, S., James, C., et al. (2007). Cross-species identification of Mendel's/locus. *Science* 315, 73–73. doi: 10.1126/science.1132912

Bao, Y., Magallenes-Lundback, M., Deason, N., and DellaPenna, D. (2020). High throughput profiling of tocochromanols in leaves and seeds of *Arabidopsis* and Maize. *Plant Methods* 16:14. doi: 10.1186/s13007-020-00671-9

Bergmüller, E., Porfirova, S., and Dormann, P. (2003). Characterization of an *Arabidopsis* mutant deficient in gamma-tocopherol methyltransferase. *Plant Mol. Biol.* 52, 1181–1190. doi: 10.1023/b:Plan.0000004307.62398.91

Bermúdez, L., del Pozo, T., Lira, B. S., de Godoy, F., Boos, I., Romano, C., et al. (2018). A tomato tocopherol-binding protein sheds light on intracellular alpha-tocopherol metabolism in plants. *Plant Cell Physiol.* 59, 2188–2203. doi: 10.1093/pcp/pcy191

Bohlmann, J., and Keeling, C. I. (2008). Terpenoid biomaterials. *Plant J.* 54, 656–669. doi: 10.1111/j.1365-313X.2008.03449.x

Butinar, B., Bucar-Miklavcic, M., Mariani, C., and Raspor, P. (2011). New vitamin E isomers (gamma-tocomonoenol and alpha-tocomonoenol) in seeds, roasted seeds and roasted seed oil from the Slovenian pumpkin variety 'Slovenska golica'. *Food Chem.* 128, 505–512. doi: 10.1016/j.foodchem.2011.03.072

Cahoon, E. B., Hall, S. E., Ripp, K. G., Ganzke, T. S., Hitz, W. D., and Coughlan, S. J. (2003). Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. *Nat. Biotechnol.* 21, 1082–1087. doi: 10.1038/nbt853

Cela, J., Tweed, J. K. S., Sivakumaran, A., Lee, M. R. F., Mur, L. A. J., and Munné-Bosch, S. (2018). An altered tocopherol composition in chloroplasts reduces plant resistance to *Botrytis cinerea*. *Plant Physiol. Biochem.* 127, 200–210. doi: 10.1016/j.plaphy.2018.03.033

Chen, M. C. M., Yang, J. H., Liu, C. H., Lin, K. H., and Yang, C. M. (2014). Molecular, structural, and phylogenetic characterization of two chlorophyllase isoforms in *Pachira macrocarpa*. *Plant Syst. Evol.* 300, 633–643. doi: 10.1007/s00606-013-0908-5

Cheng, W. Y., Kuo, Y. H., and Huang, C. J. (2007). Isolation and identification of novel estrogenic compounds in yam tuber (*Dioscorea alata* cv. Tainung No. 2). *J. Agric. Food Chem.* 55, 7350–7358. doi: 10.1021/jf0711690

Cheng, Z. G., Sattler, S., Maeda, H., Sakuragi, Y., Bryant, D. A., and DellaPenna, D. (2003). Highly divergent methyltransferases catalyze a conserved reaction in tocopherol and plastoquinone synthesis in cyanobacteria and photosynthetic eukaryotes. *Plant Cell* 15, 2343–2356. doi: 10.1105/tpc.013656

Collakova, E., and DellaPenna, D. (2001). Isolation and functional analysis of homogentisate phytyltransferase from *Synechocystis* sp PCC 6803 and arabidopsis. *Plant Physiol.* 127, 1113–1124. doi: 10.1104/pp.010421

Diepenbrock, C. H., Kandianis, C. B., Lipka, A. E., Magallanes-Lundback, M., Vaillancourt, B., Gongora-Castillo, E., et al. (2017). Novel loci underlie natural variation in vitamin E levels in maize grain. *Plant Cell* 29, 2374–2392. doi: 10.1105/tpc.17.00475

Falk, J., and Munné-Bosch, S. (2010). Tocochromanol functions in plants: Antioxidation and beyond. *J. Exp. Bot.* 61, 1549–1566. doi: 10.1093/jxb/erq030

Fang, X. F., Zhao, G. Z., Zhang, S., Li, Y. X., Gu, H. Q., Li, Y., et al. (2019). Chloroplast-to-nucleus signaling regulates microRNA biogenesis in *Arabidopsis*. *Dev. Cell* 48, 371.e–382.e. doi: 10.1016/j.devcel.2018.11.046

Fiedler, E., Soll, J., and Schultz, G. (1982). The formation of homogentisate in the biosynthesis of tocopherol and plastoquinone in spinach chloroplasts. *Planta* 155, 511–515. doi: 10.1007/bf01607575

Foyer, C. H., and Noctor, G. (2003). Redox sensing and signalling associated with reactive oxygen in chloroplasts, peroxisomes and mitochondria. *Physiol. Plant.* 119, 355–364. doi: 10.1034/j.1399-3054.2003.00223.x

Foyer, C. H., and Noctor, G. (2005). Redox homeostasis and antioxidant signaling: A metabolic interface between stress perception and physiological responses. *Plant Cell* 17, 1866–1875. doi: 10.1105/tpc.105.033589

Garcia, I., Rodgers, M., Lenne, C., Rolland, A., Sailland, A., and Matringe, M. (1997). Subcellular localization and purification of a p-hydroxyphenylpyruvate dioxygenase from cultured carrot cells and characterization of the corresponding cDNA. *Biochem. J.* 325, 761–769. doi: 10.1042/bj3250761

Godwin, J., and Farrona, S. (2020). Plant epigenetic stress memory induced by drought: A physiological and molecular perspective. *Methods Mol. Biol.* 2093, 243–259. doi: 10.1007/978-1-0716-0179-2_17

Goffman, F. D., and Mollers, C. (2000). Changes in tocopherol and plastochromanol-8 contents in seeds and oil of oilseed rape (*Brassica napus* L.) during storage as influenced by temperature and air oxygen. *J. Agric. Food Chem.* 48, 1605–1609. doi: 10.1021/jf9912755

Gutbrod, K., Romer, J., and Dörmann, P. (2019). Phytol metabolism in plants. *Prog. Lipid Res.* 74, 1–17. doi: 10.1016/j.plipres.2019.01.002

Hammann, S., Englert, M., Muller, M., and Vetter, W. (2015). Accelerated separation of GC-amenable lipid classes in plant oils by countercurrent chromatography in the co-current mode. *Anal. Bioanal. Chem.* 407, 9019–9028. doi: 10.1007/s00216-015-9068-5

Havaux, M., Eymery, F., Porfirova, S., Rey, P., and Dormann, P. (2005). Vitamin E protects against photoinhibition and photooxidative stress in *Arabidopsis thaliana*. *Plant Cell* 17, 3451–3469. doi: 10.1105/tpc.105.037036

Hershberger, J., Tanaka, R., Wood, J. C., Kaczmar, N., Wu, D., Hamilton, J. P., et al. (2022). Transcriptome-wide association and prediction for carotenoids and tocochromanols in fresh sweet corn kernels. *Plant Genome* 15:16. doi: 10.1002/tpg2.20197

Hofius, D., Hajirezaei, M. R., Geiger, M., Tschiersch, H., Melzer, M., and Sonnewald, U. (2004). RNAi-mediated tocopherol deficiency impairs photoassimilate export in transgenic potato plants. *Plant Physiol.* 135, 1256–1268. doi: 10.1104/pp.104.043927

Jacob-Wilk, D., Holland, D., Goldschmidt, E. E., Riov, J., and Eyal, Y. (1999). Chlorophyll breakdown by chlorophyllase: Isolation and functional expression of the Chlase1 gene from ethylene-treated Citrus fruit and its regulation during development. *Plant J.* 20, 653–661. doi: 10.1046/j.1365-313X.1999.00637.x

Jiang, J. S., Chen, Z. H., Ban, L. P., Wu, Y. D., Huang, J. P., Chu, J. F., et al. (2017). P-HYDROXYPHENYLPYRUVATE DIOXYGENASE from *Medicago sativa* is involved in vitamin E biosynthesis and abscisic acidmediated seed germination. *Sci. Rep.* 7:15. doi: 10.1038/srep40625

Jin, S. X., and Daniell, H. (2014). Expression of gamma-tocopherol methyltransferase in chloroplasts results in massive proliferation of the inner envelope membrane and decreases susceptibility to salt and metal-induced oxidative stresses by reducing reactive oxygen species. *Plant Biotechnol. J.* 12, 1274–1285. doi: 10.1111/pbi.12224

Kamal-Eldin, A., and Appelqvist, L.-A. (1996). The chemistry and antioxidant properties of tocopherols and tocotrienols. *Lipids* 31, 671–701. doi: 10.1007/bf02522884

Kim, S. E., Bian, X. F., Lee, C. J., Park, S. U., Lim, Y. H., Kim, B. H., et al. (2021). Overexpression of 4-hydroxyphenylpyruvate dioxygenase (IbHPPD) increases abiotic stress tolerance in transgenic sweetpotato plants. *Plant Physiol. Biochem.* 167, 420–429. doi: 10.1016/j.plaphy.2021.08.025

Kobayashi, N., and DellaPenna, D. (2008). Tocopherol metabolism, oxidation and recycling under high light stress in Arabidopsis. *Plant J.* 55, 607–618. doi: 10.1111/j.1365-313X.2008.03539.x

Krieger-Liszkay, A., and Trebst, A. (2006). Tocopherol is the scavenger of singlet oxygen produced by the triplet states of chlorophyll in the PSII reaction centre. *J. Exp. Bot.* 57, 1677–1684. doi: 10.1093/jxb/erl002

Kruk, J., Hollander-Czytko, H., Oettmeier, W., and Trebst, A. (2005). Tocopherol as singlet oxygen scavenger in photosystem II. *J. Plant Physiol.* 162, 749–757. doi: 10.1016/j.jplph.2005.04.020

Kruk, J., Szymanska, R., Cela, J., and Munné-Bosch, S. (2014). Plastochromanol-8: Fifty years of research. *Phytochemistry* 108, 9–16. doi: 10.1016/j.phytochem.2014.09.011

Lin, Y. P., and Charng, Y. Y. (2017). Supraoptimal activity of CHLOROPHYLL DEPHYTYLASE1 results in an increase in tocopherol level in mature arabidopsis seeds. *Plant Signal. Behav.* 12:3. doi: 10.1080/15592324.2017.1382797

Lin, Y. P., Wu, M. C., and Charng, Y. Y. (2016). Identification of a chlorophyll dephytylase involved in chlorophyll turnover in *Arabidopsis*. *Plant Cell* 28, 2974–2990. doi: 10.1105/tpc.16.00478

Liu, X. L., Hua, X. J., Guo, J., Qi, D. M., Wang, L. J., Liu, Z. P., et al. (2008). Enhanced tolerance to drought stress in transgenic tobacco plants overexpressing VTE1 for increased tocopherol production from *Arabidopsis thaliana*. *Biotechnol. Lett.* 30, 1275–1280. doi: 10.1007/s10529-008-9672-y

Löffelhardt, W., and Kindl, H. (1979). Conversion of 4-hydroxyphenylpyruvic acid into homogentisic acid at the thylakoid membrane of *Lemna gibba*. *FEBS Lett.* 104, 332–334. doi: 10.1016/0014-5793(79)80845-5

Luo, W., Huan, Q., Xu, Y. Y., Qian, W. F., Chong, K., and Zhang, J. Y. (2021). Integrated global analysis reveals a vitamin E-vitamin K1 sub-network, downstream of COLD1, underlying rice chilling tolerance divergence. *Cell Rep.* 36:17. doi: 10.1016/j.celrep.2021.109397

Ma, J. T., Qiu, D. Y., Gao, H. W., Wen, H. Y., Wu, Y. D., Pang, Y. Z., et al. (2020). Over-expression of a gamma-tocopherol methyltransferase gene in vitamin E pathway confers PEG-simulated drought tolerance in alfalfa. *BMC Plant Biol.* 20:16. doi: 10.1186/s12870-020-02424-1

Maeda, H., Song, W., Sage, T. L., and DellaPenna, D. (2006). Tocopherols play a crucial role in low-temperature adaptation and phloem loading in *Arabidopsis*. *Plant Cell* 18, 2710–2732. doi: 10.1105/tpc.105.039404

Matsumoto, A., Takahashi, S., Nakano, K., and Kijima, S. (1995). Identification of new vitamin E in plant oil. *J. Am. Oil. Chem. Soc.* 44, 593–597.

Mène-Saffrané, L. (2018). Vitamin E biosynthesis and its regulation in plants. *Antioxidants* 7:17. doi: 10.3390/antiox7010002

Munné-Bosch, S., Weiler, E. W., Alegre, L., Muller, M., Duchting, P., and Falk, J. (2007). alpha-Tocopherol may influence cellular signaling by modulating jasmonic acid levels in plants. *Planta* 225, 681–691. doi: 10.1007/s00425-006-0375-0

Neely, W. C., Martin, J. M., and Barker, S. A. (1988). Products and relative reaction rates of the oxidation of tocopherols with singlet molecular oxygen. *Photochem. Photobiol.* 48, 423–428. doi: 10.1111/j.1751-1097.1988.tb02840.x

Ouyang, S. Q., He, S. J., Liu, P., Zhang, W. K., Zhang, J. S., and Chen, S. Y. (2011). The role of tocopherol cyclase in salt stress tolerance of rice (*Oryza sativa*). *Sci. China Life Sci.* 54, 181–188. doi: 10.1007/s11427-011-4138-1

Pellaud, S., and Mène-Saffrané, L. (2017). Metabolic origins and transport of vitamin E biosynthetic precursors. *Front. Plant Sci.* 8:1959. doi: 10.3389/fpls.2017.01959

Pellaud, S., Bory, A., Chabert, V., Romanens, J., Chaisse-Leal, L., Doan, A. V., et al. (2018). WRINKLED1 and ACYL-COA:DIACYLGLYCEROL ACYLTRANSFERASE1 regulate tocochromanol metabolism in *Arabidopsis*. *New Phytol.* 217, 245–260. doi: 10.1111/nph.14856

Pretty, J., Sutherland, W. J., Ashby, J., Auburn, J., Baulcombe, D., Bell, M., et al. (2010). The top 100 questions of importance to the future of global agriculture. *Int. J. Agric. Sustain.* 8, 219–236. doi: 10.3763/ijas.2010.0534

Ren, W. W., Zhao, L. X., Zhang, L. D., Wang, Y. L., Cui, L. J., Tang, Y. L., et al. (2011). Molecular analysis of a homogentisate phytyltransferase gene from *Lactuca sativa* L. *Mol. Biol. Rep.* 38, 1813–1819. doi: 10.1007/s11033-010-0297-6

Riewe, D., Koohi, M., Lisec, J., Pfeiffer, M., Lippmann, R., Schmeichel, J., et al. (2012). A tyrosine aminotransferase involved in tocopherol synthesis in *Arabidopsis*. *Plant J.* 71, 850–859. doi: 10.1111/j.1365-313X.2012.05035.x

Russin, W. A., Evert, R. F., Vanderveer, P. J., Sharkey, T. D., and Briggs, S. P. (1996). Modification of a specific class of plasmodesmata and loss of sucrose export ability in the sucrose export defective1 maize mutant. *Plant Cell* 8, 645–658. doi: 10.1105/tpc.8.4.645

Sadre, R., Frentzen, M., Saeed, M., and Hawkes, T. (2010). Catalytic reactions of the homogentisate prenyl transferase involved in plastoquinone-9 biosynthesis. *J. Biol. Chem.* 285, 18191–18198. doi: 10.1074/jbc.M110.117929

Sadre, R., Gruber, J., and Frentzen, M. (2006). Characterization of homogentisate prenyltransferases involved in plastoquinone-9 and tocochromanol biosynthesis. *FEBS Lett.* 580, 5357–5362. doi: 10.1016/j.febslet.2006.09.002

Sato, Y., Morita, R., Nishimura, M., Yamaguchi, H., and Kusaba, M. (2007). Mendel's green cotyledon gene encodes a positive regulator of the chlorophyll-degrading pathway. *Proc. Natl. Acad. Sci. U.S.A.* 104, 14169–14174. doi: 10.1073/pnas.0705521104

Sattler, S. E., Gilliland, L. U., Magallanes-Lundback, M., Pollard, M., and DellaPenna, D. (2004). Vitamin E is essential for seed longevity, and for preventing lipid peroxidation during germination. *Plant Cell* 16, 1419–1432. doi: 10.1105/tpc.021360

Savidge, B., Weiss, J. D., Wong, Y. H. H., Lassner, M. W., Mitsky, T. A., Shewmaker, C. K., et al. (2002). Isolation and characterization of homogentisate phytyltransferase genes from *Synechocystis* sp PCC 6803 and *Arabidopsis*. *Plant Physiol.* 129, 321–332. doi: 10.1104/pp.010747

Schelbert, S., Aubry, S., Burla, B., Agne, B., Kessler, F., Krupinska, K., et al. (2009). Pheophytin pheophorbide hydrolase (Pheophytinase) is involved in chlorophyll breakdown during leaf senescence in Arabidopsis. *Plant Cell* 21, 767–785. doi: 10.1105/tpc.108.064089

Sen, C. K., Khanna, S., and Roy, S. (2007). Tocotrienols in health and disease: The other half of the natural vitamin E family. *Mol. Aspects Med.* 28, 692–728. doi: 10.1016/j.mam.2007.03.001

Shintani, D., and DellaPenna, D. (1998). Elevating the vitamin E content of plants through metabolic engineering. *Science* 282, 2098–2100. doi: 10.1126/science.282.5396.2098

Siehl, D. L., Tao, Y. M., Albert, H., Dong, Y. X., Heckert, M., Madrigal, A., et al. (2014). Broad 4-hydroxyphenylpyruvate dioxygenase inhibitor herbicide tolerance in soybean with an optimized enzyme and expression cassette. *Plant Physiol.* 166, 1162–1176. doi: 10.1104/pp.114.247205

Singh, R. K., Ali, S. A., Nath, P., and Sane, V. A. (2011). Activation of ethylene-responsive p-hydroxyphenylpyruvate dioxygenase leads to increased tocopherol levels during ripening in mango. *J. Exp. Bot.* 62, 3375–3385. doi: 10.1093/jxb/err006

Sozen, E., Demirel, T., and Ozer, N. K. (2019). Vitamin E: Regulatory role in the cardiovascular system. *IUBMB Life* 71, 507–515. doi: 10.1002/iub.2020

Spicher, L., Almeida, J., Gutbrod, K., Pipitone, R., Dormann, P., Glauser, G., et al. (2017). Essential role for phytol kinase and tocopherol in tolerance to combined light and temperature stress in tomato. *J. Exp. Bot.* 68, 5845–5856. doi: 10.1093/jxb/erx356

Staats, M., van Baarlen, P., and van Kan, J. A. L. (2005). Molecular phylogeny of the plant pathogenic genus *Botrytis* and the evolution of host specificity. *Mol. Biol. Evol.* 22, 333–346. doi: 10.1093/molbev/msi020

Stacey, M. G., Cahoon, R. E., Nguyen, H. T., Cui, Y. Y., Sato, S., Nguyen, C. T., et al. (2016). Identification of homogentisate dioxygenase as a target for vitamin E biofortification in oilseeds. *Plant Physiol.* 172, 1506–1518. doi: 10.1104/pp.16.00941

Stahl, E., Hartmann, M., Scholten, N., and Zeier, J. (2019). A role for tocopherol biosynthesis in *Arabidopsis* basal immunity to bacterial infection. *Plant Physiol.* 181, 1008–1028. doi: 10.1104/pp.19.00618

Surówka, E., Potocka, I., Dziurka, M., Wrobel-Marek, J., Kurczynska, E., Zur, I., et al. (2020). Tocopherols mutual balance is a key player for maintaining *Arabidopsis thaliana* growth under salt stress. *Plant Physiol. Biochem.* 156, 369–383. doi: 10.1016/j.plaphy.2020.09.008

Suzuki, Y. J., Tsuchiya, M., Wassall, S. R., Choo, Y. M., Govil, G., Kagan, V. E., et al. (1993). Structural and dynamic membrane properties of alpha-tocopherol and alpha-tocotrienol: Implication to the molecular mechanism of their antioxidant potency. *Biochemistry* 32, 10692–10699. doi: 10.1021/bi00091a020

Tian, L., DellaPenna, D., and Dixon, R. A. (2007). The *pds2* mutation is a lesion in the Arabidopsis homogentisate solanesyltransferase gene involved in plastoquinone biosynthesis. *Planta* 226, 1067–1073. doi: 10.1007/s00425-007-0564-5

Tian, Y. N., Zhong, R. H., Wei, J. B., Luo, H. H., Eyal, Y., Jin, H. L., et al. (2021). *Arabidopsis* CHLOROPHYLLASE 1 protects young leaves from long-term photodamage by facilitating FtsH-mediated D1 degradation in photosystem II repair. *Mol. Plant* 14, 1149–1167. doi: 10.1016/j.molp.2021.04.006

Tsegaye, Y., Shintani, D. K., and DellaPenna, D. (2002). Overexpression of the enzyme p-hydroxyphenolpyruvate dioxygenase in Arabidopsis and its relation to tocopherol biosynthesis. *Plant Physiol. Biochem.* 40, 913–920. doi: 10.1016/s0981-9428(02)01461-4

Tsuchiya, T., Ohta, H., Okawa, K., Iwamatsu, A., Shimada, H., Masuda, T., et al. (1999). Cloning of chlorophyllase, the key enzyme in chlorophyll degradation: Finding of a lipase motif and the induction by methyl jasmonate. *Proc. Natl. Acad. Sci. U.S.A.* 96, 15362–15367. doi: 10.1073/pnas.96.26.15362

Upadhyaya, D. C., Bagri, D. S., Upadhyaya, C. P., Kumar, A., Thiruvengadam, M., and Jain, S. K. (2021). Genetic engineering of potato (*Solanum tuberosum* L.) for enhanced alpha-tocopherols and abiotic stress tolerance. *Physiol. Plant* 173, 116–128. doi: 10.1111/ppl.13252

Valentin, H. E., Lincoln, K., Moshiri, F., Jensen, P. K., Qi, Q. G., Venkatesh, T. V., et al. (2006). The *Arabidopsis* vitamin E pathway gene5-1 mutant reveals a critical role for phytol kinase in seed tocopherol biosynthesis. *Plant Cell* 18, 212–224. doi: 10.1105/tpc.105.037077

Van Eenennaam, A. L., Lincoln, K., Durrett, T. P., Valentin, H. E., Shewmaker, C. K., Thorne, G. M., et al. (2003). Engineering vitamin E content: From *Arabidopsis* mutant to soy oil. *Plant Cell* 15, 3007–3019. doi: 10.1105/tpc.015875

Vom Dorp, K., Holzl, G., Plohmann, C., Eisenhut, M., Abraham, M., Weber, A. P. M., et al. (2015). Remobilization of phytol from chlorophyll degradation is essential for tocopherol synthesis and growth of Arabidopsis. *Plant Cell* 27, 2846–2859. doi: 10.1105/tpc.15.00395

Wang, H., Xu, S. T., Fan, Y. M., Liu, N. N., Zhan, W., Liu, H. J., et al. (2018). Beyond pathways: Genetic dissection of tocopherol content in maize kernels by combining linkage and association analyses. *Plant Biotechno. J.* 16, 1464–1475. doi: 10.1111/pbi.12889

Wang, M. M., Toda, K., and Maeda, H. A. (2016). Biochemical properties and subcellular localization of tyrosine aminotransferases in *Arabidopsis thaliana*. *Phytochemistry* 132, 16–25. doi: 10.1016/j.phytochem.2016.09.007

Wang, M. M., Toda, K., Block, A., and Maeda, H. A. (2019). TAT1 and TAT2 tyrosine aminotransferases have both distinct and shared functions in tyrosine metabolism and degradation in *Arabidopsis thaliana*. *J. Biol. Chem.* 294, 3563–3576. doi: 10.1074/jbc.RA118.006539

Whittle, K. J., Dunphy, P. J., and Pennock, J. F. (1965). Plastochromanol in the leaves of Hevea brasiliensis. *Biochem. J.* 96, 17C–19C. doi: 10.1042/bj0960017C

Wu, D., Li, X. W., Tanaka, R., Wood, J. C., Tibbs-Cortes, L. E., Magallanes-Lundback, M., et al. (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *Genetics* 14:iyac091. doi: 10.1093/genetics/iyac091

Xiang, N., Li, C. Y., Li, G. K., Yu, Y. T., Hu, J. G., and Guo, X. B. (2019). Comparative evaluation on vitamin E and carotenoid accumulation in sweet corn (*Zea mays* L.) Seedlings under temperature stress. *J. Agric. Food Chem.* 67, 9772–9781. doi: 10.1021/acs.jafc.9b04452

Xin, X. F., Kvitko, B., and He, S. Y. (2018). *Pseudomonas* syringae: What it takes to be a pathogen. *Nat. Rev. Microbiol.* 16, 316–328. doi: 10.1038/nrmicro.2018.17

Yang, W. Y., Cahoon, R. E., Hunter, S. C., Zhang, C. Y., Han, J. X., Borgschulte, T., et al. (2011). Vitamin E biosynthesis: Functional characterization of the monocot homogentisate geranylgeranyl transferase. *Plant J.* 65, 206–217. doi: 10.1111/j.1365-313X.2010.04417.x

Zbierzak, A. M., Kanwischer, M., Wille, C., Vidi, P. A., Giavalisco, P., Lohmann, A., et al. (2010). Intersection of the tocopherol and plastoquinol metabolic pathways at the plastoglobule. *Biochem. J.* 425, 389–399. doi: 10.1042/bj20090704

Zhang, C. Y., Zhang, W., Ren, G. D., Li, D. L., Cahoon, R. E., Chen, M., et al. (2015). Chlorophyll synthase under epigenetic surveillance is critical for vitamin E synthesis, and altered expression affects tocopherol levels in *Arabidopsis*. *Plant Physiol.* 168, 1503–U1729. doi: 10.1104/pp.15.00594

Zhang, W., Liu, T. Q., Ren, G. D., Hortensteiner, S., Zhou, Y. M., Cahoon, E. B., et al. (2014). Chlorophyll degradation: The tocopherol biosynthesis-related phytol hydrolase in Arabidopsis seeds is still missing. *Plant Physiol.* 166, 70–79. doi: 10.1104/pp.114.243709

Zingg, J. M. (2007). Modulation of signal transduction by vitamin E. *Mol. Aspects Med.* 28, 481–506. doi: 10.1016/j.mam.2006.12.009

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fpls.2022.998985

OPEN ACCESS

EDITED BY
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Praveen Guleria,
DAV University, India
Qinggang Yin,
Institute of Chinese Materia Medica (CACMS), China

*CORRESPONDENCE
Xian Li
xianli@zju.edu.cn

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 20 July 2022
ACCEPTED 29 August 2022
PUBLISHED 26 September 2022

CITATION
Ren C, Cao Y, Xing M, Guo Y, Li J, Xue L, Sun C, Xu C, Chen K and Li X (2022) Genome-wide analysis of UDP-glycosyltransferase gene family and identification of members involved in flavonoid glucosylation in Chinese bayberry (*Morella rubra*). *Front. Plant Sci.* 13:998985. doi: 10.3389/fpls.2022.998985

Genome-wide analysis of UDP-glycosyltransferase gene family and identification of members involved in flavonoid glucosylation in Chinese bayberry (*Morella rubra*)

Chuanhong Ren[1,2], Yunlin Cao[1,2], Mengyun Xing[1,2], Yan Guo[1,2], Jiajia Li[1,2], Lei Xue[1,2], Chongde Sun[1,2], Changjie Xu[1,2], Kunsong Chen[1,2] and Xian Li[1,2]*

[1]Zhejiang Provincial Key Laboratory of Horticultural Plant Integrative Biology, Zhejiang University, Hangzhou, China, [2]The State Agriculture Ministry Laboratory of Horticultural Plant Growth, Development and Quality Improvement, Zhejiang University, Hangzhou, China

Glycosylation was catalyzed by UDP-glycosyltransferase (UGT) and was important for enriching diversity of flavonoids. Chinese bayberry (*Morella rubra*) has significant nutritional and medical values because of diverse natural flavonoid glycosides. However, information of *UGT* gene family was quite limited in *M. rubra*. In the present study, a total of 152 *MrUGT* genes clustered into 13 groups were identified in *M. rubra* genome. Among them, 139 *MrUGT* genes were marked on eight chromosomes and 13 members located on unmapped scaffolds. Gene duplication analysis indicated that expansion of *MrUGT* gene family was mainly forced by tandem and proximal duplication events. Gene expression patterns in different tissues and under UV-B treatment were analyzed by transcriptome. Cyanidin 3-*O*-glucoside (C3Glc) and quercetin 3-*O*-glucoside (Q3Glc) were two main flavonoid glucosides accumulated in *M. rubra*. UV-B treatment significantly induced C3Glc and Q3Glc accumulation in fruit. Based on comprehensively analysis of transcriptomic data and phylogenetic homology together with flavonoid accumulation patterns, MrUFGT (MrUGT78A26) and MrUGT72B67 were identified as UDP-glucosyltransferases. MrUFGT was mainly involved in C3Glc and Q3Glc accumulation in fruit, while MrUGT72B67 was mainly involved in Q3Glc accumulation in leaves and flowers. Gln375 and Gln391 were identified as important amino acids for glucosyl transfer activity of MrUFGT and MrUGT72B67 by site-directed mutagenesis, respectively.

Transient expression in *Nicotiana benthamiana* tested the function of MrUFGT and MrUGT72B67 as glucosyltransferases. The present study provided valuable source for identification of functional UGTs involved in secondary metabolites biosynthesis in *M. rubra*.

KEYWORDS

***Morella rubra*, UGT, anthocyanin, flavonol, UDP-glucosyltransferase**

Introduction

Diverse plant secondary metabolites such as flavonoids play important roles in plant development and human health (Yin et al., 2014; Bondonno et al., 2019; Alseekh et al., 2020). Glycosylation usually occurs during later stages in many secondary metabolite biosynthesis pathways. Glycosylation could improve solubility, stability, transferability, and diversity of many plant secondary metabolites like flavonoids (Bowles et al., 2006; Yang et al., 2018; Naeem et al., 2021).

UDP-glycosyltransferase (UGT) family was the largest family in plants among GT super families reported in CAZy[1] database. It catalyzed glycosylation formation of many small molecules, including flavonoids, hormones, and xenobiotics (Vogt and Jones, 2000; Bowles et al., 2006). With the rapid development in bioinformatics and plant genomics, *UGT* gene families have been identified in many plants, from algae *Chlamydomonas reinhardtii* to vascular plants like *Selaginella moellendorffii* and *Prunus persica* (Caputi et al., 2012; Wu et al., 2017). In model plant *Arabidopsis thaliana*, 107 *UGT* members were identified in genome, and were clustered into 14 groups (A-N) based on phylogenetic relationship analysis (Ross et al., 2001). Subsequently, four new phylogenetic groups, named O, P, Q, and R, that were not presented in *Arabidopsis* were discovered in *Malus* × *domestica* (Caputi et al., 2012), *Zea mays* (Li et al., 2014), and *Camellia sinensis* (Cui et al., 2016). Gene family identification facilitates discovery of functional *UGT* genes. CsUGT78A14 and CsUGT78A15 were found to be involved in astringent taste compounds biosynthesis by analysis of *C. sinensis UGT* gene family (Cui et al., 2016). And several UGTs involved in biosynthesis of anti-diabetic plant metabolite Montbretin A were discovered based on *UGT* gene family analysis (Irmisch et al., 2018; Irmisch et al., 2020).

UDP-glycosyltransferase family contains a conserved motif close to C-terminal, named the plant secondary product glycosyltransferase (PSPG) box. Amino acids in PSPG-box were important for glycosyl transfer activity of UGTs (Shao et al., 2005; Offen et al., 2006; Osmani et al., 2009). For example, last amino acid residue of PSPG-box for UDP-glucosyltransferases usually was glutamine (Gln), and examples include VvGT1 (Ford et al., 1998), MdUGT71B1 (Xie et al., 2020), and PpUGT78T3 (Xie et al., 2022). However, other amino acids could also influence UGT sugar donor preference and more UGTs with different functions should be identified to elucidate the mechanism of sugar donor preference of UGTs.

Chinese bayberry (*Morella rubra*), a member of the Myricaceae, has significant nutritional and medical values due to high content of diverse natural flavonoids such as flavonol glycosides and anthocyanins (Sun et al., 2013; Zhang et al., 2015; Liu et al., 2022). It was reported that flavonoid-rich extracts of fruit and leaves had diverse bioactivities such as antioxidant (Sun et al., 2013; Yan et al., 2016), anti-diabetes (Sun et al., 2013; Liu et al., 2020), and anti-cancer (Sun et al., 2012). However, information of *UGT* gene family and identification of UGTs related to flavonoid glycosylation in *M. rubra* were limited. Recently, both transcriptome and genome information with high-quality have been published in *M. rubra* (Feng et al., 2013; Jia et al., 2019), which makes identification of *UGT* gene family in this plant available.

In the present study, a comprehensive genome-wide identification of *UGT* gene family was carried out in *M. rubra*. A total of 152 MrUGT putative proteins were identified from *M. rubra* genome. Genome-wide analysis was performed including phylogenetic relationship, gene structure, chromosome distribution, and gene duplication. Furthermore, expression patterns of *MrUGT* genes were analyzed by Ribonucleic Acid (RNA)-seq in different tissues and ultraviolet (UV) B-treated fruit. Base on *MrUGT* gene family analysis, MrUFGT (MrUGT78A26) and MrUGT72B67 were identified as flavonoid 3-*O*-glucosyltransferases by *in vitro* and *in vivo* investigations. In addition, important amino acids were identified for glucosyl transfer activity of MrUFGT and MrUGT72B67 by site-direct mutagenesis.

Materials and methods

Identification and phylogenetic analysis of *MrUGT* gene family

A Hidden Markov Model (HMM) profile for UGT (PF00201) downloaded from Pfam[2] database was used as a

1 http://www.cazy.org

2 http://pfam.xfam.org

query file to identify UGT proteins in *M. rubra* genome using simple HMM search program in TBtools (Jia et al., 2019; Chen et al., 2020). Multiple EM for Motif Elicitationv (MEME, suite 5.0.3) website and CDD[3] were used to check completeness of MrUGT sequences. Incomplete coding sequences were manually corrected based on RNA-Seq database (PRJNA714192). MrUGT protein sequences and other plant UGTs were aligned with MUSCLE program. Phylogenetic tree was constructed using neighbor-joining method in MEGA-X with 1000 bootstrap replicates. Genbank accession numbers could be found in **Supplementary Table 1**. Multiple sequence alignment was carried out using MUSCLE program between MrUGTs and other glucosyltransferases. Sequence alignment was visualized using GeneDoc software.

Analysis of conserved motif and gene structure

Conserved motifs in MrUGT proteins were analyzed by simple MEME Wrapper program in TBtools with default parameters. Results of conserved motifs were visualized by TBtools (Chen et al., 2020). Sequences of conserved motifs were visualized by WebLogo[4] website (Crooks et al., 2004). Intron-exon map of *MrUGT* was constructed according to genome annotation file. Gene Structure Display Server 2.0[5] was used to investigate intron-exon structure in *MrUGT* gene family using sequence format (Hu et al., 2014).

Chromosome distribution and syntenic analysis of *MrUGT* gene family

Gene Location Visualize program of TBtools was used to investigate and visualize chromosome distribution of *MrUGT* genes according to genome annotation file (Chen et al., 2020). To investigate the evolutionary relationship between MrUGTs and UGTs of other species, synteny analysis was performed within three Rosids species, i.e., *Arabidopsis*, walnut (*Juglans regia*), and peach (*P. persica*). Synteny relationship was analyzed by One Step MCScanX program and visualized by DualSyntePlot program with the help of TBtools (Chen et al., 2020). DupGen_finder program was used to analyze gene duplication events in *M. rubra* genome (Qiao et al., 2019).

Chemicals reagents

Quercetin (Q), kaempferol (K), quercetin 3-*O*-glucoside (Q3Glc), flavanones (naringenin and hesperetin), flavanols (epicatechin and catechin), flavones (apigenin and luteolin), and isoflavones (genistein and daidzein) were purchased from Aladdin (Shanghai, China). Cyanidin (C) and pelargonidin (P) were purchased from Extrasynthese (Lyon, France). Gradient grade for liquid chromatography of methanol and acetonitrile as well as cyanidin 3-*O*-glucoside (C3Glc) were purchased from Sigma-Aldrich (St. Louis, MO, USA). UDP-glucose (UDP-Glc), UDP-rhamnose (UDP-Rha), and UDP-galactose (UDP-Gal) were obtained from Yuanye Bio-Technology Co., Ltd., (Shanghai, China).

Plant materials and ultraviolet-B treatment

Flowers, leaves, and fruit of different development stages of *M. rubra* cv. Biqi were obtained from an orchard in Lanxi (Zhejiang, China). Four fruit development stages were: S1 for 45 days after flowering (DAF); S2 for 75 DAF; S3 for 80 DAF; S4 for 85 DAF. All materials were uniform in size and free from mechanical damage. Samples were cut into small pieces, frozen with liquid nitrogen immediately, and stored at −80°C for further analysis. All samples were collected for three biological replicates.

UV-B treatment was carried out as reported (Xie et al., 2020) with some modifications. Treatments were carried out at different layers in the same climatic chambers under controlled conditions with a relative humidity of 90–96% and constant temperature at 20°C. Fruit of 'Biqi' cultivar at 70 DAF were selected to treated with UV-B irradiation. Fruit were divided into two groups, and one group was exposed to UV-B irradiation (280–315 nm, 50 μW cm^{-2}) for 2 and 6 days. Fruit of control group were put in the dark. Incubator was covered with black cloth to avoid light pollution. Three biological replicates were used and each replicate contained five to eight fruits.

RNA-seq and gene expression

Total RNA was isolated using cetyltrimethylammonium bromide (CTAB) method as reported (Feng et al., 2013). Integrity of total RNA was detected using nanodrop and gel electrophoresis. RNA-Seq of UV-B-treated fruit was carried out by Novogene Technology Co., Ltd. (Beijing, China). RNA-Seq platform was Illumina Novaseq. Library was prepared using NEBNext Ultra RNA Library Prep Kit for Illumina. Gene expression levels were assessed by FPKM values. Different expression analysis was carried out using DESeq2 (1.20.0). Heatmap of transcript profiles was presented by TBtools (Chen et al., 2020). Gene expression was performed by reverse transcription quantitative PCR (RT-qPCR) as reported (Cao et al., 2019) using primers showed in **Supplementary Table 2**. *Actin* gene (*MrACT*, GQ340770) was used as internal reference

3 https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
4 http://weblogo.berkeley.edu/logo.cgi
5 http://gsds.gao-lab.org/index.php

gene. Relative gene expression was calculated using $2^{-\Delta\Delta Ct}$ method.

HPLC analysis of flavonoid glycosides

Flavonoid glycosides were extracted and analyzed as reported (Downey et al., 2007; Cao et al., 2019) with some modifications. Sample powder with 0.1 g was sonicated in 1 ml 50% methanol/water (v/v) for 30 min at room temperature. After centrifugation at 12,000 rpm for 15 min, precipitates were extracted one more time. Both supernatants were combined and then analyzed by high-performance liquid chromatography (HPLC) after centrifugation at 12,000 rpm for 15 min as previous reported (Cao et al., 2019). Standard curves were used to quantitate Q3Glc at 350 nm and C3Glc at 520 nm.

Protein recombination and purification

Coding sequences of *MrUFGT* and *MrUGT72B67* were subcloned into expression vector pET-32a (+) using specific primers listed in **Supplementary Table 3**. Recombination plasmids were transformed into *Escherichia coli* BL21 (DE3) pLysS (Promega, Madison, WI, USA). Protein recombination was carried out as reported with some modifications (Xie et al., 2020). Recombinant proteins were induced by adding 500 μM IPTG and cultured at 16°C for 20–24 h. HisTALON Gravity Columns (Takara Bio Inc., Beijing, China) was used to purify His-tagged proteins according to manual. PD-10 columns (GE Healthcare, UK) was used to desalt of His-tagged proteins. Recombinant proteins were monitored by SDS-PAGE and quantitated by BCA kit (FUDE, Hangzhou, China).

Enzyme assay

Enzymatic activity assay was carried out as reported with some modifications (Ren et al., 2022). Reactions were performed in a total volume of 100 μl mixture containing 0.1 M Tris–HCl buffer (pH 7.5), 1 mM sugar donors (UDP-Glc/UDP-Gal/UDP-Rha), 60 μM sugar acceptors (Q/C), and 1–2 μg recombinant proteins at 30°C for 20 min. Enzyme reactions were stopped by adding 100 μl methanol, and analyzed by HPLC after centrifugation (12,000 rpm for 15 min) as reported (Xie et al., 2020). Enzyme products were detected at 350 nm for flavonol glycosides and at 520 nm for anthocyanins. Enzyme products were analyzed by LC-MS/MS as reported (Ren et al., 2022).

Site-directed mutagenesis analysis

Mutant proteins were generated by overlapping PCR using primers listed in **Supplementary Table 4**. Mutant sequences were confirmed by sequencing. Recombinant mutant proteins were monitored by SDS-PAGE. Reaction for site-directed mutagenesis analysis was carried out as mentioned above, and 1–2 μg recombinant mutated proteins were contained in reaction mixture. Relative activity of mutant enzyme was quantified using HPLC.

Transient expression in *Nicotiana benthamiana*

Transient expression in *N. benthamiana* was performed as reported (Cao et al., 2019). Coding sequences of *MrUFGT* and *MrUGT72B67* were subcloned into pGreenII0029 62-SK (SK) vector. Specific primers were listed in **Supplementary Table 5**. All recombinant plasmids were electroporated into *Agrobacterium tumefaciens* strain GV3101. Bacteria were resuspended in infiltration bufefr (150 μM acetosyringone, 10 mM $MgCl_2$, 10 mM MES, pH 5.6) to OD_{600} of 0.75. Mixtures were prepared according to combination information in **Figure 8A**. Each combination contained *A. tumefaciens* strain p19. Four-week-old *N. benthamiana* leaves were infiltrated with different combination mixtures. Flavonoid glycosides were analyzed by LC-MS/MS after 5 days infiltration as previous reported (Ren et al., 2022). Data were collected from at least three independent *N. benthamiana* plants.

Statistical analysis

One-way ANOVA followed Tukey test was performed to analyze significant differences among different groups at a significance level of 0.05 using DPS 9.01. Two-tailed Student's *t*-test was used to analyze two-sample statistical significance. Experimental data were analyzed and presented by Origin 9.0 (Northampton, MA, USA) and GraphPad Prism 9 (San Diego, CA, USA). All experimental data were collected from at least three biological replicates. Error bar was presented as standard error (SE).

Results

Identification and phylogenetic analysis of *MrUGT* gene family

To identify UGT gene family in *M. rubra* genome, an HMM profile (PF00201) was used as a query file to find MrUGT proteins. The screening criteria was that the *E*-value < 1. After manual correction of incomplete sequences based on RNA-Seq database, sequences containing more than 350 amino acids were chosen for further analysis. A total of 152 predicted

amino acid sequences with conserved PSPG-box were obtained. A phylogenetic tree was constructed with other plant UGTs to investigate functional UGT in *M. rubra*. Results showed that MrUGTs were phylogenetically divided into 13 major groups, i.e., A–H, J–M, and O (**Figure 1**). Among them, 12 groups (A–H, J–M) were identified in *Arabidopsis* (Ross et al., 2001) and one group (group O) was newly identified (Caputi et al., 2012). Group I and N were absent in *M. rubra* genome (**Figure 1**).

Analysis of conserved motif and gene structure of *MrUGT* gene family

To investigate characteristics of *MrUGT* gene family, conserved motifs and intron-exon structure were analyzed. Number of MrUGT proteins was different in each group. Group E contained the largest members in MrUGT gene family, i.e., 34 MrUGT members (22%) (**Figure 2**). Followed by group L and group G, the MrUGT number was 24 (16%) and 23 (15%),

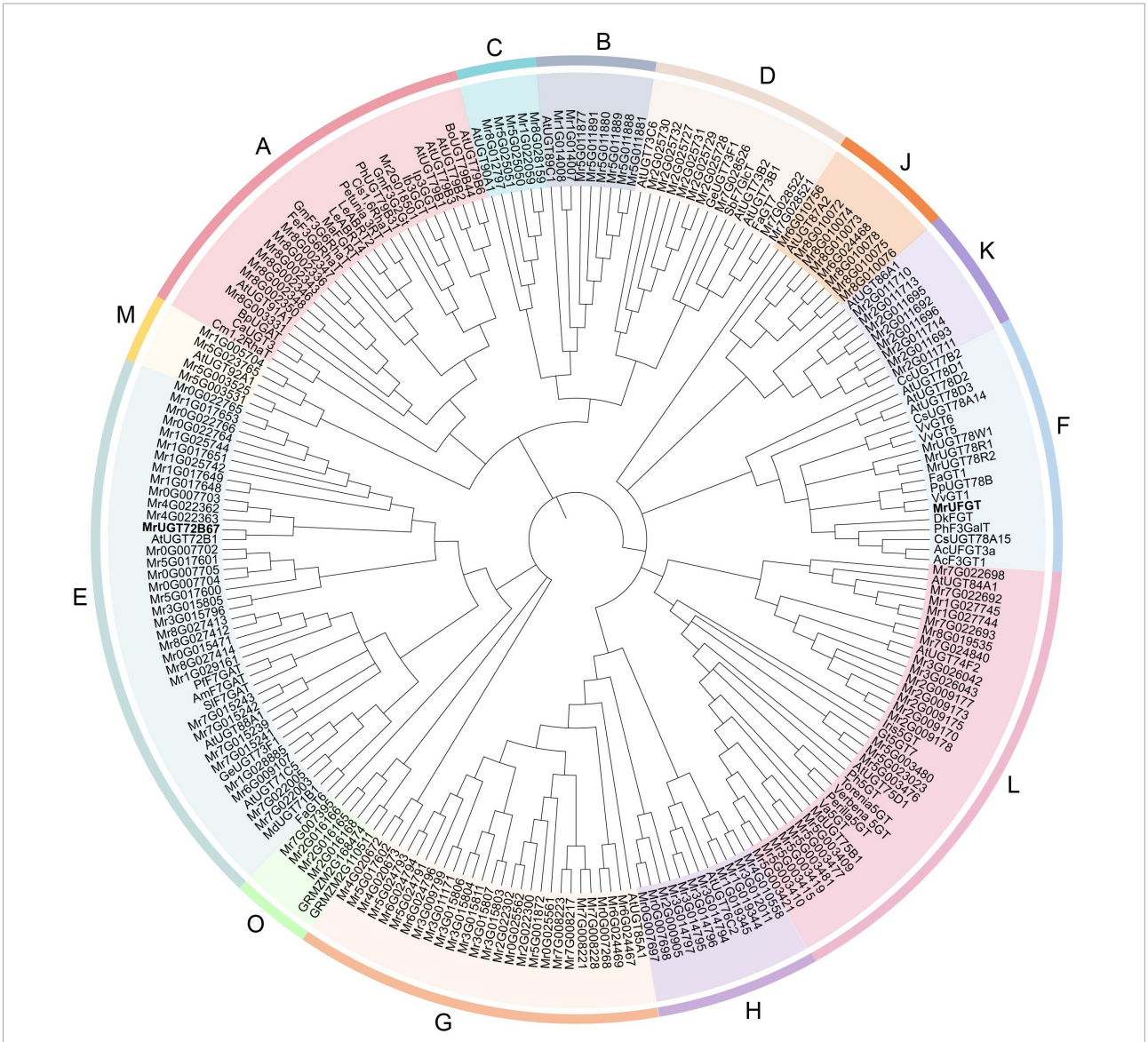


FIGURE 1
Phylogenetic analysis of *Morella rubra* UDP-glycosyltransferase (UGT) gene family. Phylogenetic tree was constructed by neighbor-joining method. Groups are shown in different colors. Abbreviations of species names are follows: AC, *Aralia cordata*; Am, *Antirrhinum majus*; At, *Arabidopsis thaliana*; Bo, *Brassica oleracea*; Bp, *Bellis perennis*; Ca, *Catharanthus roseus*; Cc, *Crocosmia* × *crocosmiiflora*; Cm, *Citrus maxima*; Cis, *Citrus sinensis*; Cs, *Camellia sinensis*; Dk, *Diospyros kaki*; Fa, *Fragaria* × *ananassa*; Fe, *Fagopyrum esculentum*; Ge, *Glycyrrhiza echinata*; Gm, *Glycine max*; Gt, *Gentiana triflora*; Iris, *Iris hollandica*; Ib, *Ipomoea batatas*; Ip, *Ipomoea nil*; Le, *Lobelia erinus*; Ma, *Morus alba*; Md, *Malus* × *domestica*; Perilla, *Perilla frutescens*; Ph, *Petunia hybrida*; Pf, *Perilla frutescens*; Pp, *Prunus persica*; Sb, *Scutellaria baicalensis*; Sl, *Scutellaria laeteviolacea*; Torenia, *Torenia hybrid*; Va, *Vitis amurensis*; Verbena, *Verbena hybrida*; Vv, *Vitis vinifera*. Accession numbers of UGTs from other species are shown in **Supplementary Table 1**.

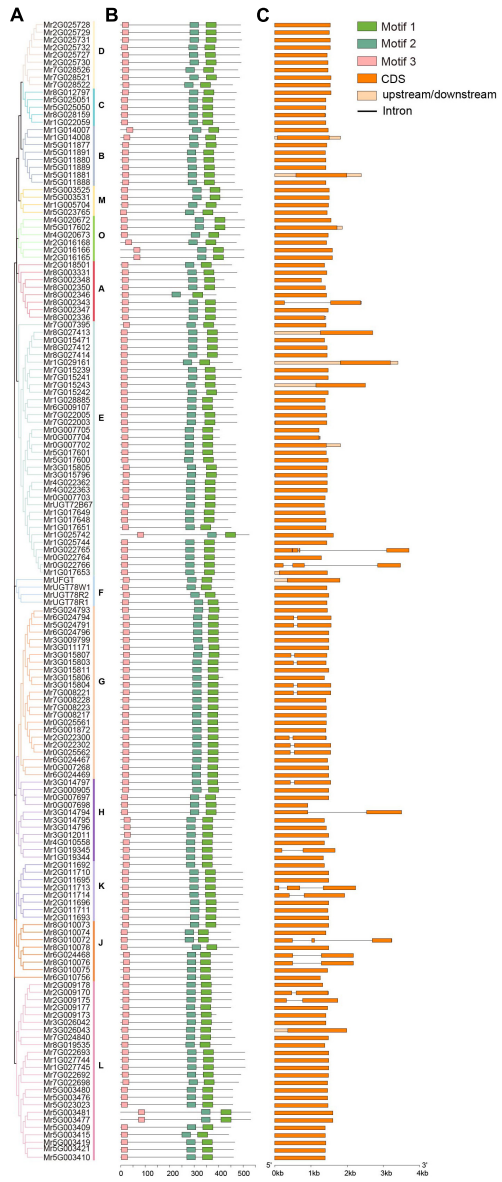


FIGURE 2
Analysis of phylogenetic relationship **(A)**, conserved motif **(B)**, and gene structure **(C)** of MrUGT gene family. The 14 groups are shown with different colors. Three conserved motifs were shown with different colors. Orange box represented exons and black lines represented introns.

respectively (Figure 2). Groups F and M contained the least MrUGT members, both were four UGT members (Figure 2). Three conserved motifs were predicted in MrUGT family based on MEME analysis. Motif 1 was conserved PSPG-box, and motif 2 and 3 were conserved in all MrUGT proteins (Figure 2 and Supplementary Figure 1). This indicating that UGT also has other conserved motif in addition to PSPG-box.

Intron-exon structure was investigated to understand gene function and evolutionary relationships within *MrUGT* gene family. Results showed that 22 *MrUGT* members contained introns, accounting for about 15% (Figure 2). In terms of intron numbers, 18 *MrUGT*s contained one intron, three *MrUGT*s had two introns, and one *MrUGT* had three introns (Figure 2). For *UGT* groups, the largest number of *UGT*s with introns was observed in group G, and that was nine members. Followed by group H and J, both groups had three *UGT*s with introns (Figure 2). Most of *MrUGT*s does not had introns, and gene structure was relatively conservative.

Chromosome distribution and synteny analysis of *MrUGT* gene family

To investigate the distribution of *MrUGT* genes, genomic positions of each *MrUGT* were marked on chromosomes (Figure 3). A total of 139 *MrUGT* genes were marked on eight chromosomes of *M. rubra* and 13 *MrUGT* genes located on unmapped scaffolds (Figure 3 and Supplementary Table 6). There were largest *MrUGT* numbers (28) located on chromosome 2, followed by 27 *MrUGTs* on chromosome 5 and 20 *MrUGTs* on chromosome 7. Only five *MrUGT* genes located on chromosome 4. For the largest MrUGT group (Group E), eight members were distributed on chromosome 1, three members were distributed on chromosome 3, two members were distributed on chromosome 4, two members were distributed on chromosome 5, one member was distributed on chromosome 6, seven members were distributed on chromosome 7, three members were distributed on chromosome 8, and eight members were distributed on unmapped scaffolds (Figure 3 and Supplementary Table 6).

Gene duplication was one of driven forces for gene family expansion (Qiao et al., 2019). Four gene duplication modes were identified in *MrUGT* gene family based on method reported by Qiao et al. (2019), including whole-genome duplication (WGD), dispersed duplication (DSD), tandem duplication (TD), and proximal duplication (PD). A total of 29 TD events were observed in MrUGT gene family, followed by 28 PD events. Only eight DSD events and three WGD events were observed in *MrUGT* gene family (Supplementary Table 7). Group L contained the largest number of gene duplication events, and it was 14. Followed by groups E and G, number of gene duplication events was nine and eight, respectively (Supplementary Table 7).

To further explore evolutionary relationships of *MrUGT*, syntenic maps were constructed between *M. rubra* and three Rosid species, including *Arabidopsis*, *J. regia*, and *P. persica* (Supplementary Figure 2). A total of 22, 41, and 40 homologous *UGT* gene pairs were identified between *M. rubra* and *Arabidopsis*, *J. regia*, and *P. persica*. It indicated that *M. rubra* has a closer evolutionary relationship with *J. regia* and *P. persica*,

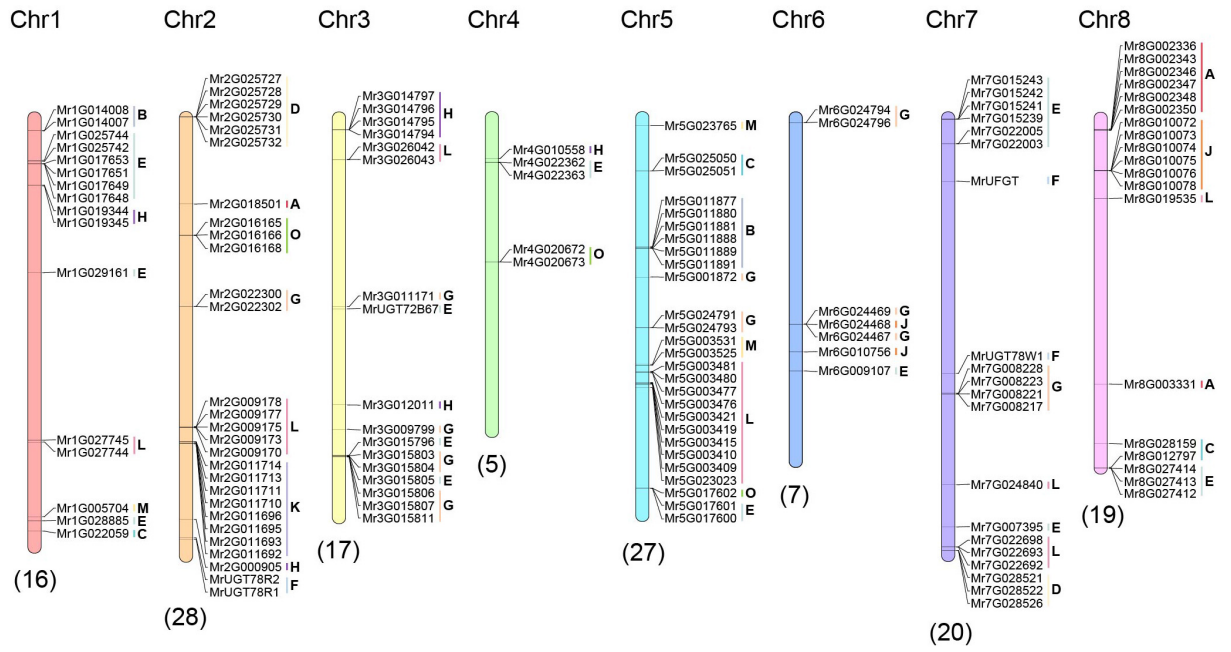


FIGURE 3
Chromosome distribution of *MrUGT* genes. Numbers of chromosomes were labeled at the top of each chromosome. Numbers of *MrUGT*s on chromosome were labeled at the bottom of each chromosome. Group of *MrUGT*s is labeled next to *MrUGT* accession numbers.

which was consistent with the study of *M. rubra* genome (Jia et al., 2019).

Tissue and temporal expression pattern of *MrUGT* genes in *Morella rubra*

RNA-seq was performed to analyze expression pattern of *MrUGT* genes in flowers, leaves, and fruit development stages of 'BQ' cultivar. A total of 29 *MrUGT* genes showed the highest expression level in flowers (**Figure 4**). All *UGT* members in group C exhibited highest expression level in flowers (**Figure 4**). The other *MrUGT* genes expressed highest in flowers were mainly from group E, G, and H (**Figure 4**). A total of 30 *MrUGT* genes showed the highest expression level in leaves. More than half of members in group D and H were mainly expressed in leaves (**Figure 4**). Notably, a total of 99 *MrUGT* members were mainly expressed in fruit, accounting for 65% of total *MrUGT*. Among them, 31 *MrUGT* members had the highest expression level in S1 stage, 23 *MrUGT*s showed the highest expression level in S2 stage, 11 *MrUGT*s showed the highest expression level in S3 stage, and 34 *MrUGT*s showed the highest expression level in S4 stage (**Figure 4**). 22 members of the largest group (group E) showed the highest expression level in fruit (**Figure 4**). All members of group K and group O had the highest expression level in fruit (**Figure 4**). Expression pattern analysis indicated that *MrUGT*s played important roles in metabolic pathways related to fruit development and ripening.

Expression pattern of *MrUGT* genes in response to ultraviolet-B irradiation

UV-B stress is an efficient treatment for induction of flavonoid glycosides accumulation in plants (Kolb et al., 2001; Stracke et al., 2010; Henry-Kirk et al., 2018; Xie et al., 2022). Therefore, we carried out UV-B treatment for investigation and identification of MrUGTs involved in flavonoid glucosylation (**Figure 5**). Based on transcriptomic analysis, gene expression of 13 *MrUGT* genes were significantly induced ($\log_2$FC >1, $p < 0.05$) by UV-B treatment. Among them, seven *MrUGT* genes were significantly induced after 2 days UV-B treatment, and ten *MrUGT* genes were significantly induced after 6 days UV-B treatment (**Supplementary Table 8**). Four *MrUGT* genes were significantly induced by UV-B treatment after both 2 and 6 days.

Based on current knowledge, UGTs in group F were closely related to flavonoid 3-*O*-glycoside formation. Among the UV-B induced MrUGTs, only four members belong to group F, i.e., MrUGT78R1, MrUGT78R2, MrUGT78W1, and MrUFGT (**Supplementary Table 8**). Recently, MrUGT78R1 and MrUGT78R2 were identified as UDP-rhamnosyltransferases while MrUGT78W1 was identified as UDP-galactosyltransferase involved in flavonol glycosylation in *M. rubra* by our group (Ren et al., 2022). Therefore, MrUFGT was chosen as one of potential candidate UGTs for flavonoid glucosylation.

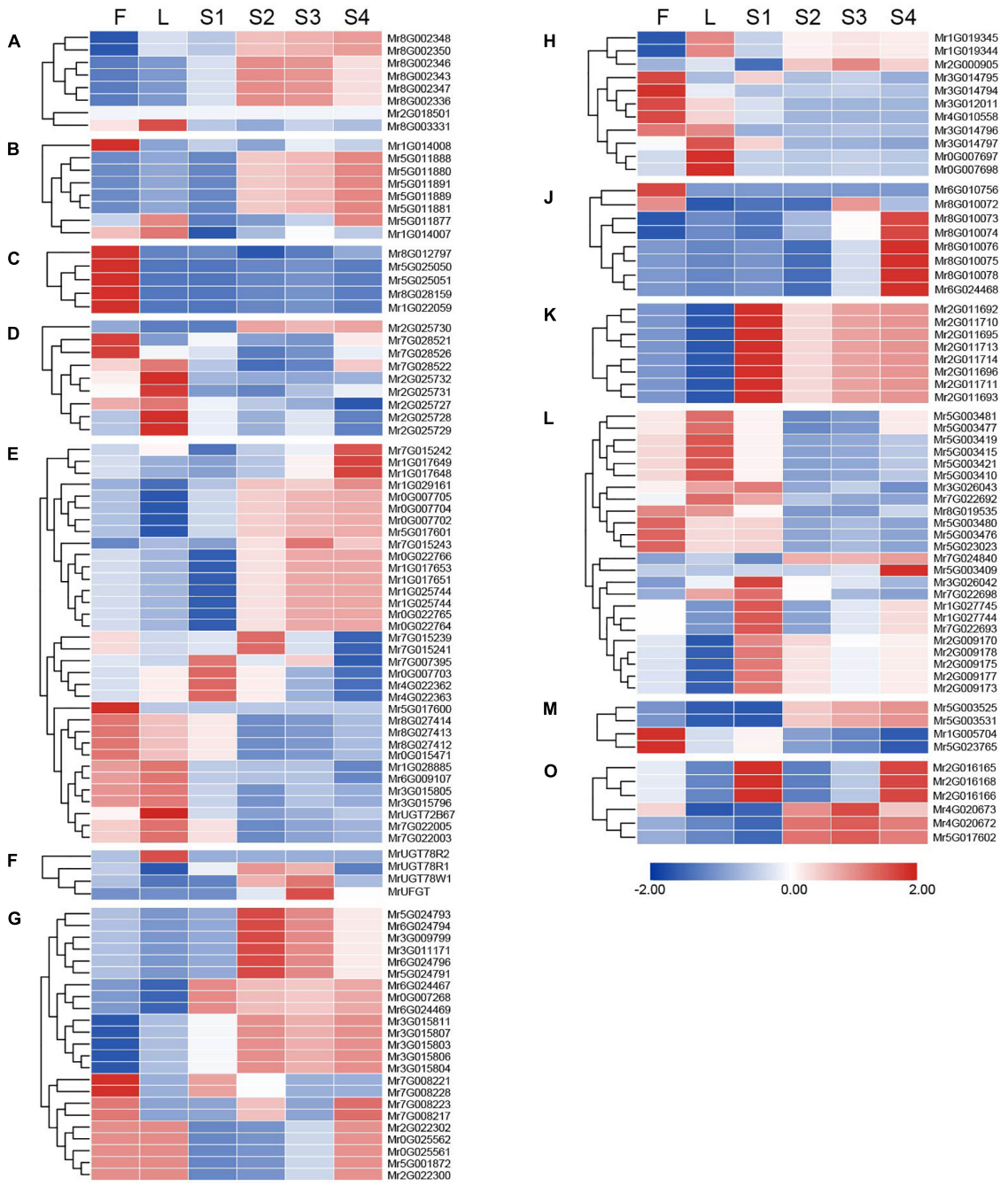


FIGURE 4
Expression pattern of *MrUGT* genes in different tissues of *Morella rubra*. Expression of *MrUGT* genes in flowers (F), leaves (L), and fruit development (S1–S4) are shown. Color scale represents −2 to 2. **(A–H)**, **(J–M)**, and **(O)** mean different phylogenetic groups of *MrUGT* genes.

Identification of MrUGTs related to flavonoid glucoside accumulation

Flavonoid glucoside profiles in different tissues of *M. rubra* were analyzed by HPLC. Flavonoid glucosides accumulation exhibited tissue specificity in *M. rubra*. C3Glc was mainly accumulated in mature fruit (S4) and flowers (Figure 6A). While Q3Glc was mainly accumulated in leaves and flowers (Figure 6A).

Correlation analysis between C3Glc content and expression of *MrUGT*s in different tissues was performed. A total of 12 *MrUGT* genes showed high correlation coefficient ($r > 0.8$)

10.3389/fpls.2022.998985

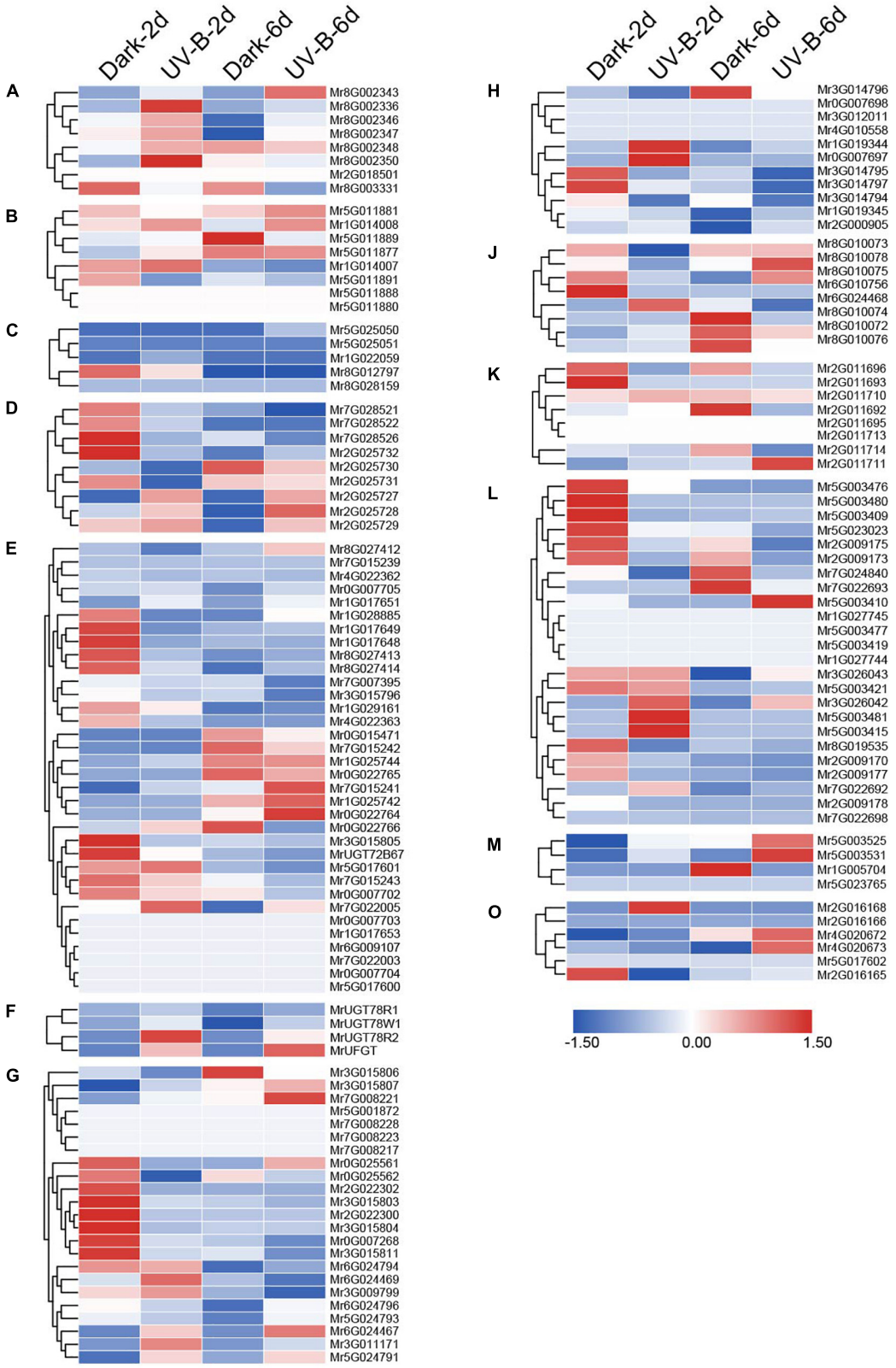


FIGURE 5

Expression pattern of *MrUGT* genes in response to UV-B irradiation. Color scale represents −1.5 to 1.5. **(A–H)**, **(J–M)**, and **(O)** mean different phylogenetic groups of *MrUGT* genes.

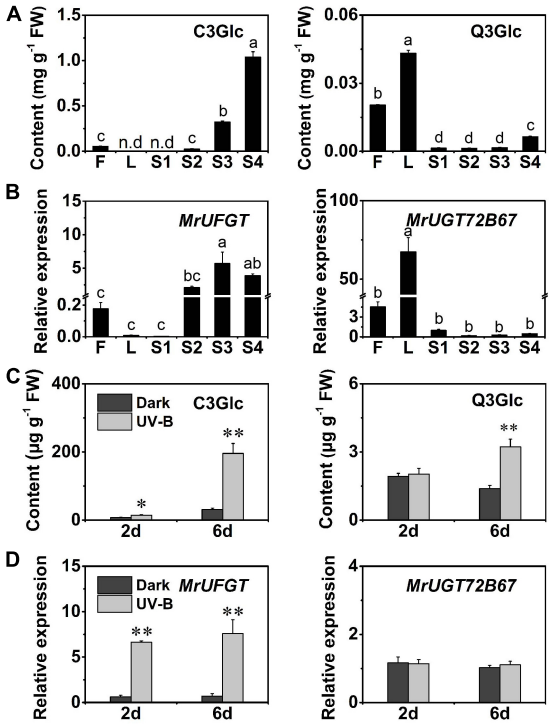


FIGURE 6
Flavonoid glucosides accumulation and gene expression of *MrUGT*s in different tissues and UV-B-treated fruit. **(A)** Accumulation of cyanidin 3-*O*-glucoside (C3Glc) and quercetin 3-*O*-glucoside (Q3Glc) in flowers (F), leaves (L), and fruit development stages (S1–S4) of *Morella rubra*. **(B)** Gene expression of *MrUFGT* and *MrUGT72B67* in different tissues. Different letters indicate significant difference between different groups ($P < 0.05$). **(C)** Effects of UV-B irradiation on content of C3Glc and Q3Glc in *M. rubra* fruit. **(D)** Gene expression of *MrUFGT* and *MrUGT72B67* in response to UV-B irradiation. Student's *t*-test is used for statistical analyses between two samples (**$P < 0.01$, *$P < 0.05$). All data are presented as the mean ± SE ($n = 3$).

with C3Glc content, where only MrUFGT belongs to group F of UGT family (Supplementary Figure 3A). Similarly, correlation analysis between Q3Glc content and expression of *MrUGT*s in different tissues was performed. A total of 9 *MrUGT* genes showed high correlation coefficient ($r > 0.8$) with Q3Glc content (Supplementary Figure 3B). However, none of these 9 MrUGTs belongs to group F of UGT family. *MrUGT72B67* in group E showed the highest expression in leaves and flowers, and was thus chosen for recombinant protein expression and enzymatic assay.

Gene expression of *MrUFGT* and *MrUGT72B67* was confirmed by RT-qPCR (Figure 6B). Results showed that *MrUFGT* was mainly expressed in fruit and flowers, and increased during fruit development (Figure 6B), which was consistent with C3Glc accumulation pattern. While *MrUGT72B67* was mainly expressed in leaves and flowers (Figure 6B), which was consistent with Q3Glc accumulation pattern in leaves and flowers. UV-B treatment could significantly induce C3Glc and Q3Glc accumulation in 'Biqi' fruit (Figure 6C). And gene expression of *MrUFGT* was significantly induced by UV-B, while *MrUGT72B67* were not (Figure 6D).

Enzymatic assays of recombinant MrUFGT and MrUGT72B67

MrUFGT and *MrUGT72B67* were isolated from cDNA libraries of 'Biqi' cultivar. ORFs of *MrUFGT* and *MrUGT72B67* were 1,389 and 1,422 bp, which encoded predicted proteins composed of 462 and 473 amino acids, respectively. Phylogenetic analysis indicated that MrUFGT and MrUGT72B67 exhibited the highest homology with VvGT1 and AtUGT72B1, respectively (Supplementary Figure 4). Sequence alignment analysis showed that PSPG-box of MrUFGT and MrUGT72B67 was conserved and closed to C-terminal (Supplementary Figure 5). Recombinant proteins of MrUFGT and MrUGT72B67 were verified by SDS-PAGE (Supplementary Figure 6). Enzymatic assays were performed to verify functions of MrUFGT and MrUGT72B67. Results showed that MrUFGT could only transfer UDP-Glc to anthocyanidin or flavonol aglycones. Product peaks with *m/z* 448 and *m/z* 463 tentatively identified as C3Glc and Q3Glc based on fragmentation information (Figures 7A,B and Supplementary Figure 7). MrUFGT could not transfer UDP-Rha or UDP-Gal to anthocyanidins or flavonol aglycones such as C or Q (Figures 7A,B). MrUGT72B67 could only transfer UDP-Glc to flavonol aglycones, resulting in formation of peak with *m/z* 463 which was tentatively identified as Q3Glc (Figure 7C and Supplementary Figure 7). MrUGT72B67 could not transfer UDP-Rha or UDP-Gal to flavonol aglycone such as Q (Figure 7C).

Enzyme activity of MrUFGT and MrUGT72B67 for different flavonoid aglycones were also investigated. For MrUFGT, C was the best substrate since MrUFGT showed the highest activity toward it. For different substrates (flavonoid aglycones), relative enzyme activities of MrUFGT were calculated by comparison of enzyme activity toward each substrate with that of C. As a result, MrUFGT showed relative lower activity for flavonol aglycones (M, Q, and K) compared to anthocyanidin aglycones (Figure 7D). MrUFGT did not exhibit glucosyl transfer activity toward naringenin, hesperetin, epicatechin, catechin, luteolin, apigenin, genistein, and daidzein (Figure 7D). For MrUGT72B67, Q was the best substrate since MrUGT72B67 showed the highest activity toward it. For different substrates (flavonoid aglycones), relative enzyme activities of MrUGT72B67 were calculated by comparison of enzyme activity toward each substrate with that of Q. As a result, MrUGT72B67 showed relative lower activity for C, P, naringenin, hesperetin, luteolin, apigenin, and daidzein

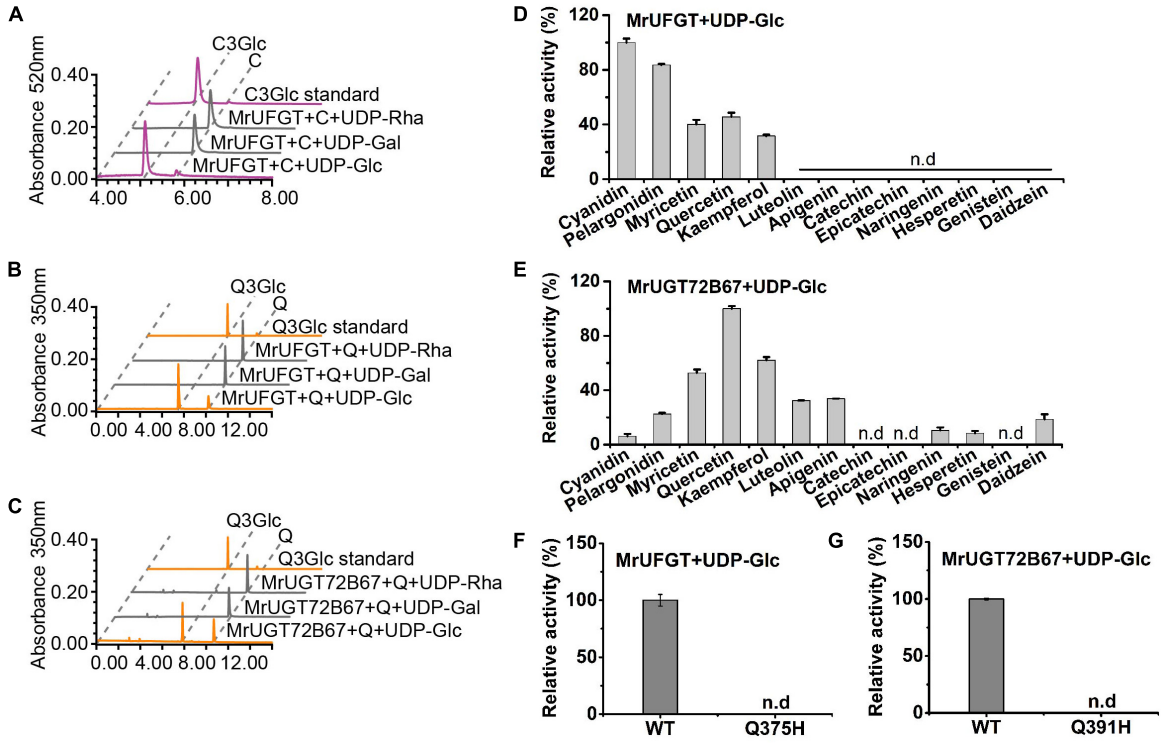


FIGURE 7
Enzymatic assay of MrUFGT and MrUGT72B67. Enzyme activity analysis of recombinant MrUFGT with cyanidin **(A)** and quercetin **(B)** as sugar acceptors, UDP-glucoside (UDP-Glc), UDP-galactoside (UDP-Gal), and UDP-rhamnoside (UDP-Rha) as sugar donors. **(C)** Enzyme activity analysis of recombinant MrUGT72B67 with quercetin as sugar acceptor, UDP-Glc, UDP-Gal, and UDP-Rha as sugar donors. Relative activities of recombinant MrUFGT with UDP-Glc **(D)** and MrUGT72B67 with UDP-Glc **(E)** toward various flavonoids. Site-directed mutagenesis analysis of MrUFGT **(F)** and MrUGT72B67 **(G)** with quercetin as acceptor and UDP-Glc as sugar donor. Data are presented as mean ± SE (n = 3). n.d, not detected.

compared to Q (Figure 7E). It indicated that MrUGT72B67 displayed a relatively broad substrate preference toward flavonoid.

To explore the role of last amino acid residue in PSPG-box for glucosyl transfer activity of MrUGTs, site-directed mutagenesis was carried out. Two mutant proteins (Q375H of MrUFGT and Q391H of MrUGT72B67) were generated by overlapping PCR (Supplementary Figure 8). Q375H mutation and Q391H mutation completely lost the glucosyltransferase activity of MrUFGT and MrUGT72B67, respectively (Figures 7F,G). No mutations resulted in additional galactosyltransferase or rhamnosyltransferase activity (Supplementary Figure 9).

Transient expression of *MrUGTs* in *Nicotiana benthamiana*

To validate functions of *MrUFGT* and *MrUGT72B67 in vivo*, transient expression was carried out in *N. benthamiana* plants. Anthocyanin- and flavonol-specific transcription factors *MrMYB1* (Niu et al., 2010) and *MrMYB12* (Cao et al., 2021) were introduced to transient expression system to enhance substrates level of UGT according to reported (Irmisch et al., 2019; Figure 8A).

Nicotiana benthamiana leaves accumulated cyanidin 3-*O*-rutinoside (m/z 593 with MS^2 fragmentation at m/z 285, C3Rut) when only expressed with MrMYB1 (combination A) (Figure 8B and Supplementary Figure 10). And level of C3Rut was significantly enhanced when *MrUFGT* was added (combination B) (Figure 8B and Supplementary Figure 10). And flavonol glucoside derivatives, i.e., Q3Rut (m/z 609 with MS^2 fragmentation at m/z 300), K3Rut (m/z 593 with MS^2 fragmentation at m/z 258), and Q3Glc (m/z 463 with MS^2 fragmentation at m/z 300), were significantly accumulated in *N. benthamiana* leaves infiltrated with *MrUFGT* (combination D) compared to infiltrated *MrMYB12* only (combination C) (Figure 8C and Supplementary Figure 10). Like MrUFGT, flavonol glucoside derivatives (Q3Rut, K3Rut, and Q3Glc) were significantly accumulated in *N. benthamiana* leaves with addition of *MrUGT72B67* (combination E) compared to expressed *MrMYB12* only (combination C) (Figure 8C and Supplementary Figure 10). Control leaves did not accumulate anthocyanins or flavonol glycosides at detectable level (Figure 8 and Supplementary Figure 10). These results

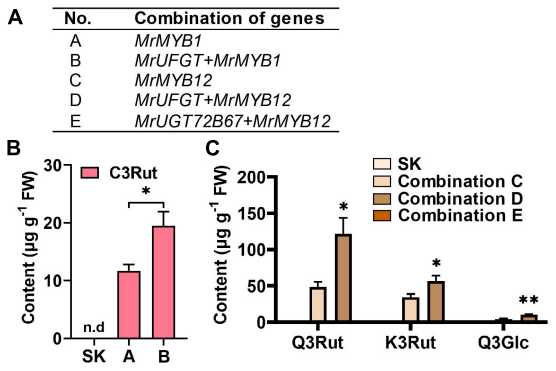


FIGURE 8
Transient expression of MrUGTs in *Nicotiana benthamiana*. **(A)** Transient expression information of combination mixtures. **(B)** Analysis of anthocyanins in *N. benthamiana* leaves infiltrated with combinations contained MrUFGT. **(C)** Analysis of flavonol glucosides in *N. benthamiana* leaves infiltrated with combinations contained MrUFGT or MrUGT72B67. Leaves infiltrated with empty SK vector were used as control. Products were confirmed by LC-MS/MS. Data are mean ± SE (n = 3).

tested the functional glucosyltransferase activity of MrUFGT and MrUGT72B67.

Discussion

UDP-glycosyltransferase gene family contribute to diversity of secondary metabolites

Morella rubra is rich in flavonoid glycosides, and different tissues have been used historically as folk medicines. Here we reported the genome-wide analysis of *UGT* gene family and identified two MrUGTs involved in the accumulation of flavonoid glucosides.

The first plant reported *UGT* gene family was *Arabidopsis* and 107 *UGT* genes was identified in genome (Li et al., 2014). In the present study, a total of 152 *UGT* genes were identified in *M. rubra* genome. Number of *MrUGT* gene was little difference compared to other species, examples include 241 *UGT*s in *M. domestica* (Caputi et al., 2012), 147 *UGT*s in *Z. mays* (Li et al., 2014), and 168 *UGT*s in *P. persica* (Wu et al., 2017). It indicated that *MrUGT* gene family did not exhibit significant expansion compared with other plants, which may be related to the lack of recent genome-wide duplication events in *M. rubra* (Jia et al., 2019). Based on phylogenetic relationship, UGTs in *Arabidopsis* were clustered into 14 groups (A-N) (Li et al., 2001). And four new groups, i.e., O, P, Q, and R, were discovered subsequently in *M. domestica* (Caputi et al., 2012), *Z. mays* (Li et al., 2014), and *C. sinensis* (Cui et al., 2016). *MrUGT* gene family contained 13 groups (**Figure 1**), including 12 groups discovered in *Arabidopsis* and one newly discovered group O, and with absent of group I, N, P, Q, and R compared to 18 groups (A-R) reported in plants. This indicated that gene loss events occurred during *UGT* gene family expansion in *M. rubra*.

UDP-glycosyltransferases prefer to be clustered by regiospecificity rather than species and sugar donor specificity (Yonekura-Sakakibara and Hanada, 2011; Hsu et al., 2017). Therefore, it is considered that sugar donor specificity differentiation was later than divergence of regiospecificity (Hsu et al., 2017). This makes it possible to predict functional UGTs by phylogenetic analysis. For example, UGT members in group A were considered to be related to biosynthesis of flavonoid disaccharides, and examples include Cis1,6RhaT (Frydman et al., 2004), Cm1,2RhaT (Frydman et al., 2013), and PpUGT79AK6 (Xie et al., 2022). UGT members in group O usually exhibited activity toward plant hormone zeatin. For example, PvZOX1 from *Phaseolus vulgaris* was identified as a zeatin *O*-xylosyltransferase involved in *O*-xylosylzeatin formation (Martin et al., 1999). And cisZOG1 from *Z. mays* was identified as a glucosyltransferase specific to *cis*-zeatin (Martin et al., 2001). Phylogenetic homology analysis of UGTs would facilitate the discovery of more functional UGTs in plants with specialized metabolites.

Gene duplication is important for gene family expansion, and results in gene clusters on chromosomes. Gene duplication events include five modes according to Qiao et al. (2019), i.e., WGD, TD, DSD, PD, and transposed duplication (TRD). Four modes except TRD were found in *MrUGT* family. About 19% and 18% *MrUGT* genes occurred TD and PD events, respectively, indicating that both TD and PD were ongoing processes throughout evolutionary of *MrUGT* gene family. In *Broussonetia papyrifera*, TD was primary driving force for expansion of *BpUGT* gene family (Wang et al., 2021). In higher plants, it was found that groups A, D, E, G, and L expanded more than other groups during plant evolution (Caputi et al., 2012). In *M. rubra*, groups E, G, and L were expanded significantly compared with other groups, which was mainly related to gene duplication events in these groups.

MrUFGT and MrUGT72B67 involved in flavonoid glucosylation

Various anthocyanins and flavonol glycosides are of interest to researchers because of their importance in plant physiology and human health. To date, many plant UGTs involved in biosynthesis of anthocyanins and flavonol glycosides are reported. Flavonoid 3-*O*-glycosyltransferase (UFGT) *bronze1* from maize was the first identified UGT in plant that only used UDP-Glc for the biosynthesis of anthocyanins, which were important for pigment accumulation in maize (Dooner and Nelson, 1977). In grape, VvGT1 was a flavonoid 3-*O*-glucosyltransferase that catalyzed anthocyanins formation

during grape fruit ripening (Ford et al., 1998). In model plant *Arabidopsis*, AtUGT78D2 was identified as flavonoid 3-*O*-glucosyltransferase by enzymatic activity analysis and T-DNA-inserted mutants (Tohge et al., 2005). Recently, PpUGT78T3 was identified as UDP-glucosyltransferase involved in regulation of flavonol glucosides in response to UV-B (Xie et al., 2022).

In this work, both transcriptomic data and phylogenetic homology of UGT subgroups together with their correlation with flavonoid accumulation patterns in different tissues or under UV-B treatment were comprehensively analyzed for screen of candidate UGTs involved in flavonoid glucosides accumulation. MrUFGT was mainly screened based on phylogenetic homology analysis with group F and correlation relationship between flavonoid glucosides contents and its expression, while MrUGT72B67 was screened based on tissue specific accumulation of flavonoid glucosides and its transcriptomic analysis. Here we demonstrated that MrUFGT was involved in C3Glc accumulation by *in vitro* and *in vivo* experimental data. In addition, MrUFGT exhibited activity toward Q resulting in Q3Glc formation. However, Q3Glc accumulation pattern in flowers and leaves was not correlation with gene expression pattern of *MrUFGT*. This indicating that there might be another UGT member involved in Q3Glc accumulation in flowers and leaves. MrUGT72B67 in group E was found to be involved in Q3Glc accumulation in leaves and flowers by gene expression analysis as well as *in vitro* and *in vivo* data. Taken together the results of C3Glc and Q3Glc induced by UV-B treatment (**Figure 6C**), we concluded that MrUFGT mainly involved in accumulation of C3Glc and Q3Glc in fruit, while MrUGT72B67 mainly involved in accumulation of Q3Glc in flowers and leaves.

UDP-glycosyltransferase members in group F were closely related to flavonoid 3-*O*-glycoside formation (Ono et al., 2010; Cheng et al., 2014; Cui et al., 2016; Xie et al., 2022). For example, VvGT5 and VvGT6 in group F from *Vitis vinifera* were identified as flavonol 3-*O*-glucuronosyltransferase and bifunctional flavonol 3-*O*-glucosyltransferase/galactosyltransferase in grapevines (Ono et al., 2010). In *C. sinensis*, CsUGT78A14 and CsUGT78A15 in group F were reported to be responsible for biosynthesis of flavonol 3-*O*-glucosides and flavonol 3-*O*-galactosides, respectively (Cui et al., 2016). PpUGT78A2 in group F was identified as a flavonoid 3-*O*-glycosyltransferase involved in different glycosylation of anthocyanin and flavonol in *P. persica* (Cheng et al., 2014; Xie et al., 2022). And in *M. rubra*, four UGT members in group F, i.e., MrUGT78R1, MrUGT78R2, MrUGT78W1, and MrUFGT in the present study, were identified as flavonoid 3-*O*-glycosyltransferases involved in accumulation of diverse flavonoid glycosides (Ren et al., 2022).

UDP-glycosyltransferase members in group E have been reported with diverse functions in many plants. In *Arabidopsis*, AtUGT72B1 was identified as a bifunctional *O*-glucosyltransferase and *N*-glucosyltransferase involved in metabolism of pollutant 3,4-dichloroaniline (Loutre et al., 2003), and it was also involved in glucose conjugation of monolignols, which play an important role in cell wall lignification in *Arabidopsis* (Lin et al., 2016). In *Lotus japonicus*, three UGTs from group E, i.e., UGT72AD1, UGT72AH1, and UGT72Z2, were identified as glucosyltransferases involved in flavonol glucoside/rhamnoside biosynthesis in *L. japonicus* seeds (Yin et al., 2017).

Key amino acids in glucosyltransferases

Crystal structure analysis of UGTs have showed that last amino acid residue in PSPG-box was critical for glycosyl transfer activity of UGT (Shao et al., 2005; Offen et al., 2006; Osmani et al., 2009). Last amino acid residue of PSPG-box in UDP-glucosyltransferases usually was glutamine (Gln), such as observed in UGT78D2 from *Arabidopsis* (Tohge et al., 2005), CsUGT78A14 from tea plant (Cui et al., 2016), and FaGT6 and FaGT7 from strawberry (Griesser et al., 2008). Some site-directed mutagenesis indicated the important role of Gln as last amino acid residue in PSPG-box. For example, by replacing Gln382 with His in UBGT from *Scutellaria baicalensis*, UBGT exhibited remarkable decrease in glucosyltransferase activity (Kubo et al., 2004). In VvGT1, Q375H mutation completely abolished glucosyl transfer activity, and did not improve galactosyl transfer activity (Offen et al., 2006). Q378H substitution for CsUGT78A14 resulted in glucosyl transfer activity markedly reduced, which indicated that Gln was important for flavonoid 3-*O*-glucosyltransferase activity (Cui et al., 2016).

In the present study, last amino acid residues in PSPG-box were both Gln in MrUFGT (Gln375) and MrUGT72B67 (Gln391). To investigate whether last amino acid residue in PSPG-box was important for glucosyl transfer activity, site-directed mutagenesis of Q375H mutation for MrUFGT and Q391H mutation for MrUGT72B67 were analyzed by enzymatic assay. Results showed that both mutation of Q375H for MrUFGT and Q391H for MrUGT72B67 abolished glucosyl transfer activity. It indicated that Gln as last amino acid residue in PSPG-box were critical for glucosyl transfer activity for MrUFGT and MrUGT72B67.

Conclusion

In the present study, genome-wide analysis was performed for *UGT* gene family in *M. rubra*, including polygenetic information, chromosomal distribution, gene duplication mode, and expression pattern. A total of 152 *UGT* family members were identified in *M. rubra* genome and clustered into 13 groups based on polygenetic analysis. 139 *MrUGT* genes marked on eight chromosomes and 13 *MrUGT* genes located on unmapped scaffolds. Gene duplication analysis indicated that both tandem

and proximal duplication were major drivers for *MrUGT* gene family expansion. Expression analysis indicated *MrUGT*s played important roles during fruit development and ripening. MrUFGT (MrUGT78A26) and MrUGT72B67 were identified as UDP-glucosyltransferases by *in vitro* and *in vivo* experiment which were involved in C3Glc and Q3Glc accumulation in different tissues of *M. rubra*. In addition, Gln375 and Gln391 were identified as important amino acids for glucosyltransferase activity of MrUFGT and MrUGT72B67, respectively.

Data availability statement

The original contributions presented in this study are publicly available. This data can be found here: NCBI, SRP386597 and SRP310482.

Author contributions

XL and CR designed the project and drafted the manuscript. CR carried out analyses and experiments with the help of YC, MX, YG, JL, and LX. CS, CX, and KC provided supports to the *M. rubra* project. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (31872067), the Key Research and Development Program of Zhejiang Province (2021C02001), and the 111 project (B17039).

Acknowledgments

We thank Prof. Liang Yan for providing *A. tumefaciens* p19 strain.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.998985/full#supplementary-material

References

Alseekh, S., Perez de Souza, L., Benina, M., and Fernie, A. R. (2020). The style and substance of plant flavonoid decoration; towards defining both structure and function. *Phytochemistry* 174:112347. doi: 10.1016/j.phytochem.2020.112347

Bondonno, N. P., Dalgaard, F., Kyro, C., Murray, K., Bondonno, C. P., Lewis, J. R., et al. (2019). Flavonoid intake is associated with lower mortality in the Danish Diet Cancer and Health Cohort. *Nat. Commun.* 10:3651. doi: 10.1038/s41467-019-11622-x

Bowles, D., Lim, E. K., Poppenberger, B., and Vaistij, F. E. (2006). Glycosyltransferases of lipophilic small molecules. *Annu. Rev. Plant Biol.* 57, 567–597. doi: 10.1146/annurev.arplant.57.032905.105429

Cao, Y., Jia, H., Xing, M., Jin, R., Grierson, D., Gao, Z., et al. (2021). Genome-wide analysis of MYB gene family in Chinese bayberry (*Morella rubra*) and identification of members regulating flavonoid biosynthesis. *Front. Plant Sci.* 12:691384. doi: 10.3389/fpls.2021.691384

Cao, Y., Xie, L., Ma, Y., Ren, C., Xing, M., Fu, Z., et al. (2019). PpMYB15 and PpMYBF1 transcription factors are involved in regulating flavonol biosynthesis in peach fruit. *J. Agric. Food Chem.* 67, 644–652. doi: 10.1021/acs.jafc.8b04810

Caputi, L., Malnoy, M., Goremykin, V., Nikiforova, S., and Martens, S. (2012). A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. *Plant J.* 69, 1030–1042. doi: 10.1111/j.1365-313X.2011.04853.x

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Cheng, J., Wei, G., Zhou, H., Gu, C., Vimolmangkang, S., Liao, L., et al. (2014). Unraveling the mechanism underlying the glycosylation and methylation of anthocyanins in peach. *Plant Physiol.* 166, 1044–1058. doi: 10.1104/pp.114.246876

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: A sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Cui, L., Yao, S., Dai, X., Yin, Q., Liu, Y., Jiang, X., et al. (2016). Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). *J. Exp. Bot.* 67, 2285–2297. doi: 10.1093/jxb/erw053

Dooner, H. K., and Nelson, O. E. (1977). Controlling element-induced alterations in UDPglucose: Flavonoid glucosyltransferase, the enzyme specified by the bronze locus in maize. *Proc. Natl. Acad. Sci. U.S.A.* 74, 5623–5627. doi: 10.1073/pnas.74.12.5623

Downey, M. O., Mazza, M., and Krstic, M. P. (2007). Development of a stable extract for anthocyanins and flavonols from grape skin. *Am. J. Enol. Vitic.* 58, 358–364. doi: 10.1111/1750-3841.12108

Feng, C., Xu, C. J., Wang, Y., Liu, W. L., Yin, X. R., Li, X., et al. (2013). Codon usage patterns in Chinese bayberry (*Myrica rubra*) based on RNA-Seq data. *BMC Genomics* 14:732. doi: 10.1186/1471-2164-14-732

Ford, C. M., Boss, P. K., and Hoj, P. B. (1998). Cloning and characterization of *Vitis vinifera* UDP-glucose: Flavonoid 3-O-glucosyltransferase, a homologue of the enzyme encoded by the maize bronze-1 locus that may primarily serve to glucosylate anthocyanidins in vivo. *J. Biol. Chem.* 273, 9224–9233. doi: 10.1074/jbc.273.15.9224

Frydman, A., Liberman, R., Huhman, D. V., Carmeli-Weissberg, M., Sapir-Mir, M., Ophir, R., et al. (2013). The molecular and enzymatic basis of bitter/non-bitter flavor of citrus fruit: Evolution of branch-forming rhamnosyltransferases under domestication. *Plant J.* 73, 166–178. doi: 10.1111/tpj.12030

Frydman, A., Weisshaus, O., Bar-Peled, M., Huhman, D. V., Sumner, L. W., Marin, F. R., et al. (2004). Citrus fruit bitter flavors: Isolation and functional characterization of the gene Cm1,2RhaT encoding a 1,2 rhamnosyltransferase, a key enzyme in the biosynthesis of the bitter flavonoids of citrus. *Plant J.* 40, 88–100. doi: 10.1111/j.1365-313X.2004.02193.x

Griesser, M., Vitzthum, F., Fink, B., Bellido, M. L., Raasch, C., Munoz-Blanco, J., et al. (2008). Multi-substrate flavonol O-glucosyltransferases from strawberry (*Fragaria* × *ananassa*) achene and receptacle. *J. Exp. Bot.* 59, 2611–2625. doi: 10.1093/jxb/ern117

Henry-Kirk, R. A., Plunkett, B., Hall, M., McGhie, T., Allan, A. C., Wargent, J. J., et al. (2018). Solar UV light regulates flavonoid metabolism in apple (*Malus* × *domestica*). *Plant Cell Environ.* 41, 675–688. doi: 10.1111/pce.13125

Hsu, Y. H., Tagami, T., Matsunaga, K., Okuyama, M., Suzuki, T., Noda, N., et al. (2017). Functional characterization of UDP-rhamnose-dependent rhamnosyltransferase involved in anthocyanin modification, a key enzyme determining blue coloration in *Lobelia erinus*. *Plant J.* 89, 325–337. doi: 10.1111/tpj.13387

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., and Gao, G. (2014). GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* 31, 1296–1297. doi: 10.1093/bioinformatics/btu817

Irmisch, S., Jancsik, S., Man Saint Yuen, M., Madilao, L. L., and Bohlmann, J. (2020). Complete biosynthesis of the anti-diabetic plant metabolite Montbretin A. *Plant Physiol.* 184, 97–109. doi: 10.1104/pp.20.00522

Irmisch, S., Jo, S., Roach, C. R., Jancsik, S., Man Saint Yuen, M., Madilao, L. L., et al. (2018). Discovery of UDP-glycosyltransferases and BAHD-acyltransferases involved in the biosynthesis of the antidiabetic plant metabolite Montbretin A. *Plant Cell* 30, 1864–1886. doi: 10.1105/tpc.18.00406

Irmisch, S., Ruebsam, H., Jancsik, S., Man Saint Yuen, M., Madilao, L. L., and Bohlmann, J. (2019). Flavonol biosynthesis genes and their use in engineering the plant antidiabetic metabolite Montbretin A. *Plant Physiol.* 180, 1277–1290. doi: 10.1104/pp.19.00254

Jia, H. M., Jia, H. J., Cai, Q. L., Wang, Y., Zhao, H. B., Yang, W. F., et al. (2019). The red bayberry genome and genetic basis of sex determination. *Plant Biotechnol. J.* 17, 397–409. doi: 10.1111/pbi.12985

Kolb, C. A., Kaäser, M. A., Kopeckyì, J., Zotz, G., Riederer, M., and Pfündel, E. E. (2001). Effects of Natural intensities of visible and ultraviolet radiation on epidermal ultraviolet screening and photosynthesis in grape leaves. *Plant Physiol.* 127, 863–875. doi: 10.1104/pp.010373

Kubo, A., Arai, Y., Nagashima, S., and Yoshikawa, T. (2004). Alteration of sugar donor specificities of plant glycosyltransferases by a single point mutation. *Arch. Biochem. Biophys.* 429, 198–203. doi: 10.1016/j.abb.2004.06.021

Li, Y., Baldauf, S., Lim, E. K., and Bowles, D. J. (2001). Phylogenetic analysis of the UDP-glycosyltransferase multigene family of *Arabidopsis thaliana*. *J. Biol. Chem.* 276, 4338–4343. doi: 10.1074/jbc.M007447200

Li, Y., Li, P., Wang, Y., Dong, R., Yu, H., and Hou, B. (2014). Genome-wide identification and phylogenetic analysis of Family-1 UDP glycosyltransferases in maize (*Zea mays*). *Planta* 239, 1265–1279. doi: 10.1007/s00425-014-2050-1

Lin, J. S., Huang, X. X., Li, Q., Cao, Y., Bao, Y., Meng, X. F., et al. (2016). UDP-glycosyltransferase 72B1 catalyzes the glucose conjugation of monolignols and is essential for the normal cell wall lignification in *Arabidopsis thaliana*. *Plant J.* 88, 26–42. doi: 10.1111/tpj.13229

Liu, Y., Wang, R., Ren, C., Pan, Y., Li, J., Zhao, X., et al. (2022). Two Myricetin-derived flavonols from *Morella rubra* leaves as potent alpha-glucosidase inhibitors and structure-activity relationship study by computational chemistry. *Oxid Med. Cell Longev.* 2022:9012943. doi: 10.1155/2022/9012943

Liu, Y., Zhan, L., Xu, C., Jiang, H., Zhu, C., Sun, L., et al. (2020). alpha-Glucosidase inhibitors from Chinese bayberry (*Morella rubra* Sieb. et Zucc.) fruit: Molecular docking and interaction mechanism of flavonols with different B-ring hydroxylations. *RSC Adv.* 10, 29347–29361. doi: 10.1039/d0ra05015f

Loutre, C., Dixon, D. P., Brazier, M., Slater, M., Cole, D. J., and Edwards, R. (2003). Isolation of a glucosyltransferase from *Arabidopsis thaliana* active in the metabolism of the persistent pollutant 3,4-dichloroaniline. *Plant J.* 34, 485–493. doi: 10.1046/j.1365-313x.2003.01742.x

Martin, R. C., Mok, M. C., and Mok, D. W. S. (1999). A gene encoding the cytokinin enzyme zeatin O-xylosyltransferase of *Phaseolus vulgaris*. *Plant Physiol.* 120, 553–558. doi: 10.1104/pp.120.2.553

Martin, R. C., Mok, M. C., Habben, J. E., and Mok, D. W. S. (2001). A maize cytokinin gene encoding an O-glucosyltransferase specific to cis-zeatin. *Proc. Natl. Acad. Sci. U.S.A.* 98, 5922–5926. doi: 10.1073/pnas.101128798

Naeem, A., Ming, Y., Pengyi, H., Jie, K. Y., Yali, L., Haiyan, Z., et al. (2021). The fate of flavonoids after oral administration: A comprehensive overview of its bioavailability. *Crit. Rev. Food Sci. Nutr.* 62, 6169–6186. doi: 10.1080/10408398.2021.1898333

Niu, S. S., Xu, C. J., Zhang, W. S., Zhang, B., Li, X., Lin-Wang, K., et al. (2010). Coordinated regulation of anthocyanin biosynthesis in Chinese bayberry (*Myrica rubra*) fruit by a R2R3 MYB transcription factor. *Planta* 231, 887–899. doi: 10.1007/s00425-009-1095-z

Offen, W., Martinez-Fleites, C., Yang, M., Kiat-Lim, E., Davis, B. G., Tarling, C. A., et al. (2006). Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO J.* 25, 1396–1405. doi: 10.1038/sj.emboj.7600970

Ono, E., Homma, Y., Horikawa, M., Kunikane-Doi, S., Imai, H., Takahashi, S., et al. (2010). Functional differentiation of the glycosyltransferases that contribute to the chemical diversity of bioactive flavonol glycosides in grapevines (*Vitis vinifera*). *Plant Cell* 22, 2856–2871. doi: 10.1105/tpc.110.074625

Osmani, S. A., Bak, S., and Moller, B. L. (2009). Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry* 70, 325–347. doi: 10.1016/j.phytochem.2008.12.009

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* 20:38. doi: 10.1186/s13059-019-1650-2

Ren, C., Guo, Y., Xie, L., Zhao, Z., Xing, M., Cao, Y., et al. (2022). Identification of UDP-rhamnosyltransferases and UDP-galactosyltransferase involved in flavonol glycosylation in Morella rubra. *Hortic. Res.* 9:uhac138. doi: 10.1093/hr/uhac138

Ross, J., Li, Y., Lim, E. K., and Bowles, D. J. (2001). Higher plant glycosyltransferases. *Genome Biol.* 2, 1–6.

Shao, H., He, X., Achnine, L., Blount, J. W., Dixon, R. A., and Wang, X. (2005). Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from Medicago truncatula. *Plant Cell* 17, 3141–3154. doi: 10.1105/tpc.105.035055

Stracke, R., Favory, J. J., Gruber, H., Bartelniewoehner, L., Bartels, S., Binkert, M., et al. (2010). The Arabidopsis bZIP transcription factor HY5 regulates expression of the PFG1/MYB12 gene in response to light and ultraviolet-B radiation. *Plant Cell Environ.* 33, 88–103. doi: 10.1111/j.1365-3040.2009.02061.x

Sun, C., Huang, H., Xu, C., Li, X., and Chen, K. (2013). Biological activities of extracts from Chinese bayberry (*Myrica rubra* Sieb. et Zucc.): A review. *Plant Foods Hum. Nutr.* 68, 97–106. doi: 10.1007/s11130-013-0349-x

Sun, C., Zheng, Y., Chen, Q., Tang, X., Jiang, M., Zhang, J., et al. (2012). Purification and anti-tumour activity of cyanidin-3-O-glucoside from Chinese bayberry fruit. *Food Chem.* 131, 1287–1294. doi: 10.1016/j.foodchem.2011.09.121

Tohge, T., Nishiyama, Y., Hirai, M. Y., Yano, M., Nakajima, J., Awazuhara, M., et al. (2005). Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* 42, 218–235. doi: 10.1111/j.1365-313X.2005.02371.x

Vogt, T., and Jones, P. (2000). Glycosyltransferases in plant natural product synthesis: Characterization of a supergene family. *Trends Plant Sci.* 5, 1360–1385. doi: 10.1016/S1360-1385(00)01720-9

Wang, F., Su, Y., Chen, N., and Shen, S. (2021). Genome-wide analysis of the UGT gene family and identification of flavonoids in *Broussonetia papyrifera*. *Molecules* 26:3449. doi: 10.3390/molecules26113449

Wu, B., Gao, L., Gao, J., Xu, Y., Liu, H., Cao, X., et al. (2017). Genome-wide identification, expression patterns, and functional analysis of UDP glycosyltransferase family in peach (*Prunus persica* L. Batsch). *Front. Plant Sci.* 8:389. doi: 10.3389/fpls.2017.00389

Xie, L., Cao, Y., Zhao, Z., Ren, C., Xing, M., Wu, B., et al. (2020). Involvement of MdUGT75B1 and MdUGT71B1 in flavonol galactoside/glucoside biosynthesis in apple fruit. *Food Chem.* 312:126124. doi: 10.1016/j.foodchem.2019.126124

Xie, L., Guo, Y., Ren, C., Cao, Y., Li, J., Lin, J., et al. (2022). Unravelling the consecutive glycosylation and methylation of flavonols in peach in response to UV-B irradiation. *Plant Cell Environ.* 45, 2158–2175. doi: 10.1111/pce.14323

Yan, S., Zhang, X., Wen, X., Lv, Q., Xu, C., Sun, C., et al. (2016). Purification of flavonoids from Chinese bayberry (*Morella rubra* Sieb. et Zucc.) fruit extracts and alpha-glucosidase inhibitory activities of different fractionations. *Molecules* 21:1148. doi: 10.3390/molecules21091148

Yang, B., Liu, H., Yang, J., Gupta, V. K., and Jiang, Y. (2018). New insights on bioactivities and biosynthesis of flavonoid glycosides. *Trends Food Sci. Technol.* 79, 116–124. doi: 10.1016/j.tifs.2018.07.006

Yin, Q., Shen, G., Chang, Z., Tang, Y., Gao, H., and Pang, Y. (2017). Involvement of three putative glucosyltransferases from the UGT72 family in flavonol glucoside/rhamnoside biosynthesis in *Lotus japonicus* seeds. *J. Exp. Bot.* 68, 597–612. doi: 10.1093/jxb/erw420

Yin, R., Han, K., Heller, W., Albert, A., Dobrev, P. I., Zazimalova, E., et al. (2014). Kaempferol 3-O-rhamnoside-7-O-rhamnoside is an endogenous flavonol inhibitor of polar auxin transport in *Arabidopsis* shoots. *New Phytol.* 201, 466–475. doi: 10.1111/nph.12558

Yonekura-Sakakibara, K., and Hanada, K. (2011). An evolutionary view of functional diversity in family 1 glycosyltransferases. *Plant J.* 66, 182–193. doi: 10.1111/j.1365-313X.2011.04493.x

Zhang, X., Huang, H., Zhang, Q., Fan, F., Xu, C., Sun, C., et al. (2015). Phytochemical characterization of Chinese bayberry (*Myrica rubra* Sieb. et Zucc.) of 17 cultivars and their antioxidant properties. *Int. J. Mol. Sci.* 16, 12467–12481. doi: 10.3390/ijms160612467

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fpls.2022.1024239

OPEN ACCESS

EDITED BY
Peipei Wang,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Chen Junfeng,
Shanghai University of Traditional Chinese Medicine, China
Yuliang Wang,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Xuebo Hu
xuebohu@mail.hzau.edu.cn
Jingying Chen
cjy6601@163.com

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 21 August 2022
ACCEPTED 11 October 2022
PUBLISHED 03 November 2022

CITATION
Liu B, Chen J, Zhang W, Huang Y, Zhao Y, Juneidi S, Dekebo A, Wang M, Shi L and Hu X (2022) The gastrodin biosynthetic pathway in *Pholidota chinensis* Lindl. revealed by transcriptome and metabolome profiling. *Front. Plant Sci.* 13:1024239. doi: 10.3389/fpls.2022.1024239

The gastrodin biosynthetic pathway in *Pholidota chinensis* Lindl. revealed by transcriptome and metabolome profiling

Baocai Liu[1,2,3,4,5], Jingying Chen[2]*, Wujun Zhang[2], Yingzhen Huang[2], Yunqing Zhao[2], Seifu Juneidi[6], Aman Dekebo[7,8], Meijuan Wang[9], Le Shi[1,3,4,5] and Xuebo Hu[1,3,4,5]*

[1]Institute for Medicinal Plants, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [2]Institute of Agricultural Bioresource, Fujian Academy of Agricultural Sciences, Fuzhou, China, [3]Innovation Academy of International Traditional Chinese Medicinal Materials, Huazhong Agricultural University, Wuhan, China, [4]National-Regional Joint Engineering Research Center in Hubei for Medicinal Plant Breeding and Cultivation, Huazhong Agricultural University, Wuhan, China, [5]Medicinal Plant Engineering Research Center of Hubei Province, Huazhong Agricultural University, Wuhan, China, [6]Department of Applied Biology, School of Natural Science, Adama Science and Technology University, Adama, Ethiopia, [7]Applied Chemistry Department, School of Applied Natural Sciences, Adama Science and Technology University, Adama, Ethiopia, [8]Institute of Pharmaceutical Sciences, Adama Science and Technology University, Adama, Ethiopia, [9]Shengnongjia Academy of Forestry, Shengnongjia, Hubei, China

Pholidota chinensis Lindl. is an epiphytic or lithophytic perennial herb of Orchidaceae family used as a garden flower or medicinal plant to treat high blood pressure, dizziness and headache in traditional Chinese medicine. Gastrodin (GAS) is considered as a main bioactive ingredient of this herb but the biosynthetic pathway remains unclear in *P. chinensis*. To elucidate the GAS biosynthesis and identify the related genes in *P. chinensis*, a comprehensive analysis of transcriptome and metabolome of roots, rhizomes, pseudobulbs and leaves were performed by using PacBio SMART, Illumina Hiseq and Ultra Performance Liquid Chromatography Tandem Mass Spectrometry (UPLC-MS/MS). A total of 1,156 metabolites were identified by UPLC-MS/MS, of which 345 differential metabolites were mainly enriched in phenylpropanoid/phenylalanine, flavone and flavonol biosynthesis. The pseudobulbs make up nearly half of the fresh weight of the whole plant, and the GAS content in the pseudobulbs was also the highest in four tissues. Up to 23,105 Unigenes were obtained and 22,029 transcripts were annotated in the transcriptome analysis. Compared to roots, 7,787, 8,376 and 9,146 differentially expressed genes (DEGs) were identified in rhizomes, pseudobulbs and leaves, respectively. And in total, 80 Unigenes encoding eight key enzymes for GAS biosynthesis, were identified. Particularly, glycosyltransferase, the key enzyme of the last step in the GAS biosynthetic pathway had 39 Unigenes candidates, of which, transcript28360/f2p0/1592, was putatively identified as the most likely candidate based on analysis of co-expression, phylogenetic analysis, and

10.3389/fpls.2022.1024239

homologous searching. The metabolomics and transcriptomics of pseudobulbs versus roots showed that 8,376 DEGs and 345 DEMs had a substantial association based on the Pearson's correlation. This study notably enriched the metabolomic and transcriptomic data of *P. chinensis*, and it provides valuable information for GAS biosynthesis in the plant.

KEYWORDS

***Pholidota chinensis*, gastrodin, metabolome, transcriptome, molecular mechanism**

Introduction

Pholidota chinensis Lindl, a member of the Orchidaceae family, is commonly known as "Shi Xian Tao" in China (Figures 1A–G). It is an epiphytic or lithophytic perennial herb widely distributed in southern China (Want et al., 2010; Dunn et al., 2011). The whole plants or pseudobulbs are used as ornamental flowers or folklore medicine in treating high blood pressure, dizziness and headache (Medincine, 2006). It is also orally administered in treating cough, tuberculosis, scrofula, diuresis, and infantile malnutrition as a traditional medicine by the Maonan tribal minorities in Guangxi province of China (Hong et al., 2015). Researchers have shown that polysaccharides, stilbenoids, dihydrophenanthrenes and triterpenoids are the main bioactive components in *P. chinensis* (Yao et al., 2008). These compounds exhibited multiple therapeutic activities including anti-tumor (Luo et al., 2018), anti-oxidant (Yang et al., 2016), anti-bacterial (Ti et al., 2020), anti-diabetic (Ren et al., 2020), anti-inflammatory (Wang et al., 2006), anti-pain and inhibit central nervous system (Liu et al., 2007; Rueda et al., 2014; Wang et al., 2016). Extensive chemical and pharmacological studies have laid a solid foundation for further application of these ingredients as medicine (Yang et al., 2016; Ti et al., 2020).

A Chinese patented medicine "Toutongding Syrup" is made up of *P. chinensis* for treating neurological headaches and concussion sequelae, and the gastrodin (4-hydroxymethylphenyl-β-Dglucopyranoside, GAS) and gastrodigenin (4-hydroxybenzyl alcohol, HBA) were shown to be the primary active ingredients (Weng, 2006; Zou et al., 2017). According to the established high performance liquid chromatography (HPLC) fingerprinting of *P. chinensis*, GAS was one of its analytical markers and its content in *P. chinensis* was higher than another traditional Chinese medicine *Gastrodia elata* Blume. (Zhang et al., 2019; Zhang et al., 2020). *G. elata* is the major source of GAS and HBA that is widely used to treat

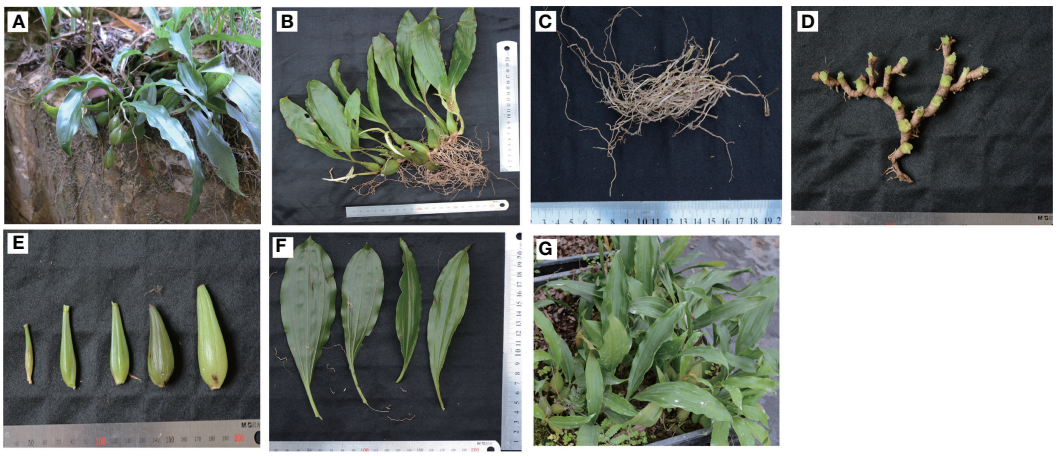


FIGURE 1
The morphological characteristics and growing environment of *P. chinensis*. **(A)**, Wild growth environment and plants growing on stone surfaces; **(B)**, the whole plant; **(C)**, Roots (designated as B1 for the rest of the sample analysis); **(D)**, Rhizomes (designated as B2 for the rest of the sample analysis); **(E)**, pseudobulbs (designated as B3 for the rest of the sample analysis); and **(F)** leaves (designated as B4 for the rest of the sample analysis) were analyzed; **(G)**, Artificially cultivated plants in a garden.

neurological disorders for centuries in China (Yuan et al., 2018; Zhang et al., 2019; Bae et al., 2022).

The GAS, a phenolic glycoside, is widely used to treat sedative, hypnotic, anticonvulsive and neuroprotective diseases in clinics (Liu and Yang, 2022). Synthesis of GAS is accomplished through glycosylation by a glycosyltransferase (GT) which transforms HBA with different glucose donors (Bai et al., 2016). Toluene was considered as the biosynthetic precursor for HBA that catalyzed by monooxygenase of cytochrome P450 (Tsai et al., 2016). The biosynthetic pathway of phenolic components, including GAS and HBA, were synthesized from phenylalanine through the phenylpropanoid pathway, which was speculated in *G. elata* by transcriptome analysis (Shan et al., 2021), and the biosynthetic pathway of 4-hydroxylbenzaldehyde and vanillin had been well studied in *Vanilla* spp. (Gallage et al., 2014; Wang et al., 2018). Interestingly, whether the precursor is toluene or phenylalanine, the last step is a GT that catalyzes the conversion of HBA to GAS in the GAS biosynthetic pathway (Tsai et al., 2016; Yin et al., 2020). However, the full, native biosynthetic pathway of GAS in *P. chinensis* has still not yet been documented.

To date, the transcriptome and metabolome studies provide effective strategies for understanding the molecular mechanisms of active ingredient formation (Hu et al., 2021; Chen et al., 2022). The combination of transcriptome and metabolome makes it possible to identify genes in any complex biological process with high sensitivity and accuracy (Song et al., 2022). The next-generation sequencing merges short reads into longer fragments by computation and it unavoidably affects the accuracy and integrity in fragments assembly (Cheng et al., 2021). In contrast, the third-generation sequencing technology has an advantage of sequencing reads as long as 100 kb but with lower sequencing accuracy (Liu et al., 2022). Therefore, the combination of next-generation sequencing and third-generation sequencing may assist to make up the shortcomings of each sequencing tool.

In the present study, based on multi-omics comparison, the GAS biosynthesis pathway and the genes involved in *P. chinensis* were elucidated. To the best of our knowledge, this study is the first to dissect the genes for GAS biosynthesis in *P. chinensis* and the same genus.

Materials and methods

Plant materials

P. chinensis was collected in June 2018 from Lingxia Village, Dongzhang Town, Fuqing City, Fujian Province of China (with 25°41.221′ N; 119°08.358′E and altitude 259 m). The plant sample was authenticated by Prof. Xuebo Hu (College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China), and Prof. Jingying Chen (Institute of Agricultural Bioresource Fujian Academy of Agricultural Sciences Fuzhou, China). The samples were collected from a wild forest (Figures 1A, B), and the plant part subjected to study was immediately separated into roots (B1, Figure 1C), rhizomes (B2, Figure 1D), pseudobulbs (B3, Figure 1E) and leaves (B4, Figure 1F). The samples with six independent biological replicates were washed clean, surface dried, and flash-frozen in liquid nitrogen, and then stored at -80°C until chemical composition analysis and RNA extraction. The rest plants were relocated to a greenhouse (Figure 1G).

UPLC-MS/MS conditions

Liquid chromatography-mass spectrometry (LC-MS) was used to analyze phytochemical constitutes *of P. chinensis*. The fresh materials of roots, rhizomes, pseudobulbs and leaves (0.1 g) was ground and extracted with 0.5 ml 80% (v/v) MeOH (LC-MS Grade, Thermo Fisher, USA). Samples were sonicated with a Vortex (Kylin-Bell, Jiangsu, China) and centrifuged for 20 min at 15, 000 g. The obtained supernatant was filtered through a 0.22 μm organic nylon needle filter (SCAA-104, ANPEL, Shanghai, China) and stored in a sample bottle (Want et al., 2010; Dunn et al., 2011). The extraction was performed in 6 replicates for statistical analysis. The metabolites were extracted and identified by the Novogene Bioinformatics Technology Co., Ltd.

Metabolite identification and quantification

The raw data of the mass spectrometry detection were imported into Compound Discoverer 3.1 (CD) software (Hao et al., 2018), used for extraction of metabolite feature. The characteristics of metabolites were obtained based on simple screening of data with their retention time, mass-to-charge ratio and peak alignment, molecular weight of the compound, and the mass deviation and adduct ion information. By matching fragment ions, collision energy and other information of each compound in the mzCloud database, the metabolites in the biological system were identified. Then, the QC Compounds with a CV (Coefficient of Variance) value less than 30% (Dai et al., 2017) were selected and used for final identification. Data quality control was performed to ensure the accuracy and reliability of the data. These metabolites were annotated using the Kyoto encyclopedia of genes and genomes (KEGG) database (http://www.genome.jp/kegg/), human metabolome database (HMDB) (http://www.hmdb.ca/) and Lipidmaps database (http://www.lipidmaps.org/). Principal components analysis (PCA) and Partial least squares discriminant analysis (PLS-DA) were performed with metaX (Wen et al., 2017). Metabolites with significant differences in content were identified according to the thresholds of variable importance

in projection (VIP) >1, fold change >2 or <0.5 and P value <0.05. Hierarchical clustering (HCA) and metabolite correlation analysis to reveal the relationship among the samples and metabolites (Chen et al., 2015). The metabolic pathway enrichment of differential metabolites (DEMs) was performed, when the ratio x/n > y/N (x, number of differential metabolites associated with this pathway; y, number of background (all) metabolites associated with this pathway; n: number of differential metabolites annotated by KEGG; N, number of KEGG-annotated background (all) metabolites), metabolic pathway was considered as enriched. When the P-value of metabolic pathway < 0.05, metabolic pathway was considered as statistically significant enrichment.

HPLC analyses of GAS and HBA

The major constituents of *P. chinensis*, GAS and HBA, were analyzed by HPLC system. GAS and HBA were extracted from dried and fresh *P. chinensis* tissues (roots and rhizomes, pseudobulbs as well as leaves) and measured, as described previously (Zhang et al., 2019), with slight modifications. Briefly, dried (0.5 g) and fresh powder of each tissue was extracted in 25 mL of 50% (v/v) methanol by ultrasonication for 30 min. Using the following chromatographic conditions, injection volume, 10 uL; column, Agilent SB-aq (5 μm, 4.6 mm × 250 mm); temperature, 30°C; flow rate, 1.0 mL min^{-1}; detector and UV-VIS detector at 220 nm. The mobile phases were containing 99.9% acetonitrile (A) and 0.05% phosphoric acid (B).

RNA extraction and Illumina sequencing

Frozen tissues were transferred to a mortar pre-cooled by liquid nitrogen and ground with a pestle. Total RNA was extracted from roots, rhizomes, pseudobulbs and leaves (4 tissues× 3 biological replications) by using the RNAprep Pure Plant Kit 264 (Tiangen, Beijing, China), following the manufacturer's instructions. The quality and quantity of RNA was checked by agarose gel electrophoresis and spectrophotometry (IMPLEN, CA, USA) and Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA), respectively. The RNA samples with A260/A280 of 1.8-2.2 were selected for cDNA synthesis.

An Illumina Hiseq platform was conducted using the NGS sequencing. The sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's recommendations, and index codes were added to attribute sequences to each sample. The RNA-seq experiment was performed at Novogene Bioinformatics Technology Co., Ltd. The raw data were uploaded to Sequence Read Archive (http://www.ncbi.nlm.nih.gov/) as accession PRJNA841044.

RNA extraction and PacBio ISO-Seq

To obtain a complete information of all transcripts, the full-length transcriptome sequencing was adopted in the present study. In order to reduce experimental error, the best RNA sample of three replicates was selected from each sample used in Illumina sequencing, and then mixed together in an equal quantity, as one sample, for SMRT sequencing. The Iso-Seq library was prepared according to the Isoform Sequencing protocol (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol as described by Pacific Biosciences (PN 100-092-800-03). The generated cDNA was re-amplified by PCR. A Qubit fluorometer (Life Technologies, Carlsbad, CA, USA) was used to determine fragment size distribution. The quality of the libraries was assessed using the Agilent Bioanalyzer 2100 system. The SMRT sequencing was performed using the Pacific Biosciences' real time sequencer using C2 sequencing reagents. The RNA-seq experiment was performed at Novogene Bioinformatics Technology Co., Ltd. The raw data were deposited to Sequence Read Archive (http://www.ncbi.nlm.nih.gov/) with accession. PRJNA806713.

The sequence data were processed using the SMRTlink 5.0 software (https://www.pacb.com/support/software-downloads/). Circular consensus sequence (CCS) was generated from subread BAM files parameters: min_length 200, max_drop_fraction 0.8, no_polish TRUE, min_zscore -9999, min_passes 1, min_predicted_accuracy 0.8, max_length 18000. The CCS.BAM files were output, which were then classified into full length and non-full length reads using pbclassify.py script, ignore polyA false, minSeq Length 200. Non-full length and full-length fasta files produced were then fed into the cluster step, which does isoform-level clustering, followed by final Arrow polishing, hq_quiver_min_accuracy 0.99, bin_by_primer false, bin_size_kb 1, qv_trim_5p 100, qv_trim_3p 30. Additional nucleotide errors in consensus reads were corrected using the Illumina RNA- seq data with the software LoRDEC (Salmela and Rivals, 2014). After all redundancy corrected, the consensus reads were removed by CD-HIT (Fu et al., 2012), and the final consensus isoforms were obtained for the subsequent analysis.

Functional annotation

Final consensus isoforms were searched used diamond v0.8.36 software against NCBI non-redundant (Nr), Swiss-Prot, euKaryotic Ortholog Groups (KOG)/Cluster of Orthologous Groups and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases with an E value threshold of $1e^{-5}$. The BLAST software with E-value ≤1e−5 was used for NT database analysis. The Hmmscan procedure was used in the Pfam database, and GO function categories were performed by

Blast2 GO v2.5 based on Pfam annotation. We use the confidence protein sequences of R. ferrugineus or closely related species for ANGLE training, and then run the ANGLE predictions for given sequences (Shimzu et al., 2006). Transcription factors (TF) were performed by the iTAK software (Zheng et al., 2016). Coding Potential Calculator (CPC) (Kang et al., 2017), and Pfam-scan (Finn et al., 2016) to predict the coding potential of transcripts.

RNA-seq read mapping and expression analysis

The consensus after de-redundancy correction was used the reference sequence (ref), and the clean reads of each sample obtained by Illumina sequencing were aligned to the ref using RSEM software (Li and Dewey, 2011). Further, RSEM software was used to count the comparison results of bowtie2, obtained the read count value of each sample compared to each gene, performed reads per kilo base of transcript per million mapped reads (FPKM) normalization, and then analyzed the expression level of the gene.

Differential expression analysis

Differential expression analysis of two conditions/groups was performed using the DESeq R package (Love et al., 2014). The DESeq provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P-values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted P-value <0.05 found by DESeq were assigned as differentially expressed. Gene Ontology (GO) enrichment analysis of differentially expressed genes or lncRNA target genes were implemented by the GOseq R package (http://www.bioconductor.org/packages/release/bioc/html/goseq.html), in which gene length bias was corrected. The KOBAS software (http://kobas.cbi.pku.edu.cn/download.php) was used to test the statistical enrichment of differentially expressed genes or lncRNA target genes in KEGG pathways.

Identification of candidate genes involved in the GAS biosynthesis pathway

Candidate genes belonging to the GAS biosynthetic pathway in *P. chinensis* were manually identified according to the annotated sequences in the above databases. Protein coding sequences (CDS) were acquired by Angel software (Shimzu et al., 2006), and multiple sequence alignment was carried out by MEGA7.0 (Kumar et al., 2016).

Phylogenetic analysis

Amino acid alignments were performed using Clustal W, and phylogenetic trees were built using MEGA7.0 (Kumar et al., 2016), employing the neighbor joining method with 1000 bootstrap replicates, and applying the default settings for other parameters. The GenBank accession numbers/transcript numbers for all sequences are shown in Supplementary Table S1.

Quantitative Real-Time PCR validation

To verify the accuracy of transcriptomic data, 6 DEGs in roots, rhizomes, pseudobulbs and leaves were selected for qRT-PCR verification. Primers were designed using Primer-BLAST on the NCBI website (Supplementary Table S2). RNA was reverse transcribed using a TransScript® RT/R1 reagent kit according to the manufacturer's instructions. The qRT-PCR was performed on an ABI QuantStudio 3. There were three biological and three technical replicates for each sample. The qRT-PCR reaction system (20 μL) consisted of 10 μL of Universal SYBR Green Fast qPCR Mix SYBR Green Master Mix, 1 μL of cDNA, 0.4 μL each of forward and reverse primer, and 8.2 μL of sterile water. The qRT-PCR procedure included 3 min of initiation, followed by 40 cycles at 95°C for 5 s, 60°C for 30 s, and 72°C for 12 s. Relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method and normalized according to the actin gene of β-tubulin.

Results

Tissue specific metabolites analysis

To explore the metabolite differences in roots, rhizomes, pseudobulbs and leaves of fresh *P. chinensis*, the samples were analyzed by UPLC-MS/MS (Figure 2A). A total of 1,156 (positive: 711, negative: 445) metabolites were identified (Supplementary Table S3). They were subsequently annotated in the KEGG, HMDB and Lipdmaps database, and annotations of 375, 462, and 129 metabolites were obtained, respectively (Figure 2B). The results of HCA showed that the DEMs were significantly varied in different organs, which were divided into five clusters (Figure 2C). The metabolites were comparable in leaves and pseudobulbs as well as roots and rhizomes. The relative content of GAS, Com_2638 in negative metabolites, was the highest in B3 and the lowest in B1 (Supplementary Table S4). Therefore, the comparison group of B3 and B1 was profiled and the results showed (Figures 2D, E) that 345 DEMs were mainly enriched in phenylpropanoid biosynthesis and phenylalanine metabolism (positive ion model), secondary metabolite biosynthetic process and flavone and flavonol

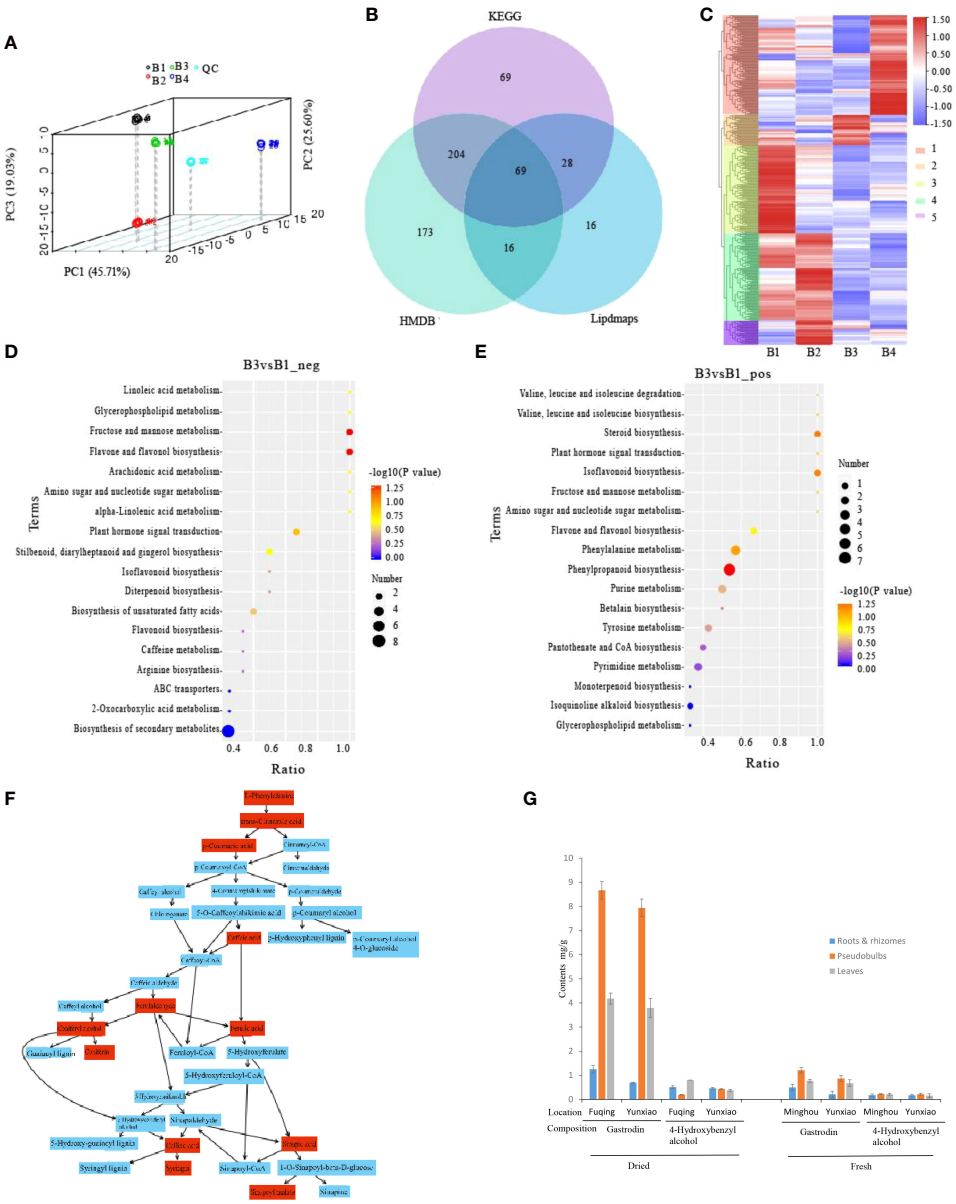


FIGURE 2
Metabolomes and differential expression of metabolomes (DEMs) of different tissues in *P. chinensis* by UPLC-MS analysis. **(A)**, PCA analysis of all samples. Scattered dots in different colors represent samples from different experimental groups; **(B)**, Venn diagram of annotations in KEGG, HMDB and Lipdmaps database; **(C)**, DEMs clustering heatmap of roots, rhizomes, pseudobulbs and leaves, and divided into five clusters in different color on the left heatmap and marked 1, 2, 3, 4, 5 on the right heatmap. Expression value was calculated based on Log2 Fold change. **(D)**, DEMs of B3 vs B1 on negative ion mode in KEGG pathway enrichment; **(E)**, DEMs of B3 vs B1 on positive ion mode in KEGG pathway enrichment; **(F)**, Phenylpropanoid biosynthesis by Metaboanalyst on line analyses. Red boxes were detected and annotated KEGG components; **(G)**, Gastrodin and 4-Hydroxybenzyl alcohol contents of different tissues by HPLC in dry and fresh *P. chinensis*. PCA, Principal component analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; Roots (B1), rhizomes (B2), pseudobulbs (B3) and leaves (B4).

biosynthesis (negative ion model). Phenylpropanoid biosynthesis, including C00079 (L-Phenylalanine), C00423 (trans-Cinnamate), C00811 (4-Coumarate), C01197 (Caffeate), C02666 (Coniferyl aldehyde), C00590 (Coniferyl alcohol), C00761 (Coniferin), C01494 (Ferulate), C02325 (Sinapyl alcohol), C01533 (Syringin), C00482 (Sinapate) and C02887 (Sinapoyl malate), were shown by Metaboanalyst on line analyses (Figure 2F). Phenylalanine or its derivatives may be precursors for the biosynthesis of GAS biosynthesis in *P. chinensis*.

GAS and HBA contents

To investigate GAS and HBA contents in roots, rhizomes, pseudobulbs and leaves, crude MeOH extracts of dried or fresh *P. chinensis* samples from different sites were analyzed by HPLC. The results indicated that both in dried and fresh samples, the GAS of pseudobulbs was the highest (of 0.867 and 0.794% in dried samples, of 0.121 and 0.087% in fresh samples), followed by leaves. The lowest content was detected in roots and rhizomes (Figure 2G). However, HBA did not show significant difference among sampled plant organs. This result was consistent with the result of UPLC-MS/MS.

Sequencing and analysis of RNA-Seq

To obtain the transcriptome expression profiles in *P. chinensis*, the RNA was extracted from roots, rhizomes, pseudobulbs and leaves, and mixed together in an equal quantity, as one sample for PacBio Sequel sequencing. As a result, 26.66 Gigabytes Polymerase Read Bases from PacBio Sequel were produced. A total of 506,905 circular consensus sequences (CCS) with an average length of 2,195 bp was obtained after filtration with full passes ≥ 1 and quality > 0.90 (Table 1). To further improve the accuracy, >6 Gb of raw reads were obtained for each sample from Illumina Hiseq platform performed using NGS sequencing (Supplementary Table S5). The redundant and similar sequences were removed using CD-HIT software. Finally, 23,105 Unigenes were obtained with an average length of 2,186 bp. It was taken as the reference transcriptome (Table 1 and Supplementary Figure S1).

A total of 22,029 transcripts were annotated functionally in this analysis by searching against the GO, KEGG, COG/KOG, NT, Pfam, NR, and Swiss-Prot databases with transcripts 15,307 (69.49%), 21,322(96.79%), 14,155 (64.26%), 15,664 (71.11%), 15,307 (69.49%), 21,512 (97.65%), and 18,731 (85.03%), respectively (Figure 3A and Supplementary Table S5). However, 8,577 (38.94%) transcripts were annotated in all seven databases (Figure 3B and Supplementary Table S6). Based on the homologous sequence alignment with NR database and statistical analysis, *Elaeis guineensis* was the most homology species (6,867 transcripts, 31.92%) (Figure 3C and Supplementary Table S7). In KEGG database annotation, the transcripts were grouped into six main categories: Cellular processes (1,406 transcripts), Environmental information processing (1,269 transcripts), Genetic information processing (2,336 transcripts), Human diseases (2,668 transcripts), Metabolism (4,795 transcripts), and Organismal systems (2,215 transcripts). In the metabolism of phenylalanine and terpenoid backbone biosynthesis, 58 and 63 transcripts were involved, respectively. (Figure 3E and Supplementary Table S7). A group of 128 transcripts were matched to phenylpropandoid biosynthesis (ko00940), including: phenylalanine ammonia-lyase (PAL), 4-coumarate-CoA ligase (4CL), trans-cinnamate 4-monooxygenase (CYP73A) (Supplementary Figure S2). GO analysis showed that 15,307 transcripts could be classified into three categories: cellular component, molecular function and biological process. However, the GO terms of metabolic process (7,492 transcripts, 48.94%) were the most annotated transcripts in the Biological process (Figure 3D and Supplementary Table S7). In KOG classifications, the transcripts yielded 26 functional categories (Figure 3F and Supplementary Table S7). Up to 611 transcripts were annotated in amino acid transport and metabolism and 501 transcripts were annotated in secondary metabolites biosynthesis, transport and catabolism. The number of annotated transcripts identified using the NT, Pfam, and Swiss-Prot databases were summarized in Supplementary Table S7. These transcripts involved in amino acid metabolism or secondary metabolism might be partially involved in GAS biosynthesis in *P. chinensis*.

Analysis of differentially expressed genes (DEGs)

To identify genes differently expressed in different tissues of *P. chinensis*, 12 cDNA libraries, were mapped to reference sequence (CD-HIT software de-redundant and corrected consensus sequence). The cDNA libraries were generated with mRNA from roots, rhizomes, pseudobulbs and leaves. The matched rate of all the clean reads was >45% (Supplementary Table S8). The expression level per sample was shown with read count and FPKM in Supplementary Table S9. Compared to roots, the 7,787 DEGs (2,907 up-regulated and 4,880 down-

TABLE 1 The characteristics of transcriptome sequences of *P. chinensis* by PacBio sequencing and Illumina.

Item	Number	Average length (bp)	N50	Min_Length	Max_Length
Subreads	14573961	1756	2075	51	——
Circular consensus sequences (CCS)	506905	2195	2542	53	14966
Full-Length non-chimeric Read (FLNC)	450366	2074	2431	56	14705
Polished consensus reads	45157	2054	2388	65	8171
Transcripts of after Illumina correction	45157	2054	2388	65	8171
Unigenes of after CD-HIT De-redundancy	23105	2186	2525	109	8171

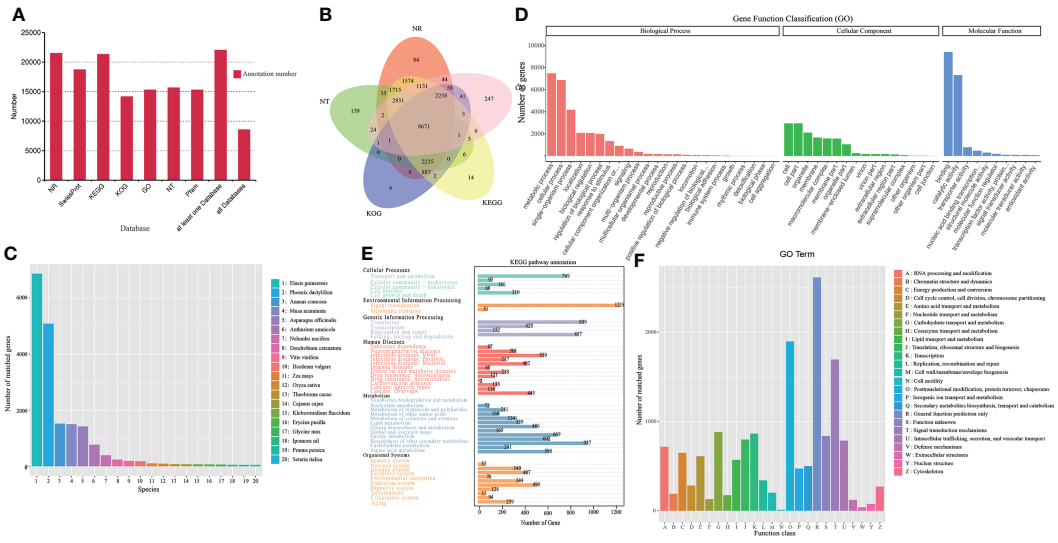


FIGURE 3
Transcripts functional annotation of *P. chinensis* in NR, NT, Pfam, KOG/COG, Swiss-prot, KEGG, GO databases and analysis. **(A)**: Statistics of the transcripts annotated in different databases. **(B)**: Venn diagram of annotations in NR, GO, KEGG, KOG, and NT databases. **(C)**: Distribution of the top 20 species with matched transcripts in the NR database. 1. *Elaeis guineensis*, 2. *Phoenix dactylifera*, 3. *Ananas comosus*, 4. *Musa acuminate*, 5. *Asparagus officinalis*, 6. *Anthurium amnicola*, 7. Nelumbo nucifera, 8. Dendrobium catenatum, 9. Vitis vinifera, 10. Hordeum vulgare, 11. Zea mays, 12. Oryza sativa, 13. Theobroma cacao, 14. Cajanus cajan, 15. Klebsormidium flaccidum, 16. Erycina pusilla, 17. Glycine max. 18. Ipomoea nil, 19. Prunus persica, 20. Setaria italic. **(D)**: Distribution of GO terms for all annotated transcripts in biological process, cellular component, and molecular function. **(E)**: KEGG pathways annotation by all transcripts. **(F)**: KOG categories of the annotation transcripts. NR, Non-Redundant Protein Sequence Database; NT, Nucleotide Sequence Database; Pfam, database of a large collection of protein families; KOG/COG, EuKaryotic Orthologous Groups of proteins/Clusters Orthologous Groups of proteins; Swiss-prot, annotated protein database and as such an absolute requirement in the toolbox of any protein chemist; KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, gene ontology.

regulated), 8,376 DEGs (3,210 up-regulated and 5,166 down-regulated) and 9,146 DEGs (3,581 up-regulated and 5,565 down-regulated) were identified in rhizomes, pseudobulbs and leaves extracts, respectively (Figure 4A). And in total, 16,175 DEGs unigenes in all four tissues were identified. Among different tissues DEGs, only 357 common genes were expressed in all four compared tissues (Figure 4B).

To reveal the biological significance of these DEGs, function annotation and enrichment analysis were performed by GO and KEGG database. The analysis of GO functional classification indicated that all the DEGs of B3 and B1 comparison group were grouped into 34 functional groups, including 15 molecular function categories, 15 biological processes, and 4 cellular components. (Figure 4C). Metabolic process and single-organism in the biological processes, and catalytic activity in the molecular function were the most enriched terms. However, in almost all terms, down-regulated genes were higher than up-regulated genes. To further illustrate the alterations of gene expression between B3 and B1, the KEGG analysis of all the DEGs of B3 and B1 comparison group was made. The DEGs were enriched in linoleic acid metabolism, flavonoid biosynthesis, phenylpropanoid biosynthesis and others (Figure 4D). However, while the up-regulated DEGs of B3 and B1 were mainly enriched in photosynthesis - antenna proteins, phenylpropanoid biosynthesis; the down-regualted DEGs of B3 and B1 were mainly enriched in flavonoid biosynthesis, linoleic acid metabolism and others terms (Supplementary Figures 3-4.). 28 transcripts were up-regulated in phenylpropanoid biosynthesis of B3 and B1 comparison group, including encoding 4CL, cinnamyl-alcohol dehydrogenase, peroxidase, and beta-glucosidase. These enzymes might be critical for the synthesis of GAS precursors. In addition, the transcripts transcript28360/f2p0/1592, transcript25791/f2p0/1719, etc. had significant expression differences in B3 and B1 comparison group and the *p* value was close to zero.

The candidate genes involved in GAS biosynthesis pathway

Based on the KEGG pathway (map00940, map00996) analysis as reported in *G. elata* (Shan et al., 2021), the putative GAS biosynthetic pathway of *P. chinensis* is shown in Figure 6. The biosynthesis of GAS primarily initiated from the L-phenylalanine, which is derived from the common phenylpropanoid biosynthesis pathway that is broadly distributed in plants (Zhang et al., 2020; Rai et al., 2021). A total of 80 unigenes were identified that encoding eight key

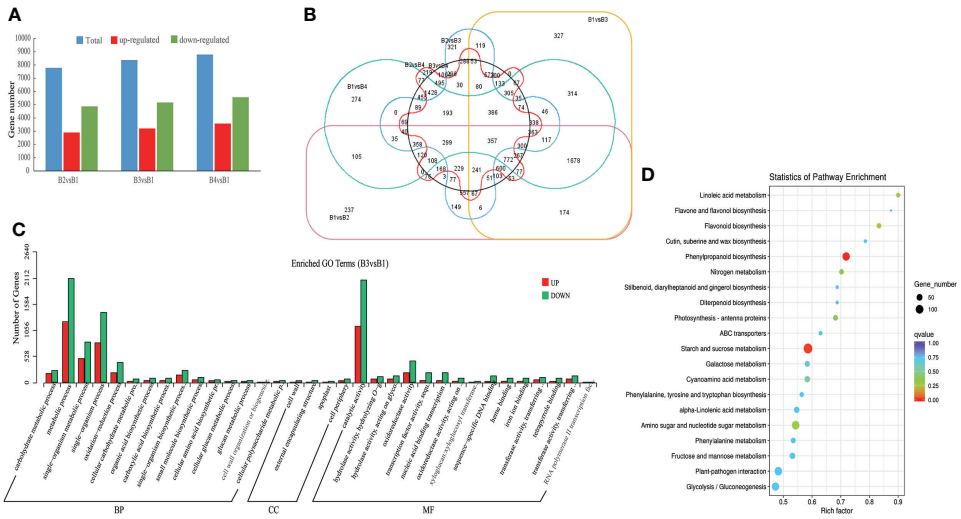


FIGURE 4
Different expression genes (DEGs) of roots (B1), rhizomes (B2), pseudobulbs (B3) and leaves (B4) in *P. chinensis*. **(A)**, DEGs statistics of B2 vs B1, B3 vs B1, B4 vs B1. The blue bar represents all DEGs, red bar represents up-regulated DEGs, and green bar represents down-regulated DEGs; **(B)**, Venn diagram of DEGs in different comparison groups. The circle color of pink, orange, green, blue, red and black represents B1 vs B2, B1 vs B3, B1 vs B4, B2 vs B3, B2 vs B4 and B3 vs B4, respectively; **(C)**, Enriched GO terms of DEGs in B3 vs B1. The red bar represents up regulation and the blue bar represents down regulation. **(D)**, Enriched KEGG pathway of DEGs in B3 vs B1. The size of the dots represented the number of DEGs. Red and blue represented high and low expression levels, respectively. GO, gene ontology; DEGs, different expression genes.

enzymes controlling GAS biosynthesis: phenylalanine ammonia-lyase (PAL), trans-cinnamate 4-monooxygenase (CYP73A), 4CL, shikimate O-hydroxycinnamoyltransferase (HCT), 5-O-(4-coumaroyl)-D-quinate 3'-monooxygenase (C3H), caffeoyl coenzyme A-O-methyltransferase (CCoAOMT), alcohol dehydrogenases (ADH) and GT. The relative expression levels of the DEGs in the different tissues were showed in the heatmap (Figure 5). However, as the last key enzyme, GT, which catalyzed the GAS synthesis from HBA with UDP-glucose, had 39 Unigenes. These unigenes were divided into four types according to the types of encoded enzymes, including GT1 (3 beta-glucosyltransferase (2.4.1.173)), GT2 (cis-zeatin O-glucosyltransferase (2.4.1.215)), GT3 (hydroquinone glucosyltransferase (2.4.1.218)) and GT4 (others glucosyltransferase (2.4.1-)). Some annotated transcripts GTs were differently expressed in targeted tissues of the studied plant: transcript28360/f2p0/1592, transcript16563/f4p0/2237, transcript19586/f3p0/2041 and transcript25251/f2p0/1759 were highly expressed in pseudobulbs, and lower in leaves, least in roots and rhizomes (Figure 5).

To verify the accuracy of RNA-seq data, the quantitative real-time PCR (qRT-PCR) was used to validate differential gene expression levels of roots, rhizomes, pseudobulbs and leaves with gene-specific primers (Supplementary Table S2). The results showed that the gene relative expression profile was almost consistent with the RNA-seq data. It further demonstrated the credibility of the data generated in the present study (Figure 6).

Identification of glucosyltransferase

The previous results strongly suggest that the AsUGT, a serpentwood-derived GT convert HBA to GAS with high catalytic efficiency in yeast, compared with UGT73B6 from *Rhodiola sachalinensis* in *Escherichia coli* (Bai et al., 2016; Yin et al., 2020). In *P. chinensis*, significant homology was found using queries from the encoding sequences of the AsUGT and UGT73B6 genes. The most similar transcripts were transcript17418/f3p0/2174 (55.72% identical), transcript28360/f2p0/1592 (47.38% identical), transcript27824/f2p0/1641 (49.90% identical), and transcript29686/f2p0/1549 (48.86% identical) (Supplementary Table S10). To further characterize the GTs, a phylogenetic tree was constructed based on *P. chinensis* and other plant GT protein sequences, including *Zea mays, Arabidopsis thaliana, Apostasia shenzhenica, Rhodiola sachalinensis, Dendrobium catenatum, Rhodiola sachalinensis* and *Phalaenopsis equestris* (Supplementary Table S1). All GT members in *P. chinensis* were divided into eight phylogenetic groups. Among them, AsUGT (Rse_Q9AR73.1), the reference UGT was clustered in the same clade with transcript17418/f3p0/2174, transcript28360/f2p0/1592, transcript24635/f2p0/1797 and transcript28208/f27p0/1567, whereas UGT73B6 (Rsa_AAS55083.1) was clustered in the same clade with transcript19941/f2p0/2048 (Figure 6). The gene expression level of these transcripts exhibited great tissue-specific tendency. While others were highly expressed in the roots, transcript28360/f2p0/1592 was the only one with high expression in the pseudobulbs. Meanwhile, homologous searching by using

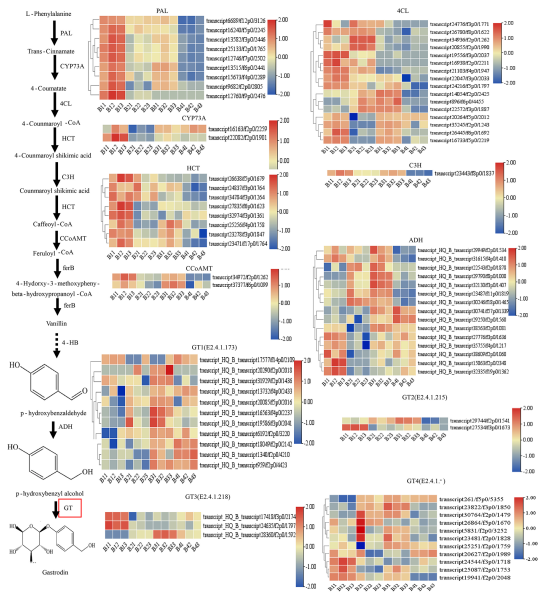


FIGURE 5
Putative gastrodin biosynthesis pathway in *P. chinensis*. This pathway was constructed based on the KEGG pathway (map00940, map01061) and literature references. The expression levels deduced from the RMPK of each Unigene that encodes the relevant enzyme, were shown as heat map, whereas roots (triplicates as B11, B12, B13), rhizomes (triplicates as B21, B22, B23), pseudobulbs (triplicates as B31, B32, B33) and leaves (triplicates as B41, B42, B43) were separately analyzed. The expression value was calculated based on the Log2 Fold change. Red and blue represented high and low expression levels, respectively. Non-dashed line arrows represent identified enzymatic reactions, and dashed line arrows represent multiple enzymatic reactions through multiple steps.

queries from the transcript TRINITY_DN50323_c0_g1(which might participate in glucosylation in the GAS biosynthetic pathway of *G. elata*) (Tsai et al., 2016), showed significant homology with transcript20627/f2p0/1989 (84.29% identical) in *P. chinensis* and XM_020720017.1 (83.4% identical) in *P. equestris*, respectively (Supplementary Table S11). However, the expression level of transcript20627/f2p0/1989 was the highest in leaves and lower in pseudobulbs, which might be not the best candidate gene (Figure 7).

Integrative analysis of transcriptomics and metabolomics

To obtain a deeper understanding, a multi-omics analysis was performed. These analyses integrated the metabolomics with the transcriptomic data. In negative/positive ion mode, top 50 DEMs (sorted by p value from small to large) and top 100 DEGs (sorted by *p* value from small to large) of B2 and B1 comparison group, B3 and B1 comparison group, B4 and B1 comparison group were shown in Supplementary Figures 5–10. These DEMs and DEGs had a stronger positive or negative connection (R>0.9). To identify the major biochemical pathways and signal transduction involved pathways of DEMs and DEGs, all DEMs and DEGs were matched to the KEGG pathway. The results revealed that the DEMs and DEGs were main enriched in phenylpropanoid biosynthesis and linoleic acid metabolism of B2 and B1 comparison group, phenylpropanoid biosynthesis and amino sugar and nucleotide sugar metabolism of B3 and B1 comparison group, phenylpropanoid biosynthesis and flavonoid biosynthesis of B4 and B1 comparison group (Figure 8). The phenylpropanoid biosynthesis may be critical for GAS

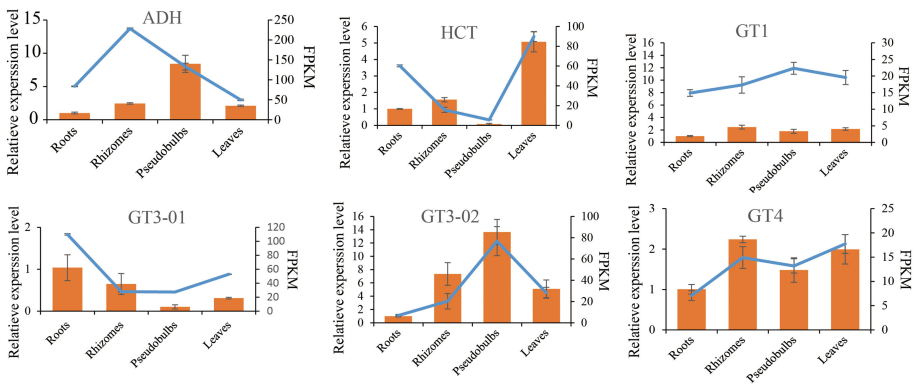


FIGURE 6
Verification of six selected DEGs by qRT-PCR. Comparison of RNA-seq data (Blue line chart) with qRT-PCR data (Yellow bar graph). The relative qRT-PCR expression level of selected DEGs is shown on the y-axis to the left. *β-tubulin* (TUB) was used as the internal control. Three biological replicates were used. The normalized expression level (FPKM; expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) of RNA-seq is indicated on the *y*-axis to the left.

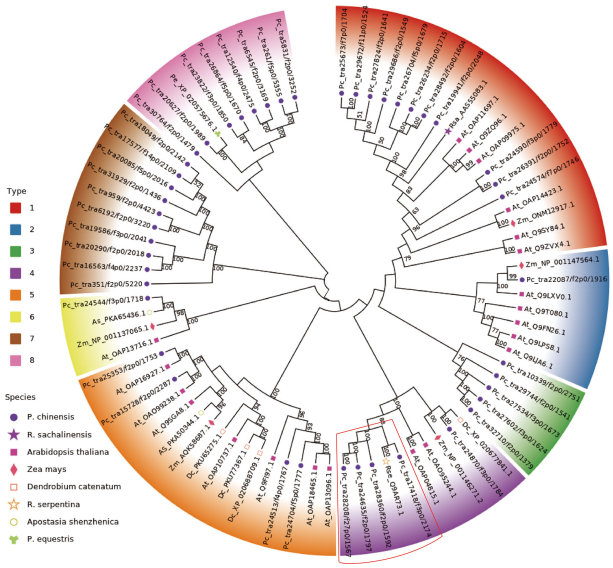


FIGURE 7
Phylogenetic tree analysis of glucosyltransferase. The constructed phylogenetic tree includes the amino acid sequences of 80 UGT enzymes of 8 species represented by different shapes on the left legends (*Pholidota chinensis, Rhodiola sachalinensis, Arabidopsis thaliana, Zea mays, Dendrobium catenatum, Rauvolfia serpentina serpentine, Apostasia shenzhenica, Phalaenopsis equestris*). The 80 UGT enzymes are clustered into 8 types and displayed with different colors on the left legends and chart. The candidate gene clustering positions obtained in this study are in the red box.

biosynthesis in *P. chinensis*. There were a notable association (R>0.9) between 8,376 DEGs and 345 DEMs based on the Pearson's correlation coefficient in B3 and B1 comparison group (Supplementary Table S12). And the results showed that no matter in positive or negative ion mode, there were metabolites that had significant correlation with most genes, such as coniferyl aldehyde, butylparaben, 3_4_5-trimethoxycinnamic acid, monobenzyl phthalate etc. Coniferyl aldehyde, a natural non-toxic and anti-inflammatory phenolic compound extracted from edible and medicinal plants (Wang et al., 2020), might be involved in the biosynthesis of GAS (Huccetogullari et al., 2019) and it had significant associations (R>0.9) with 6266 DEGs (Supplementary Table S13).

To further understand the relationship between metabolites and genes in common pathway, DEGs and DEMs of B3 and B1 comparison group were simultaneously mapped to the KEGG pathway. The results in negative ion mode showed that 2 DEMs and 42 DEGs (fructose and mannose metabolism), 2 DEMs and 7 DEGs (flavone and flavonol biosynthesis), 2 DEMs and 115 DEGs (plant hormone signal transduction), 2 DEMs and 11 DEGs (stilbenoid, diarylheptanoid and gingerol biosynthesis) enriched the corresponding biological processes (Figure 8C). However, the results in positive ion mode showed that 7 DEMs and 92 DEGs, 2 DEMs and 19 DEGs, 4 DEMs and 31 DEGs, 2 DEMs and 7 DEGs were enriched phenylpropanoid biosynthesis, steroid biosynthesis, phenylalanine metabolism, and flavone and flavonol biosynthesis, respectively (Figure 8D). Furthermore, the putative candidate gene transcript28360/f2p0/1592 in GT3 was significantly negative correlated with coniferylaldehyde and spermidine, but had a significantly positive correlation with GAS, isoeugenol and sinapoyl malate (Supplementary Table S14). These results are consistent with transcriptome or metabolome results.

Discussion

The GAS is the second compound identified from the plant *G. elata* after vanilyalcohol. It is a phenolic glycoside that chemically known as 4-hydroxybenzyl alcohol-4-*O*-β-*D*-glucopyranoside. And it is also the main bioactive constituent of another TCM *Rhizoma Gastrodiae* (Tao et al., 2009; Liu et al., 2018). Being the largest and the most widespread class of plant secondary metabolites, phenolics have been extensively researched due to the diverse health benefits. Some of these include flavonoids, lignans, coumarins, chalcones, and phenolic acids, which participate in the regulation of plant growth, seed germination, and in defense responses (Acosta-Estrada et al., 2014; Naikoo et al., 2019). The GAS content is the most appreciated analytical marker for the quality standardization of *P. chinensis* (Zhang et al., 2019). The mechanism of GAS action is gradually being understood and recognized (Chen et al., 2022; Yang et al., 2022). Several reports have shown that the content of GAS in *P. chinensis* was higher than the content in *G. elata*, one of the major sources of GAS (Zhang et al., 2019; Wang et al., 2022).

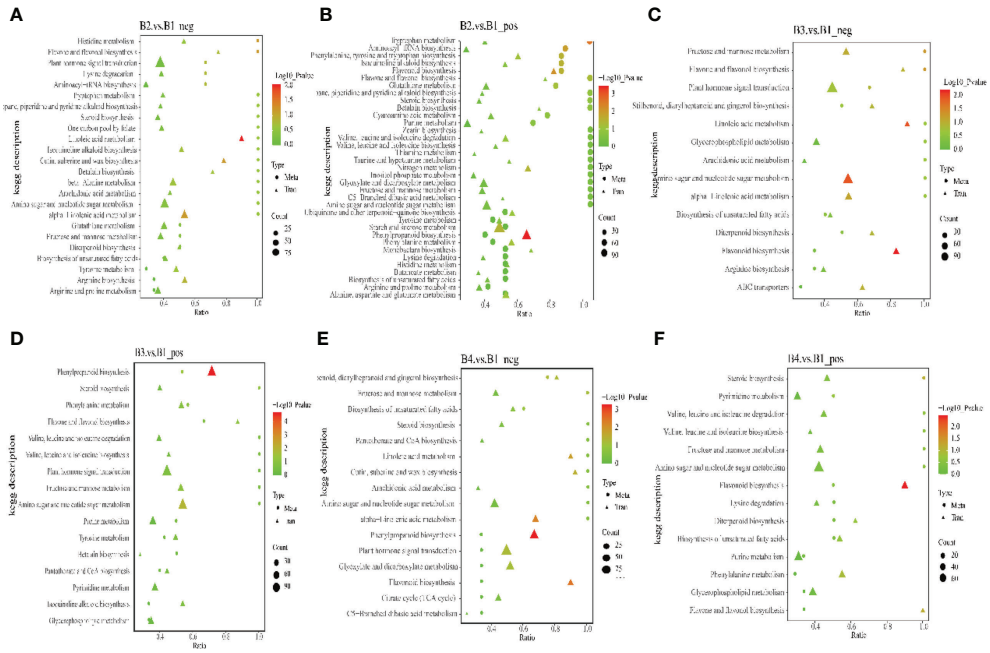


FIGURE 8
The KEGG enrichment bubble chart of co-expression DEMs and DEGs in B2 vs B1, B3 vs B1 and B4 vs B1 group. Dots represent DEMs. Triangles represent DEGs. Dots or triangles size represent enriched in the pathway number of metabolites or genes. "P value" is the *p* value of the transcription or metabolism pathway enrichment. **(A, C, E)**, negative ion mode, **(B, D, F)**, positive ion mode.

Metabolomics, especially untargeted metabolomics using LC-MS, is considered to be the best omics technique to represent the phenotypes because its data analysis based on the state of biochemical activity in the living organism (Fraisier-Vannier et al., 2020; Zeki et al., 2020). In this study, 1,156 metabolic differences were identified in roots, rhizomes, pseudobulbs and leaves of fresh *P. chinensis* using the UPLC-MS/MS. Many of the metabolites were phenolic compounds that were mainly enriched in phenylpropanoid biosynthesis, flavone and flavonol biosynthesis, and phenylalanine metabolism. And some of these compounds were known to exhibit antioxidant, antidiabetic, and anti-inflammatory activities (Ren et al., 2020). Simultaneously, HPLC analysis revealed that pseudobulbs were the primary tissue of GAS, followed by leaves, roots and rhizomes.

Some medicinal plants lack genomic information due to the wide variety, limiting some research, such as the integrity of transcriptome assembly (Cheng et al., 2021). The combination of third-generation sequencing and second-generation sequencing is an effective method for gene mining without reference genome (Liu et al., 2022). In the present study, the mean length of predicted unigenes (2,186 bp) the N50 length (2,525 bp) were longer than some other traditional medicine including *G. elata* (Tsai et al., 2016), *Dendrobium officinale* (Li et al., 2021) and *Dendrobium sinense* (Zhang et al., 2021), indicated that the transcriptome assembly were of high reliability and quality.

The GAS was synthesized from HBA with UDP-glucose *via* glycosylation catalyzed through GT, and HBA was synthesized from cresols (toluene) degradation through two steps of hydroxylation *via* monooxygenase in *G. elata* (Tsai et al., 2016). The monooxygenase 1.14.13.-, which were reported in *G. elata* (Tsai et al., 2016), were not discovered in *P. chinensis*. However, it was also reported that GAS could be synthesized *via* the phenylpropanoid pathway, and the PAL, C4H, 4CL and GT are the key genes in biosynthetic pathway of GAS in *G. elata* (Shan et al., 2021). It is worth noting that GAS was synthesized from glucose by an artificial microbial pathway with key genes of ADHs and GT in *Saccharomyces cerevisiae* and *Escherichia coli*, respectively (Bai et al., 2016; Yin et al., 2020). Glycosylation is often the last step in the biosynthesis of natural products in plants and plays an important role in a variety of biosynthetic pathways (He et al., 2022). According to the above analysis, although the starting point of GAS synthesis is different, but the last step is same, that is, GT catalyzes the conversion of HBA to GAS. In this study, putatively 80 unigenes involved in the biosynthetic pathway of GAS in *P. chinensis* were identified including genes for PAL, CYP73A, 4CL, HCT, C3H, CCoAOMT, ADH and GT. The GT (39 unigenes) were divided into four subgroups according to the types of encoded enzymes. Among all transcripts being found, transcript28360/f2p0/1592, transcript16563/f4p0/2237, transcript19586/f3p0/

2041 and transcript25251/f2p0/1759 were highly expressed in pseudobulbs, and lower in other targeted plant parts (leaves, roots and rhizomes), which could be key candidates. Based on phylogenetic tree analysis, the transcript28360/f2p0/1592 in GT3 was deduced as the best candidate gene because it shares a highly homologous sequence with AsUGT, which was identified as the plant-derived GT that converts HBA to GAS with high catalytic efficiency in yeast (Yin et al., 2020).

Integrated analysis of transcriptome and metabolome provides an efficient approach for the research of metabolic networks and key genes. For example, multi-omics were applied for flavonoid biosynthesis in a purple tea plant cultivar (Song et al., 2022), the response of *Zanthoxylum bungeanum* and apple to different stresses (Li et al., 2021; Sun et al., 2021). It is worth noting that the proteome (Camp et al., 2022) is also often analyzed together with the transcriptome or metabolome, but this study has not yet performed a proteome of *P. chinensis*. Coniferylaldehyde, coniferyl alcohol, isoeugenol and sinapoyl malate were members of dominant group of volatile compounds, and those volatile phenyl propene formation might takes two enzymatic steps with lignin, and they were perhaps involved in the synthesis of GAS precursors (Ramya et al., 2020).

Other than our focus on the dissection of transcriptomic and metabolic profiles of the extremely less studied traditional medicine for understanding of the GAS biosynthetic pathway, our study, as the first in the genus, also provides useful data for other basic researches on *P. chinensis* and related species. For example, the plant undergoes a well differentiated developmental stage of pseudobulb and our data could be lent for mining of the molecular mechanisms of pseudobulb development.

Conclusion

The GAS is an important active component of a traditional Chinese medicine, but its biosynthetic pathway in *P. chinensis* is still unclear. In the present study, the biosynthetic pathway of GAS in *P. chinensis* was speculated by combination of transcriptome and metabolome analysis. Unigenes involved in the biosynthetic pathways, as well as the metabolites, were identified. Besides commonly known unigenes in the synthetic pathway for PAL, CYP73A, 4CL, HCT, C3H, CCoAOMT, and ADH, best candidates for the last synthetic step of GAS, the transcript28360/f2p0/1592, were assured by bioinformatics. Since the growth of the plant is extremely slow and it is not practical to cultivate it in large area to gain sufficient yield and profit, the biosynthetic pathway disclosed in this study, especially the last unique GT for GAS, will be extremely useful for possible biosynthetic engineering of the chemical in microbial platforms. To the best of our knowledge, this study is the first exploration of the genes involved in the GAS biosynthesis in *P. chinensis* and plants in the same genus.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA841044 and PRJNA806713.

Author contributions

XH conceived, supervised, and wrote-reviewed the manuscript. SJ, AD, and JC reviewed the draft. BL wrote and reviewed the draft. BL, WZ, YH, YZ, MW, and LS performed the experiments and carried out the analysis. BL and XH designed the experiments. XH and JC co-founded and co-administrated the project. All authors approved the final version.

Funding

This research financially supported by Fundamental scientific research projects of non-profit scientific research institutes in Fujian Province (2020R1034003), Collaborative Innovation of the Fujian Provincial People's Government (XTCXGC2021003), Fujian Provincial Finance Special Project to Science and Technology Innovation Team of Fujian Academy of Agricultural Sciences (CXTD2021014-2), Young Talents in Science and Technology Project of Fujian Academy of Agricultural Sciences (YC2021005). This research is also funded by Shengnongjia Academy of Forestry, Hubei, China (No. SAF202102), Hubei Technology Innovation Center for Agricultural Sciences - '2020 key technology research and demonstration project of safe and efficient production of genuine medicinal materials' (No. 2020-620-000-002-04), China-Bulgaria science and technology exchange meeting for traditional medicine (year of 2020), Pinghu Municipal Bureau of Agriculture and Rural Affairs (PH2020005) to XH.

Conflict of interest

The 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1024239/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Assemble Unigene length distribution of *P. chinensis* by PacBio sequencing and Illumina data correction, Abscissa (length/bp), ordinate (number of genes).

SUPPLEMENTARY FIGURE 2
The transcripts were matched to phenylpropandoid biosynthesis (ko00940)

SUPPLEMENTARY FIGURE 3
Up-regulated DEGS of B3 vs B1. The size of the dots indicates the number of DEGS in this pathway, and the colors of the dots correspond to different q value ranges, which is closer zero (red color) and more significant enrichments.

SUPPLEMENTARY FIGURE 4
Down-regulated DEGS of B3 vs B1. The size of the dots indicates the number of DEGS in this pathway, and the colors of the dots correspond to different q value ranges, which is closer zero (red color) and more significant enrichments.

SUPPLEMENTARY FIGURE 5
The correlation heat map in negative ion mode of co-expression B2 vs B1 DEMs & B2 vs B1 DEGs.

SUPPLEMENTARY FIGURE 6
The correlation heat map in positive ion mode of co-expression B2 vs B1 DEMs & B2 vs B1 DEGs.

SUPPLEMENTARY FIGURE 7
The correlation heat map in negative ion mode of co-expression B3 vs B1 DEMs & B3 vs B1 DEGs.

SUPPLEMENTARY FIGURE 8
The correlation heat map in positive ion mode of co-expression B3 vs B1 DEMs & B3 vs B1 DEGs.

SUPPLEMENTARY FIGURE 9
The correlation heat map in negative ion mode of co-expression B4 vs B1 DEMs & B4 vs B1 DEGs.

SUPPLEMENTARY FIGURE 10
The correlation heat map in positive ion mode of co-expression B4 vs B1 DEMs & B4 vs B1 DEGs.

SUPPLEMENTARY TABLE 3
Species of components in positive and negative ion modes by UPLC-MS/MS.

SUPPLEMENTARY TABLE 4
The GAS relative content of different tissue by UPLC-MS/MS.

SUPPLEMENTARY TABLE 5
Sequencing data quality from Illumina Hiseq.

SUPPLEMENTARY TABLE 6
Annotation and classification of detected unigenes by the seven datasets.

SUPPLEMENTARY TABLE 7
The annotation of NR, GO, KEGG, KOG, NT, Pfam, Swissprot.

SUPPLEMENTARY TABLE 8
The comparison rate of the second-generation data map to the third-generation data.

SUPPLEMENTARY TABLE 9
The expression level of each sample.

SUPPLEMENTARY TABLE 10
Based on amino acid blast results in *P. chinensis* with AsUGT and UGT73B6 genes as query.

SUPPLEMENTARY TABLE 11
Blast results of GT in *P. chinensis* and *Phalaenopsis equestris* with candidate gene TRINITY_DN50323_c0_g1 of *Gastrodia elata* as query.

SUPPLEMENTARY TABLE 12
The correlation in positive and negative ion mode of DEMs & DEGs based on r^2>0.91 and p<0.01.

SUPPLEMENTARY TABLE 13
Coniferyl aldehyde correlation DEGs based on r^2>0.91 and p<0.01.

SUPPLEMENTARY TABLE 14
The Correlation putative candidate gene transcript28360/f2p0/1592 in GT3 with metabolites.

References

Acosta-Estrada, B. A., Gutierrez-Uribe, J. A., and Serna-Saldivar, S. O. (2014). Bound phenolics in foods, a review. *Food Chem.* 152, 46–55. doi: 10.1016/j.foodchem.2013.11.093

Bae, E. K., An, C., Kang, M. J., Lee, S. A., Lee, S. J., Kim, K. T., et al. (2022). Chromosome-level genome assembly of the fully mycoheterotrophic orchid gastrodia elata. *G3 Genes|Genomes|Genetics (Bethesda)* 12 (3), 3. doi: 10.1093/g3journal/jkab433

Bai, Y. F., Yin, H., Bi, H. P., Zhuang, Y. B., Liu, T., and Ma, Y. H. (2016). De novo biosynthesis of gastrodin in escherichia coli. *Metab. Eng.* 35, 138–147. doi: 10.1016/j.ymben.2016.01.002

Camp, E. F., Kahlke, T., Signal, B., Oakley, C. A., Lutz, A., Davy, S. K., et al. (2022). Proteome metabolome and transcriptome data for three symbiodiniaceae under ambient and heat stress conditions. *Sci. Data* 9, 153. doi: 10.1038/s41597-022-01258-w

Cheng, Q. Q., Ouyang, Y., Tang, Z. Y., Lao, C. C., Zhang, Y. Y., Cheng, C. S., et al. (2021). Review on the development and applications of medicinal plant genomes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.791219

Chen, J., Tang, Y. J., Kohler, A., Lebreton, A., Xing, Y. M., Zhou, D. Y., et al. (2022). Comparative transcriptomics analysis of the symbiotic germination of d. officinale (Orchidaceae) with emphasis on plant cell wall modification and cell wall-degrading enzymes. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.880600

Chen, X., Wang, J., He, Z., Liu, X., Liu, H., and Wang, X. (2022). Analgesic and anxiolytic effects of gastrodin and its influences on ferroptosis and jejunal microbiota in complete freund's adjuvant-injected mice. *Front. Microbiol.* 13. doi: 10.3389/fmicb.2022.841662

Chen, X., Xie, C., Sun, L., Ding, J., and Cai, H. (2015). Longitudinal metabolomics profiling of parkinson's disease-related alpha-synuclein A53T transgenic mice. *PloS One* 10, e0136612. doi: 10.1371/journal.pone.0136612

Dai, W., Xie, D., Lu, M., Li, P., Lv, H., Yang, C., et al. (2017). Characterization of white tea metabolome: Comparison against green and black tea by a nontargeted metabolomics approach. *Food Res. Int.* 96, 40–45. doi: 10.1016/j.foodres.2017.03.028

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). 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. doi: 10.1038/nprot.2011.335

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Fraisier-Vannier, O., Chervin, J., Cabanac, G., Puech, V., Fournier, S., Durand, V., et al. (2020). MS-CleanR: A feature-filtering workflow for untargeted LC-MS based metabolomics. *Anal. Chem.* 92, 9971–9981. doi: 10.1021/acs.analchem.0c01594

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gallage, N. J., Hansen, E. H., Kannangara, R., Olsen, C. E., Motawia, M. S., Jorgensen, K., et al. (2014). Vanillin formation from ferulic acid in vanilla planifolia is catalysed by a single enzyme. *Nat. Commun.* 5, 4037. doi: 10.1038/ncomms5037

Hao, L., Wang, J., Page, D., Asthana, S., Zetterberg, H., Carlsson, C., et al. (2018). Comparative evaluation of MS-based metabolomics software and its application to preclinical alzheimer's disease. *Sci. Rep.* 8, 9291. doi: 10.1038/s41598-018-27031-x

He, B., Bai, X., Tan, Y., Xie, W., Feng, Y., and Yang, G. Y. (2022). Glycosyltransferases: Mining, engineering and applications in biosynthesis of glycosylated plant natural products. *Synth Syst. Biotechnol.* 7, 602–620. doi: 10.1016/j.synbio.2022.01.001

Hong, L., Guo, Z., Huang, K., Wei, S., Liu, B., Meng, S., et al. (2015). Ethnobotanical study on medicinal plants used by maonan people in China. *J. Ethnobiol. Ethnomed.* 11, 32. doi: 10.1186/s13002-015-0019-1

Huccetogullari, D., Luo, Z. W., and Lee, S. Y. (2019). Metabolic engineering of microorganisms for production of aromatic compounds. *Microb. Cell Fact* 18, 41. doi: 10.1186/s12934-019-1090-4

Hu, H. C., Fei, X. T., He, B. B., Luo, Y. L., Qi, Y. C., and Wei, A. Z. (2021). Integrated analysis of metabolome and transcriptome data for uncovering flavonoid components of zanthoxylum bungeanum maxim. leaves under drought stress. *Front. Nutr.* 8. doi: 10.3389/fnut.2021.801244

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., et al. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45, W12–W16. doi: 10.1093/nar/gkx428

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* 12 (1), 323. doi: 10.1186/1471-2105-12-323

Li, N., Dong, Y., Lv, M., Qian, L., Sun, X., Liu, L., et al. (2021). Combined analysis of volatile terpenoid metabolism and transcriptome reveals transcription factors related to terpene synthase in two cultivars of dendrobium officinale flowers. *Front. Genet.* 12. doi: 10.3389/fgene.2021.661296

Li, P., Ruan, Z., Fei, Z., Yan, J., and Tang, G. (2021). Integrated transcriptome and metabolome analysis revealed that flavonoid biosynthesis may dominate the resistance of zanthoxylum bungeanum against stem canker. *J. Agric. Food Chem.* 69, 6360–6378. doi: 10.1021/acs.jafc.1c00357

Liu, Y., Gao, J., Peng, M., Meng, H., Ma, H., Cai, P., et al. (2018). A review on central nervous system effects of gastrodin. *Front. Pharmacol.* 9. doi: 10.3389/fphar.2018.00024

Liu, X., Gong, X., Liu, Y., Liu, J., Zhang, H., Qiao, S., et al. (2022). Application of high-throughput sequencing on the Chinese herbal medicine for the data-mining of the bioactive compounds. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.900035

Liu, S., and Yang, S. (2022). Cardiovascular protective properties of gastrodin. *Asian Pacific J. Trop. Biomed.* 12 (4), 4. doi: 10.4103/2221-1691.340558

Liu, Y., Zhang, Y., Jin, Y., and Chen, Y. (2007). Advance on the chemical and pharmacological studies on plants of pholidota genus. *Lishizhen Med. Mater. Med. Res.* 18, 1631–1633. doi: 1008-0805(2007)-1631-03

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Luo, D., Wang, Z., Li, Z., and Yu, X. (2018). Structure of an entangled heteropolysaccharide from pholidota chinensis lindl and its antioxidant and anti-cancer properties *Int. J. Biol. Macromol.* 112, 921–928. doi: 10.1016/j.ijbiomac.2018.02.051

Medincine, N. J. (2006). *Dictionary of traditional Chinese medicine* (Shanghai: Science Press), 837.

Naikoo, M. I., Dar, M. I., Raghib, F., Jaleel, H., Ahmad, B., Raina, A., et al. (2019). Role and regulation of plants phenolics in abiotic stress tolerance. *Plant Signaling Mol.* 9, 157–168. doi: 10.1016/b978-0-12-816451-8.00009-5

Rai, A., Hirakawa, H., Nakabayashi, R., Kikuchi, S., Hayashi, K., Rai, M., et al. (2021). Chromosome-level genome assembly of ophiorrhiza pumila reveals the evolution of camptothecin biosynthesis. *Nat. Commun.* 12, 405. doi: 10.1038/s41467-020-20508-2

Ramya, M, Jang, S, An, HR, Lee, SY, Park, PM, Park, PH, et al. (2020). Volatile Organic Compounds from Orchids: From Synthesis and Function to Gene Regulation. *Int. J Mol. Sci.* 21 (3)

Ren, M., Xua, W., Zhang, Y., Ni, L., Lin, Y., Zhang, X., et al. (2020). Qualitative and quantitative analysis of phenolic compounds by UPLC-MS/MS and biological activities of pholidota chinensis lindl. *J. Pharm. Biomed. Anal.* 187, 1–11. doi: 10.1016/j.jpba.2020.113350

Rueda, D. C., Schoffmann, A., De Mieri, M., Raith, M., Jahne, E. A., Hering, S., et al. (2014). Identification of dihydrostilbenes in pholidota chinensis as a new scaffold for GABAA receptor modulators. *Bioorg. Med. Chem.* 22, 1276–1284. doi: 10.1016/j.bmc.2014.01.008

Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Shan, T., Yin, M., Wu, J., Yu, H., Liu, M., Xu, R., et al. (2021). Comparative transcriptome analysis of tubers, stems, and flowers of gastrodia elata blume reveals potential genes involved in the biosynthesis of phenolics. *Fitoterapia* 153, 104988. doi: 10.1016/j.fitote.2021.104988

Shimzu, K., Adachi, J., and Muraoka, Y. (2006). ANGLE_ a sequencing errors resistant program for:predicting protein coding regions in unfinished cDNA. *J. Bioinf. Comput. Biol.* 4, 649–664. doi: 10.1142/s0219720006002260

Song, S. S., Tao, Y., Gao, L., Liang, H., Tang, D., Lin, J., et al. (2022). An integrated metabolome and transcriptome analysis reveal the regulation mechanisms of flavonoid biosynthesis in a purple tea plant cultivar. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.880227

Sun, M., Zhao, Y., Shao, X., Ge, J., Tang, X., Zhu, P., et al. (2021). doi: 10.21203/rs.3.rs-415397/v1

Tao, J., Luo, Z. Y., Msangi, C. I., Shu, X. S., Wen, L., Liu, S. P., et al. (2009). Relationships among genetic makeup, active ingredient content, and place of origin of the medicinal plant gastrodia tuber. *Biochem. Genet.* 47, 8–18. doi: 10.1007/s10528-008-9201-7

Ti, H., Zhuang, Z., Li, Y., Wei, G., and Wang, F. (2020). Three new phenanthrenes from pholidota chinensis lindl. and their antibacterial activity. *Natural Product Res.* 11, 1–7. doi: 10.1080/14786419.2020.1845168

Tsai, C. C., Wu, K. M., Chiang, T. Y., Huang, C. Y., Chou, C. H., Li, S. J., et al. (2016). Comparative transcriptome analysis of gastrodia elata (Orchidaceae) in response to fungus symbiosis to identify gastrodin biosynthesis-related genes. *BMC Genomics* 17, 212. doi: 10.1186/s12864-016-2508-6

Wang, S., Bilal, M., Hu, H., Wang, W., and Zhang, X. (2018). 4-hydroxybenzoic acid-a versatile platform intermediate for value-added compounds. *Appl. Microbiol. Biotechnol.* 102, 3561–3571. doi: 10.1007/s00253-018-8815-x

Wang, Y., Gao, Y. J., Li, X., Sun, X. L., Wang, Z. Q., Wang, H. C., et al. (2020). Coniferyl aldehyde inhibits the inflammatory effects of leptomeningeal cells by suppressing the JAK2 signaling. *BioMed. Res. Int.*, 2020, 4616308, 12. doi: 10.1155/2020/4616308

Wang, X. Y., Li, L., Zhu, H., et al (2016). Advance studies on plants of pholidota chinensis. *Asia-Pacfic Tradit. Med.* 12, 42–44. doi: 10.11954/ytcty.201601016

Wang, J., Matsuzaki, K., and KiItanakaba, T. (2006). Stilbene derivatives from pholidota chinensis and their antiinflammatory activity. *Chem. Pharm. Bull.* 54, 1216–1218. doi: 10.1248/cpb.54.1216

Wang, C., Xu, Q., Chen, Z., Hou, J., Li, H., Zhang, X., et al. (2022). Quality evaluation of gastrodia elata in different producing areas. *Cjhinese Tradit. Patent Med.* 44, 487–492. doi: 10.3969/j.issn.1001-1528.2922.02.028

Want, E. J., Wilson, I. D., Gika, H., Theodoridis, G., Plumb, R. S., Shockcor, J., et al. (2010). Global metabolic profiling procedures for urine using UPLC-MS. *Nat. Protoc.* 5, 1005–1018. doi: 10.1038/nprot.2010.50

Weng, S. (2006). Advances in pholidota chinensis. *Modern Chin. Med.* 8, 35–36. doi: 10.13313/j.issn.1673-4890.2006.06.013

Wen, B., Mei, Z., Zeng, C., and Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinf.* 18, 183. doi: 10.1186/s12859-017-1579-y

Yang, F., Li, G., Lin, B., and Zhang, K. (2022). Gastrodin suppresses pyroptosis and exerts neuroprotective effect in traumatic brain injury model by inhibiting NLRP3 inflammasome signaling pathway. *J. Integr. Neurosci.* 21, 72. doi: 10.31083/j.jin2102072

Yang, H., Wu, Y., Gan, C., Yue, T., and Yuan, Y. (2016). Characterization and antioxidant activity of a novel polysaccharide from pholidota chinensis lindl. *Carbohydr Polym.* 138, 327–334. doi: 10.1016/j.carbpol.2015.11.071

Yao, S., Tang, C. P., Ye, Y., Tibor, K., Attila, K. S., Sándor, A., et al. (2008) Stereochemistry of atropisomeric 9,10-dihydrophenanthrene dimers from pholidota chinensis *Tetrahedron: Asymm.* 19 (17) doi: 10.1016/j.tetasy.2008.08.013

Yin, H., Hu, T., Zhuang, Y., and Liu, T. (2020). Metabolic engineering of saccharomyces cerevisiae for high-level production of gastrodin from glucose. *Microb. Cell Fact.* 19 (1). doi: 10.1186/s12934-020-01476-0

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., et al. (2018). The gastrodia elata genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9, 1615. doi: 10.1038/s41467-018-03423-5

Zeki, O. C., Eylem, C. C., Recber, T., Kir, S., and Nemutlu, E. (2020). Integration of GC-MS and LC-MS for untargeted metabolomics profiling. *J. Pharm. BioMed. Anal.* 190, 113509. doi: 10.1016/j.jpba.2020.113509

Zhang, C., Chen, J., Huang, W., Song, X., and Niu, J. (2021). Transcriptomics and metabolomics reveal purine and phenylpropanoid metabolism response to drought stress in dendrobium sinense, an endemic orchid species in hainan island. *Front. Genet.* 12. doi: 10.3389/fgene.2021.692702

Zhang, M., Chen, L., Zhu, H., Li, L., Da, F., Long, L., et al. (2019). Establishment of HPLC fingerprint of pholidota chinensis and its cluster analysis. *China Pharm.* 30, 1792–1795. doi: 10.6039/j.issn.1001-0408.2019.13.13

Zhang, Z. L., Gao, Y. G., Zang, P., Gu, P. P., Zhao, Y., He, Z. M., et al. (2020). Research progress on mechanism of gastrodin and p-hydroxybenzyl alcohol on central nervous system. *China J. Chin. Mater. Med.* 45, 312–320. doi: 10.19540/j.cnki.cjcmm.20190730.401

Zhang, D., Song, Y. H., Dai, R., Lee, T. G., and Kim, J. (2020). Aldoxime metabolism is linked to phenylpropanoid production in camelina sativa. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00017

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zou, Z., Huang, P., Zeng, H., Xu, Q., Li, J., and Liu, Y. (2017). The research progress of Chinese pholidota pseudobulb or herb's biological activity function. *Guangdong Chem. Industry* 44, 61–62. doi: 1007-1865 (2017)02-0061-02

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 December 2022
DOI 10.3389/fpls.2022.1086335

OPEN ACCESS

EDITED BY
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Zhiqiang Wu,
Agricultural Genomics Institute at Shenzhen (CAAs), China
Dahui Liu,
Hubei University of Chinese Medicine, China

*CORRESPONDENCE
Xiaohui Yan
yanxh@tjutcm.edu.cn
Chunhua Wang
pharmwch@126.com
Xueshuang Huang
xueshuanghuang@126.com

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 01 November 2022
ACCEPTED 01 December 2022
PUBLISHED 20 December 2022

CITATION
Yang Y, Sun Y, Wang Z, Yin M, Sun R, Xue L, Huang X, Wang C and Yan X (2022) Full-length transcriptome and metabolite analysis reveal reticuline epimerase-independent pathways for benzylisoquinoline alkaloids biosynthesis in *Sinomenium acutum*. *Front. Plant Sci.* 13:1086335. doi: 10.3389/fpls.2022.1086335

Full-length transcriptome and metabolite analysis reveal reticuline epimerase-independent pathways for benzylisoquinoline alkaloids biosynthesis in *Sinomenium acutum*

Yufan Yang[1,2], Ying Sun[1,3], Zhaoxin Wang[1,2], Maojing Yin[4,5], Runze Sun[6], Lu Xue[1,2], Xueshuang Huang[6]*, Chunhua Wang[5]* and Xiaohui Yan[1,2]*

[1]State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, China, [2]Haihe Laboratory of Modern Chinese Medicine, Tianjin, China, [3]WuXi AppTec (Tianjin) Co., Ltd., Tianjin, China, [4]College of Pharmaceutical Engineering of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, China, [5]School of Medicine, Foshan University, Foshan, Guangdong, China, [6]Hunan Provincial Key Laboratory for Synthetic Biology of Traditional Chinese Medicine, Hunan University of Medicine, Huaihua, Hunan, China

Benzylisoquinoline alkaloids (BIAs) are a large family of plant natural products with important pharmaceutical applications. *Sinomenium acutum* is a medicinal plant from the Menispermaceae family and has been used to treat rheumatoid arthritis for hundreds of years. *Sinomenium acutum* contains more than 50 BIAs, and sinomenine is a representative BIA from this plant. Sinomenine was found to have preventive and curative effects on opioid dependence. Despite the broad applications of *S. acutum*, investigation on the biosynthetic pathways of BIAs from *S. acutum* is limited. In this study, we comprehensively analyzed the transcriptome data and BIAs in the root, stem, leaf, and seed of *S. acutum*. Metabolic analysis showed a noticeable difference in BIA contents in different tissues. Based on the study of the full-length transcriptome, differentially expressed genes, and weighted gene co-expression network, we proposed the biosynthetic pathways for a few BIAs from *S. acutum*, such as sinomenine, magnoflorine, and tetrahydropalmatine, and screened candidate genes involved in these biosynthesis processes. Notably, the reticuline epimerase (REPI/STORR), which converts (*S*)-reticuline to (*R*)-reticuline and plays an essential role in morphine and codeine biosynthesis, was not found in the transcriptome data of *S. acutum*. Our results shed light on the biogenesis of the BIAs in *S. acutum* and may pave the way for the future development of this important medicinal plant.

KEYWORDS

transcriptome, *Sinomenium acutum*, benzylisoquinoline alkaloid, biosynthetic pathway, weighted gene co-expression network analysis

1 Introduction

Benzylisoquinoline alkaloids (BIAs) are a large and diverse group of specialized plant secondary metabolites with important medicinal values. BIAs are widely distributed in the Papaveraceae, Ranunculaceae, Berberidaceae, and Menispermaceae families (Narcross et al., 2016). Pharmacological studies have revealed that BIAs have diverse biological activities, such as anticancer, analgesia, antitumor, and anti-inflammatory (Dang et al., 2012). *Papaver somniferum* is a medicinal plant that produces several pharmaceutically important BIAs, including morphine, codeine, papaverine, and noscapine. It is used as a model organism to investigate the biosynthesis of BIAs (Beaudoin and Facchini, 2014). The recently reported genomes of *Papaver somniferum* (Guo et al., 2018) and *Macleaya cordata* (Liu et al., 2017) provide deep insight into the BIA biosynthetic gene clusters in these medicinal plants and allow the functional characterization of key genes involved in BIA metabolism. Morphine is commonly used as an analgesic drug for pain relief. However, long-term use of morphine can lead to severe health problems, such as drug addiction and abuse (Listos et al., 2019).

Previous studies have shown that various BIAs share similar steps in the early stage of their biosynthesis. The biosynthesis of BIAs begins with the conversion of *L*-tyrosine to two precursors, dopamine and 4-hydroxyphenylacetaldehyde (Sato and Kumagai, 2013). These two compounds are converted by a series of enzymes, such as tyrosine aminotransferase (TyrAT), tyrosine decarboxylase (TYDC), tyrosine/tyramine 3-hydroxylase (3OHase), norcoclaurine synthase (NCS), coclaurine N-methyltransferase (CNMT), and 3′-hydroxyl-*N*-methylcoclaurine 4′-*O*-methyltransferase (4′OMT), to afford (*S*)-reticuline which is the divergent point for the biosynthesis of many BIAs (Hagel and Facchini, 2013). Modification of (*S*)-reticuline by various *O*-methyltransferase (OMTs) and cytochrome P450 oxygenases (CYPs) can lead to the formation of more than 2,500 BIAs. Reticuline epimerase (REPI) is responsible for the conversion of (*S*)-reticuline to (*R*)-reticuline, which is essential for the formation of morphine and codeine (Catania et al., 2022). Apart from simple methylation and oxidation, biosynthesis of some BIAs also undergoes complex reactions, including methylenedioxy bridge formation and phenol coupling. For example, CYP80G2 from *Coptis japonica* can catalyze the intramolecular C–C phenol coupling reaction of (*S*)-reticuline to generate (*S*)-corytuberine. CYP719B1 from *P. somniferum* can catalyze the C–C phenol-coupling reaction in morphine biosynthesis (Gesell et al., 2009).

Sinomenium acutum is a traditional Chinese herb of the Menispermaceae family. The root and stem of *S. acutum* have long been used to treat rheumatoid arthritis in south China (Liu et al., 2018). Compounds from *S. acutum* have shown anti-inflammatory, analgesic, sedative, and immunosuppressive effects (Liu et al., 2019; Xu et al., 2021a). Sinomenine is the representative compound of *S. acutum*, structurally similar to morphine (Jiang et al., 2020). Sinomenine has preventive and therapeutic effects on opioid dependence without detectable drug addiction (Jin et al., 2008; Bhambhani et al., 2021). The studies on *S. acutum* mainly focused on the chemical constituents and the biological activities of the isolated BIAs (Jiang et al., 2019; Liu et al., 2021), while the biosynthesis of BIAs in *S. acutum* remains untouched, mainly due to the lack of genomic or transcriptomic data.

With the development of high-throughput sequencing technology, transcriptome sequencing has become a powerful tool for studying the regulation of gene expression and the biosynthetic pathways for secondary metabolites in plants (Wang et al., 2019; Zhang et al., 2020a). Combining transcriptomics data with metabolomic analysis is an effective way to identify genes involved in secondary metabolite biosynthesis. This approach has been successfully used in several plant species, such as *C. deltoidei* (Zhong et al., 2020), *Stephania tetrandra* (Zhang et al., 2020b), *Polygonatum cyrtonema* Hua (Wang et al., 2019), and *Corydalis yanhusuo* (Xu et al., 2021b).

In this study, we combined the next-generation sequencing (NGS) and single-molecule real-time (SMRT) sequencing techniques to obtain the *de novo* transcriptome assembly of *S. acutum*, aiming to identify the candidate genes involved in the biosynthesis of various BIAs in this plant. The contents of alkaloids in the root, stem, leaf, and seed were determined to help the screening of candidate genes involved in BIA biosynthesis. The transcripts were annotated using various public databases and evaluated by the weighted gene co-expression network analysis (WGCNA) and differentially expressed gene (DEG) analysis to screen candidate genes involved in BIA biosynthesis in *S. acutum*.

2 Materials and methods

2.1 Plant materials

Sinomenium acutum was collected in September 2019 from Huangyan District, Huaihua, Hunan Province, China (27°27'N, 110°40'E) and identified by Dr. Wang Chunhua. Three biological replicates of root (YXH-ZQ-G), stem (YXH-ZQ-J), leaf (YXH-ZQ-Y), and seed (YXH-ZQ-Z) were collected randomly from fresh plants (Figures 1A–D). After the collection, the fresh samples were frozen immediately in liquid nitrogen, stored at -80°C, and subjected to sequencing and metabolomic analysis.

2.2 Identification and quantification of BIAs

The BIAs in *S. acutum* were identified and quantified using the Waters ACQUITY Ultra-Performance Liquid Chromatography

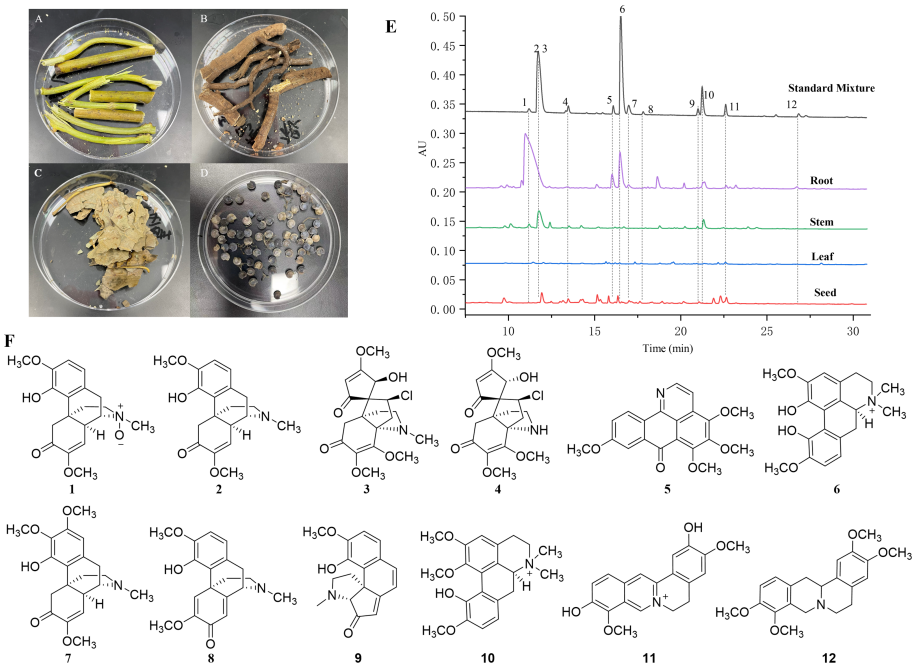


FIGURE 1
Photos and HPLC profiles of fresh tissues of *S. acutum*. **(A–D)**: the photos of fresh root **(A)**, stem **(B)**, leaf **(C)**, and seed **(D)**. **(E)**: HPLC profiles of the four tissues and the mixed standards of 12 BIAs from *S. acutum*. **(F)**: Chemical structures of 12 selected BIAs, including sinomenine *N*-oxide (1), sinomenine (2), acutumine (3), dauricumidine (4), bianfugenine (5), magnoflorine (6), 8-demethoxyrunanine (7), sinoacutine (8), sinoracutine (9), menisperine (10), stepharanine (11), and tetrahydropalmatine (12).

(UPLC) System (Waters, Milford, MA, United States) and the Q Exactive Plus Orbitrap Mass Spectrometer (Thermo Fisher Scientific., Waltham, MA, United States). Dried powder samples (0.25 g) of the root, leaf, stem, and seed of *S. acutum* were accurately weighted and subsequently extracted with 20 mL of methanol solution for 30 min in an ultrasonic bath. Chromatographic separation was carried out on the Waters ACQUITY UPLC system equipped with a BEH C18 column (2.1 × 100 mm, 1.7 μm), and the column temperature was kept at 30°C. The mobile phases were water with 0.1% formic acid (A) and acetonitrile (B) at a flow rate of 0.3 mL/min. The solvent gradients for B were 0–6 min, 5%; 6–20 min, 5–16%; 20–22 min, 16–95%; 22 –25 min, 95%; and 25–26 min, 5%. The Q Exactive MS system used the electrospray ionization source (ESI) in the positive ion mode. The spray voltage and capillary temperature were 3.0 kV and 320 °C, respectively. The flow rates of the atomization gas and heating auxiliary gas were set at 35 arb and eight arb, respectively. The auxiliary gas heating temperature was set at 350°C.

2.3 RNA extraction, library preparation, and Illumina Hiseq sequencing

Total RNA from the root, leaf, stem, and seed of *S. acutum* was isolated from the tissue using the TRIzol® Reagent according to the manufacturer's instructions (Invitrogen, Carlsbad, CA, United States), and genomic DNA was removed using DNase I (Takara, Akasaka, Tokyo, Japan). Then the RNA quality was determined using 2100 Bioanalyser (Agilent, Palo Alto, CA, United States) and quantified using the ND-2000 (NanoDrop Technologies, Wilmington, DE, United States). High-quality RNA samples with $OD_{260/280} \geq 1.8$ and the RNA Integrity Number ≥ 6.5 were used to construct the sequencing library and to calculate the Q20, Q30, and GC percentages for each sample. RNA-seq transcriptome libraries were prepared following the TruSeq™ RNA sample preparation Kit from Illumina (San Diego, CA, United States), using 1 μg of total RNA. Shortly, messenger RNA was isolated with polyA selection by oligo(dT) beads and fragmented using a fragmentation buffer. cDNA synthesis, end repair, A-base addition, and ligation of the Illumina-indexed adaptors were performed according to Illumina's protocol. Libraries were then size-selected for cDNA target fragments of 200–300 bp on 2% Low Range Ultra Agarose followed by PCR amplified using Phusion DNA polymerase (New England Biolabs, Ipswich, MA, United States) for 15 PCR cycles. After quantified by TBS380, the paired-end libraries (150 bp*2) were sequenced using the Illumina NovaSeq 6000 system (Biozeron, Shanghai, China).

2.4 *De novo* sequence assembly and gene annotation

The raw paired-end reads were trimmed and quality controlled by Trimmomatic (Bolger et al., 2014). RNA was then assembled *de novo* using Trinity on the clean data from all samples (Grabherr et al., 2011). All assembled transcripts were searched with BlastX against the NCBI protein non-redundant (NR), String, and KEGG databases to identify proteins with the highest sequence similarity to a given transcript to retrieve its functional annotation. A typical *E* value cutoff of less than 1.0×10^{-5} was used to define a hit. The Blast2GO program (Conesa et al., 2005) was used to obtain gene ontology (GO) annotations of uniquely assembled transcripts to describe biological processes, molecular functions, and cellular components. Metabolic pathway analysis was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG).

2.5 Analysis of differentially expressed genes

To identify differentially expressed genes (DEGs) between two different samples, the expression level of each transcript was calculated according to the reads per kilobase of exon per million mapped reads (RPKM) method. The RSEM software was used to quantify gene and isoform abundances (Li and Dewey, 2011). The *R* statistical package software was utilized for differential gene expression analysis. In addition, functional-enrichment analysis, including GO and KEGG, was performed to identify which DEGs were significantly enriched in GO terms and metabolic pathways at Bonferroni-corrected *P*-value ≤ 0.05 compared with the whole-transcriptome background. GO functional enrichment and KEGG pathway analysis were carried out by Goatools (Klopfenstein et al., 2018) and KOBAS 2.1.1 (Bu et al., 2021).

2.6 Phylogenetic analysis

Phylogenetic analysis of the CYPs and OMTs (4′-OMT, 6-OMT, CoOMT, and SOMT1) involved in BIA biosynthesis in *S. acutum* and other plants were performed based on the deduced amino acid sequences. The alignment was implemented using the MUSCLE algorithm with default parameters. MEGA6 was used to construct the neighbor-joining trees using a bootstrap method with a Poisson model and pairwise deletion (1000 replications).

2.7 Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) was performed to identify the key genes involved in BIA biosynthesis in *S. acutum*. WGCNA constructs networks using the absolute value of Pearson's correlation coefficient as the measure of gene co-expression, which is raised to a power to create the adjacency matrix. The topological overlap distance calculated from the adjacency matrix is clustered with the average linkage hierarchical clustering. Our modules were defined using the cutreeDynamic function with a minimum module size of 30 genes. A module eigengene distance threshold of 0.25 was used to merge highly similar modules using the mergeCloseModules function. We empirically set the cutoff of the weight value as 0.10 to determine biologically significant edges for each module. Correlation analysis between module eigengenes and measured agronomic traits was applied to explore the biological significance of each module.

2.8 Quantitative real-time PCR analysis

cDNA was synthesized from 1 μg of total RNA using the 5x PrimerScript RT reagent Kit (Takara Bio Inc). In brief, 10 μL of genomic DNA-removed template, 4 μL of 5x PrimerScript buffer, 1 μL of RT enzyme mix, 1 μL of RT primer, and 4 μL of nuclease-free water were added to result in a 20 μL reverse transcription solution. The mixture was incubated for 30 min at 37°C, followed by inactivated for 5 s at 85°C to afford the cDNA for quantitative real-time PCR (qRT-PCR) experiments. The qRT-PCR was performed in triplicate on 8-strip PCR tubes (Axygen) on a SLAN-96P Real Time PCR System (StrongMed Corporation) using the SYBR Green Mix (Life Technologies). Each PCR solution consisted of 10 μL of SYBR Green Mix, 2 μL of primer mix (10 mM), 5 μL of template cDNA, and 3 μL of nuclease-free water, to obtain a 20 μL system for each PCR reaction. The PCR conditions were: 15 min at 95°C, followed by 40 cycles of 30 s at 95°C, 20 s at 55–65°C, 20 s at 72°C. Table S2 provides the detailed information about the primer sequences and temperatures. The β-actin gene was used as an internal control for the qRT-PCR analysis.

3 Results

3.1 Quantification analysis of BIAs in different parts of *S. acutum*

It has been reported that *S. acutum* contains more than one hundred alkaloids, including morphinans, aporphines, protoberberines, benzylisoquinolines, and other compounds. We collected fresh samples of the root, stem, leaf, and seed of *S. acutum* (Figures 1A–D). In order to correlate the distribution of alkaloids with the expression of different biosynthetic genes, we first determined the concentration of representative alkaloids in various tissues of *S. acutum* by Q Exactive high-resolution MS and UPLC analysis, respectively (Figure 1E). The contents of

sinomenine *N*-oxide (**1**, 0.96 ± 0.08 mg/g, DW), sinomenine (2, 22.11 ± 0.09 mg/g, DW), acutumine (3, 0.94 ± 0.02 mg/g, DW), magnoflorine (6, 5.04 ± 0.08 mg/g, DW), 8-demethoxyrunanine (7, 1.26 ± 0.53 mg/g, DW), and tetrahydropalmatine (12, 4.00 ± 0.53 mg/g, DW) were much higher in the root than the other three parts (Figure 1F and Table 1). Besides the root, sinomenine is also accumulated in the stem. The stem contains most menisperine (10), whereas the concentrations of dauricumidine (4) and sinoracutine (9) are the highest in the seed. Compared to the root, stem, and seed, the leaf contains the least BIAs. We could only detect a trace amount of acutumine (3, 0.07 ± 0.01 mg/g, DW), sinoracutine (9, 0.08 ± 0.01 mg/g, DW), and stepharanine (11, 0.06 ± 0.01 mg/g, DW) in the leaves of *S. acutum*. These results confirmed that the biosynthesis and accumulation of different alkaloids vary significantly in different tissues of *S. acutum*, thus enabling the prediction of genes involved in alkaloid biosynthesis in this plant.

3.2 Transcriptome analysis of *S. acutum* samples

After the quantitative analysis, we obtained the transcriptome data of *S. acutum* by performing the single-molecule real-time (SMRT) sequencing on the PacBio Sequel platform and the second-generation sequencing (SGS) on the Illumina platform. Twelve high-quality RNA samples from the root, stem, leaf, and seed of *S. acutum* were sequenced using the Illumina Hiseq system. The SGS sequencing produced average raw reads ranging from 42,406,048 to 46,677,278 and clean reads from 39,799,353 and 43,421,740, respectively. After filtering out the adapters and low-quality reads, the PacBio Sequel platform generated 25,660,172 subreads (34.72 Gb subreads base). The Q20 and Q30 statistics of the clean reads in all samples were higher than 98.2% and 94.2%, respectively, indicating that the sequencing data were of high quality and met the analysis requirements. The average length of the transcripts was 1,353 bp, and the N50 length of all subreads was 1,514 bp. A total of 458,090 circular consensus sequences (CCSs) were detected, with an average read length of 1,645 bp and an N50 of 1,856 bp. We detected 280,999 full-length nonchimeric (FLNC) reads with a mean length of 1,620 bp. Correction of data from the PacBio Sequel platform using the Illumina platform generated 110,876 polished consensus sequences, with an average length of 1,453 bp, an N90 of 861 bp, and an N50 of 1,658 bp, respectively. To analyze the functions of the unigenes from the *S. acutum* transcripts, the redundant sequences were removed *via* the CD-Hit program, and the consensus transcripts were finally clustered into 60,675 non-redundant unigenes, with a mean length of 1,614 bp and an N50 length of 1,840 bp (Table 2). Most unigenes are distributed between 800 to 2,800 nt, indicating a high-quality transcriptome of *S. acutum* for future studies (Figure 2A).

3.3 Functional annotation of the *S. acutum* transcriptome data

The unigenes from the *S. acutum* transcriptome were annotated with the commonly used databases, including the NCBI non-redundant (NR), Swiss-Prot, Clusters of Orthologous

TABLE 1 The BIA contents in different tissues of *S. acutum*.

Alkaloids	Contents (mg/g)			
	Root	Stem	leave	Seed
sinomenine *N*-oxide (**1**)	0.96 ± 0.08	ND	ND	ND
sinomenine (**2**)	22.11 ± 0.25	2.50 ± 0.03	ND	0.09 ± 0.01
acutumine (**3**)	0.94 ± 0.02	0.63 ± 0.05	0.07 ± 0.01	0.64 ± 0.06
dauricumidine (**4**)	0.25 ± 0.01	0.21 ± 0.01	ND	0.48 ± 0.03
bianfugenine (**5**)	0.04 ± 0.01	ND	ND	ND
magnoflorine (**6**)	5.04 ± 0.08	ND	ND	ND
8-demethoxyrunanine (**7**)	1.26 ± 0.11	ND	ND	ND
sinoacutine (**8**)	0.14 ± 0.01	0.03 ± 0.01	ND	ND
sinoracutine (**9**)	0.11 ± 0.01	0.16 ± 0.01	0.08 ± 0.01	0.29 ± 0.01
menisperine (**10**)	0.90 ± 0.01	1.37 ± 0.03	ND	ND
stepharanine (**11**)	0.10 ± 0.01	ND	0.06 ± 0.01	0.19 ± 0.01
tetrahydropalmatine (**12**)	4.00 ± 0.53	0.26 ± 0.01	ND	ND

ND, not detected.

Groups of proteins (COGs), Kyoto Encyclopedia of Genes and Genomes (KEGG), and GO (Gene Ontology), using BLASTX (*e* value < 10^{-5}). Protein function was predicted according to the annotations of the most similar proteins in these databases. Out of the 60,675 unigenes, 19,675 (32.4%) were annotated by all databases, and 52,624 (86.7%) were annotated by at least one of these databases. The unigenes annotated in the NR, GO, COG, KEGG, and SWSS databases were 52,530 (86.6%), 49,844 (82.2%), 38,814 (63.9%), 22,715 (37.4%), and 40,351 (66.5%), respectively (Table 2 and Figure 2B). KEGG analysis revealed that 5,897 unigenes were predicted to participate in the metabolic pathways, and 3,132 unigenes were involved in the biosynthesis of secondary metabolites (Figure S1). KEGG metabolic pathway analysis showed that the unigenes from *S. acutum* could be assigned to 322 pathways (Supplementary Table 1). Based on the results of the aligned transcripts in the NR database, 25.9%, 21.2%, and 12.3% of the annotated unigenes had high sequence similarities to the unigenes from *Macleaya cordata*, *Nelumbo nucifera*, and *Aquilegia coerulea*, respectively (Figure 2C). It has been reported that these plants all contain a series of alkaloids, such as sanguinarine, chelerythrine, nuciferine, magnoflorine, and aporphine alkaloids (Hagel and Facchini, 2013). The 49,844 unigenes annotated with the GO database were divided into three categories: biological process, molecular function, and cellular component. The largest GO groups in the "biological process" ontology were the "cellular process" and "metabolic process," with 32,556 and 25,989 annotated unigenes, respectively (Figure S2).

3.4 Analysis of differentially expressed genes among different tissues of *S. acutum*

To investigate the mechanism behind the tissue-specific distribution of various alkaloids, we performed the differentially expressed genes (DEGs) analysis among the root, stem, leaf, and seed in *S. acutum*. The transcript abundance of all the unigenes was evaluated by converting read counts to reads per kilobase of exon model per million mapped reads (RPKM). RNA-seq analysis showed that 35,959, 36,927, 36,086, and 36,975 unigenes in the full-length transcriptome were expressed with RPKM > 1 in the

TABLE 2 Summary of the assembly and annotation of *S. acutum* transcriptome.

	Root	Stem	Leaf	Seed
Illumina sequencing				
Total reads	46,105,271	46,677,278	44,909,644	42,406,048
Nucleotides (nt)	6,915,790,700	7,001,591,800	6,736,446,600	6,360,907,200
Clean reads	42,879,224	43,421,740	42,014,966	39,799,353
Clean nucleotides (nt)	6,404,310,071	6,484,998,541	6,275,854,255	5,948,595,887
PacBio sequencing				
Subreads number	25,660,172			
Average subreads length (nt)	1,353			
N50 length (nt)	1,514			
CCS reads number	458,090			
Average CCS length	1,645			
FLNC reads	280,999			
Unigene	60,675			
Total length (nt)	97,935,561			
N50 (nt)	1,840			
Nr	52,530 (86.6%)			
GO	48,944 (82.2%)			
COG	38,814 (63.9%)			
KEGG	22,715 (37.4%)			
SWSS	40,351 (66.5%)			
All annotated unigenes	52,624 (86.7)			

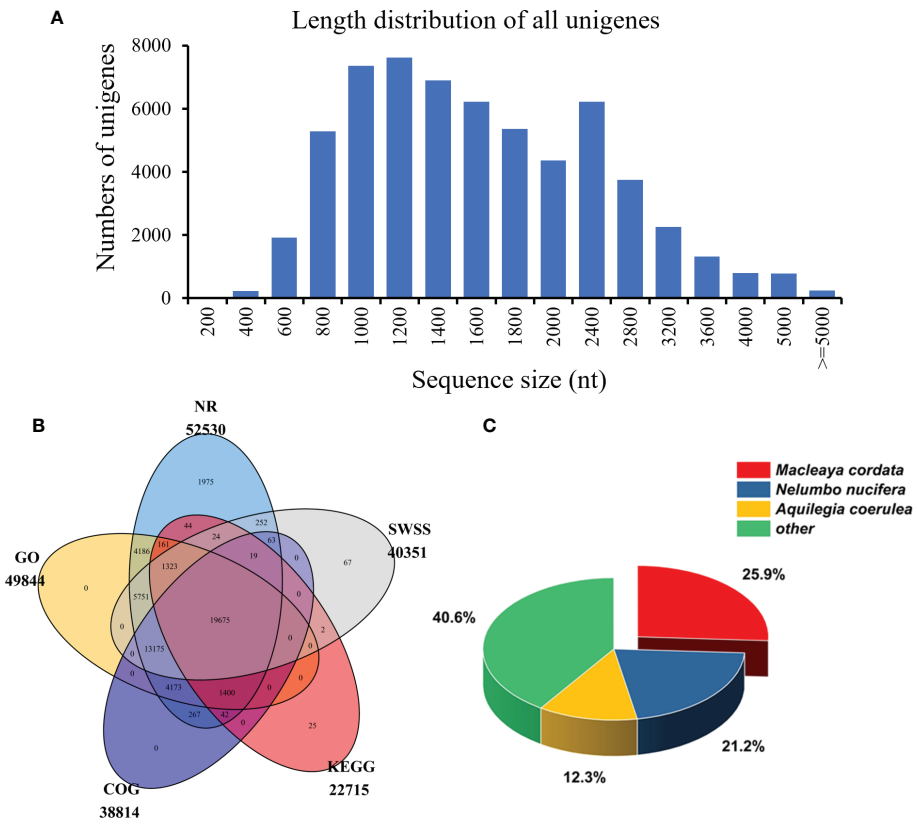


FIGURE 2
The length distribution and functional annotation of the unigenes in *S. acutum*. **(A)** Length distribution of the 60,675 unigenes. **(B)** Annotated unigenes from different public databases. **(C)** NR homologous species distribution analysis.

root, stem, leaf, and seed, respectively (Table S3 and Figure S3). A total of 21,594 DEGs were screened under the condition of false discovery rate (FDR) metric with adjusted *p-value* <= 0.05 and | log2(FoldChange) | >= 1. As the quantitative studies have shown that the root of *S. acutum* contains the most alkaloids and the leaf comprises the least alkaloids, we analyzed the DEGs in the root versus the stem and leaf. There are 8,720 DEGs in the root vs. leaf group and 10,026 DEGs in the root vs. seed group. Most of the DEGs are down-regulated in the root. In contrast, only 2,848 DEGs were screened in the root vs. stem group, indicating that the gene expression patterns in the root and stem are much more similar than the other two tissues (Figure S4). Of the 8,720 DEGs in the root vs. leaf group, 5,099 genes were down-regulated, and 3,621 genes were up-regulated in the root. Moreover, 330 DEGs were found in the three comparison groups (root vs. stem, root vs. leaf, and stem vs. leaf) (Figure 3A). The identified DEGs from different tissues were further analyzed by KEGG enrichment analysis. The significantly enriched pathways are protein processing in the endoplasmic reticulum, carbon metabolism, biosynthesis of amino acids, and plant hormone signal transduction (Figures 3B–D). Based on KEGG analysis, 22,715 unigenes were annotated to 6 main categories and 322 biological pathways (Supplementary Table 1). In addition, a total of 672 unigenes were assigned to the subcategory "Other secondary metabolite biosynthesis," of which 247 genes were assigned to "Phenylpropane biosynthesis" (ko00940) and 47 genes to "Isoquinoline alkaloid biosynthesis" (ko00950) (Table S1). According to the distribution characteristics of the target alkaloids and the gene expression patterns, the candidate genes involved in alkaloid biosynthesis could be retrieved. These results also align with the previous results that the root and the stem had similar alkaloid contents, and the seed and leaf contain much fewer alkaloids than the root and stem.

3.5 Weighted gene co-expression network analysis of the DEGs in *S. acutum*

In order to explore key genes and co-expression networks that play essential roles in the biosynthesis of alkaloids, we analyzed 15,580 DEGs from 12 samples by weighted gene co-expression network analysis (WGCNA). The optimal soft

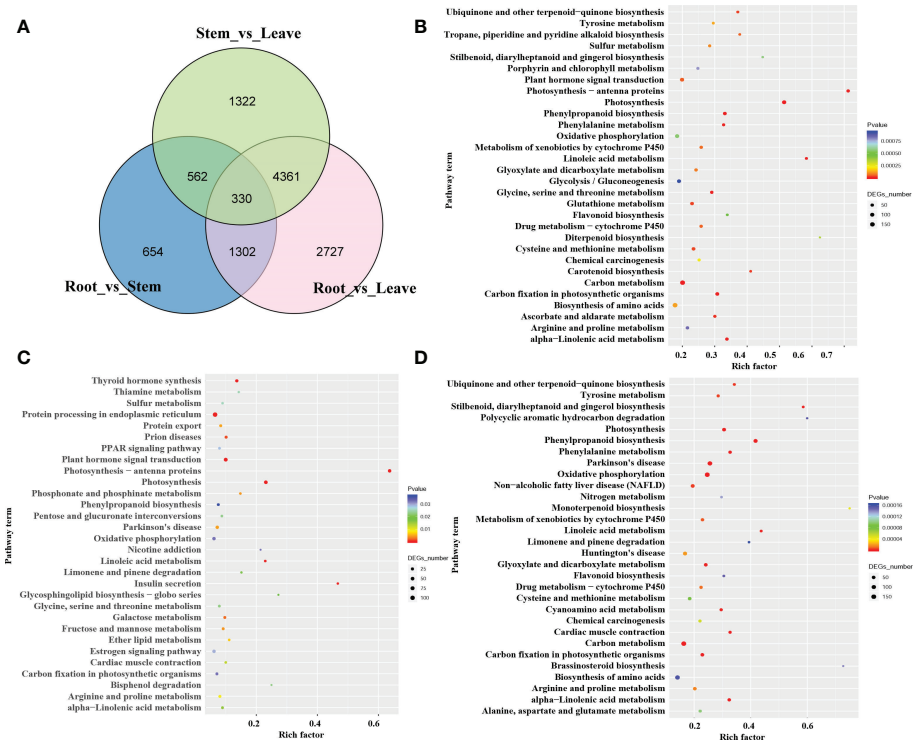


FIGURE 3
The number and pathway terms of differentially expressed genes among the root, stem, and leaf. **(A)** Venn diagram of DEGs in the three tissues. **(B)**, **(C)**: pathway enrichment of genes predominantly expressed in the root **(B)**, stem **(C)**, and leaf **(D)**.

threshold was set at 12 to construct a scale-free network. The adjacency and tom overlap matrix were established using the function adjacency and tom similarity, respectively [33]. The modules were divided according to the dynamic cutting tree, and the modules with high similarities were merged, leading to the formation of seven modules (Figures 4A, B). Seven characteristic BIAs of *S. acutum*, including sinomenine, menisperine, tetrahydropalmatine, sinomenine *N*-oxide, sinoacutine, magnoflorine, and acutumine, were used as traits to correlate with the seven gene modules generated by WGCNA analysis. A labeled heatmap function was used to visualize and analyze the relationship between the modules and BIAs. As shown in Figure 4C, the pink, purple, and dark red modules were positively correlated with the seven BIAs. The pink, purple, and dark red modules contain 2314, 867, and 307 genes, respectively. The pink, purple, and dark red modules had correlation coefficients of 0.98, 0.63, and 0.53 to sinomenine, and the correlation coefficients of the pink, purple, and dark red modules to magnoflorine were 0.96, 0.67, and 0.56, respectively (Figure 4C). The eigenvector correlation analysis between each module was performed for all seven modules. It can be seen that the pink and dark red modules and the pink and purple modules have significant correlations (Figure 4D). We proposed that the genes in these three modules were more likely to participate in the biosynthesis of BIAs in *S. acutum*.

3.6 Identification of candidate genes involved in BIA biosynthesis in *S. acutum*

In order to reveal the BIA biosynthetic pathways in *S. acutum*, we divided the pathways into two parts: 1) the common pathway that converts tyramine and 4-hydroxyphenylpyruvate to (*S*)-reticuline and 2) the downstream pathways that drive (*S*)-reticuline to the various BIAs. Based on the results from KEGG pathway analysis and WGCNA screening of the transcriptome data, as well as BLASTP analysis of the unigenes using the benzylisoquinoline biosynthetic enzymes from *P. somniferum* as queries, 24 unigenes encoding enzymes of the common pathway in *S. acutum* were obtained (Figure 5). For the classification of cytochrome P450 enzymes, proteins with greater than 40% sequence similarity are grouped into the same family. We also used this criteria to mine the protein homologs for BIA biosynthesis in *S. acutum*. All of the encoded enzymes showed high sequence similarities (similarities > 40%) to the query enzymes from *P. somniferum*. Five candidate genes were found

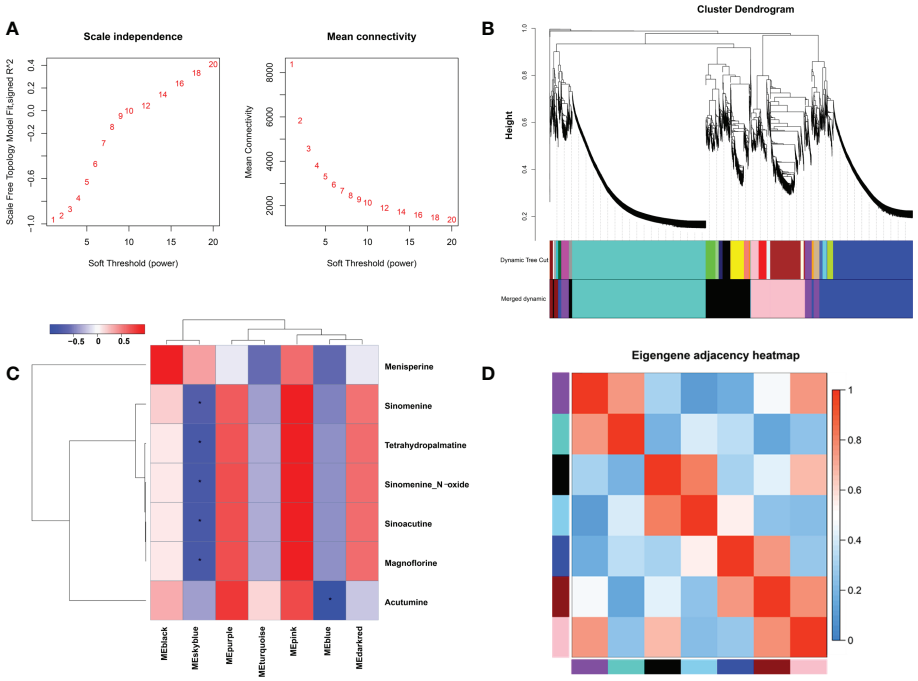


FIGURE 4
Weighted gene co-expression network analysis (WGCNA) of DEGs identified in the root, stem, leaf, and seed of *S. acutum*. **(A)** Determination of soft-thresholding power in WGCNA. **(B)** Hierarchical cluster tree showing seven modules of co-expressed genes. **(C)** Correlations of the modules and the seven BIAs. **(D)** Heatmap of the correlation between different modules in the weighted gene co-expression network. The symbol * denotes correlation coefficient smaller than -0.5.

to encode proteins of the TyrAT family and the (*S*)-*N*-methylcoclaurine 3′-hydroxylase (NMCH) family, respectively. In contrast, only one candidate gene (gene4475) was found for the 6OMT.

The cytochrome P450 (CYP450) oxygenases and O-methyltransferases were screened for the downstream pathways using Pfam annotation. A total of 262 CYP450-encoding unigenes were found in the transcriptome data of *S. acutum*. Three CYP families, i.e., CYP80, CYP82, and CYP719, have been reported to participate in BIA biosynthesis in plants (Dastmalchi et al., 2018). We searched for CYP450 proteins in these three families from *S. acutum* and found 19 members. The 19 screened enzymes were subjected to phylogenetic analysis, together with their characterized homologs from *P. somniferum*, *C. japonica*, and *Berberis stolonifera* (Figure 6A). Seven unigenes encode proteins of the CYP80B subfamily, which catalyzes the hydroxylation of the 3′-position of (S)-*N*-methylcoclaurine. Two genes (gene40122 and gene27149) encode proteins of the CYP80G2 subfamily. Six genes (gene1888, gene24702, gene2351, gene2423, gene1693, and gene25895) encode the CYP82Y/N subfamily enzymes that are involved in the formation of 1,2-dehydroreticuline, protopine, dihydrosanguinarine. Gene46351 and gene30689 encode proteins of the CYP719A/B subfamilies. Salutaridine synthase (SalSyn) is a representative enzyme of the CYP719B1 family. It is responsible for the bridge-forming C–C phenol coupling of (*R*)-reticuline to yield the promorphinan alkaloid salutaridine. As sinomenine is also a member of the morphinan alkaloid with a different bridge from morphine, thebaine, and codeine, one of the enzymes encoded by gene46351 and gene30689 is proposed to catalyze the C–C phenol coupling of (*S*)-reticuline, instead of (*R*)-reticuline, to form sinomenine and its derivatives. The expression patterns of these genes were analyzed. It could be seen that genes encoding proteins of the CYP80B family were actively expressed in the root and leaf, while their expression levels were low in the stem (Figure 6B). The CYP719A/B family of proteins were expressed at a higher level in the root and stem than in the leaf and stem.

Plenty of *O*-methyltransferases (OMTs) and *N*-methyltransferases (NMTs) are involved in modifying the alkaloid scaffolds during the biosynthetic process of BIAs. We performed phylogenetic analysis and conserved motif structure prediction of the methyltransferases from *S. acutum* to mine enzymes catalyzing the methylation reactions in BIA biosynthesis (Figure 6C). The characterized methyltransferases from the BIA biosynthetic pathways were added to the phylogenetic tree for comparison. Three genes (gene41327, gene42607, and gene38655) probably encode 4′-OMTs that convert (*S*)-3-hydroxy-*N*-methylcoclaurine to (*S*)-reticuline. Gene41327 and gene38655 each encode a protein with a dimerization domain in the N-terminal. In contrast, no dimerization domain was detected in the enzyme encoded by

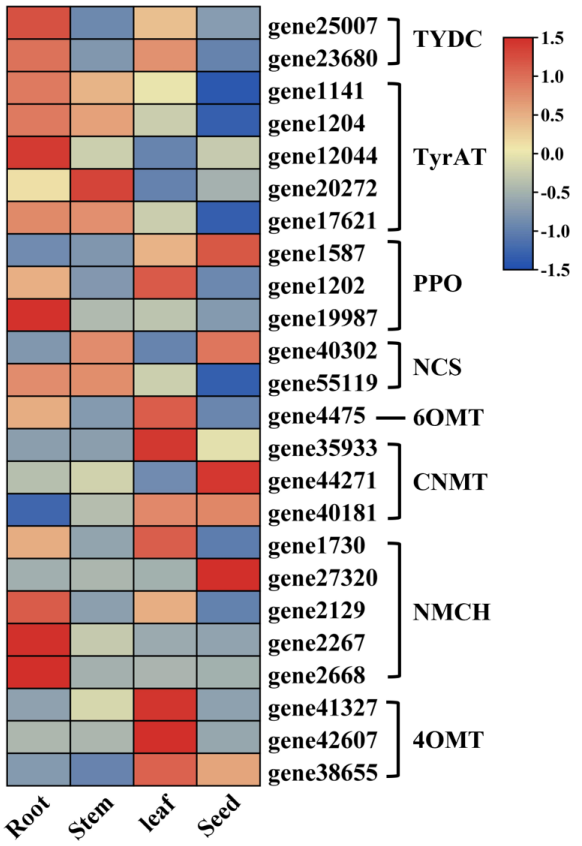


FIGURE 5
Expression patterns of candidate unigenes in the common biosynthetic pathway of BIAs.

gene42607. Gene expression analysis revealed that all these three genes are highly expressed in the leaf. Gene4475 encodes a 6-OMT that catalyzes the methylation of (*S*)-norcoclaurine to afford (*S*)-coclaurine. Six genes encode enzymes similar to CjCoOMT, a columbamine O-methyltransferase from *C. japonica*. Gene expression patterns showed that the expression levels of gene4552 and gene48928 were much higher in the root and stem than in the other two tissues. The scoulerine 9-*O*-methyltransferase (SOMT) converts (*S*)-scoulerine to (*S*)-tetrahydrocolumbamine in the formation of phthalideisoquinoline alkaloids. It can be seen from the phylogenetic analysis that three genes (gene 23249, gene36725, and gene42393) encode SOMTs in *S. acutum*. Gene36725 and gene42393 had similar expression patterns and were both actively expressed in the root and the leaf (Figure 6D). Gene23249 had the highest level in the leaf. Coclaurine *N*-methyltransferase (CNMT) catalyzes the methylation of (*S*)-coclaurine to generate (S)-N-methylcoclaurine. Three genes (gene35933, gene44271, and gene40181) encode proteins homologous to CNMTs from *P. somniferum* and *C. japonica*. Gene expression pattern analysis revealed that the root and stem of *S. acutum* had lower expression levels of these three genes, while the seed had the highest expression level (Figure 6D).

In order to check the reliability of the transcriptome and the DEG data, we selected four genes (gene4552, gene4645, gene29764, and gene30689) that encode enzymes of the CYP450 families or methyltransferases to compare their expression levels determined by qRT-PCR analysis and by the transcriptome data. For the qRT-PCR analysis, gene4552 had highest expression level in the stem and minimal levels in the other tissues. Gene4645, gene29764, and gene30689 all exhibited noticeable expression in two tissues (Figure S5). These results were consistent with the gene expression levels revealed in the transcriptome data (Figures 6B, D), showing that the gene expression levels determined by transcriptome sequencing were reliable.

The epimerization of (*S*)-reticuline to (*R*)-reticuline is catalyzed by the reticuline epimerase (REPI, also known as STORR), and it is an important branch point in the biosynthesis of promorphinan/morphinan subclass of BIAs. Interestingly, we could not find orthologs of REPI from the transcriptome data of *S. acutum*, which is consistent with the lack of (*R*)-reticuline-derived alkaloids in this medicinal plant (Figure 7).

From the quantitative analysis of alkaloids, DEG- and function-based candidate gene screening, as well as the characterized biosynthetic pathways for BIAs, we were able to propose the biosynthetic pathways for a few BIAs in *S. acutum*, including the common biosynthetic pathway (route I), the pathway from (*S*)-reticuline to sinomenine and sinomenine *N*-oxide (route II), magnoflorine (route III), and tetrahydropalmatine (route IV) (Figure 7). Only a few candidate genes were proposed for each step of BIA biosynthesis in *S. acutum*, which should greatly facilitate the expression and function characterization of their encoded enzymes.

4 Discussion

4.1 A combination of SMRT and NGS generated a high-quality full-length transcriptome of *S. acutum*

BIAs are a large and diverse group of specialized plant secondary metabolites with important research and medicinal value (Weber and Opatz 2019). *Sinomenium acutum* is a with a long history of therapeutical applications. Sinomenine is a BIA from *S. acutum* and has been used clinically as an antirheumatic drug (Zhao et al., 2012). Recently, various pharmacological studies have shown the potential of sinomenine in treating cardiovascular diseases (Jiang et al., 2019; Yuan et al., 2021; Zhang et al., 2021). Besides sinomenine, more than 50 BIAs have been isolated from *S. acutum*. However, the biosynthetic

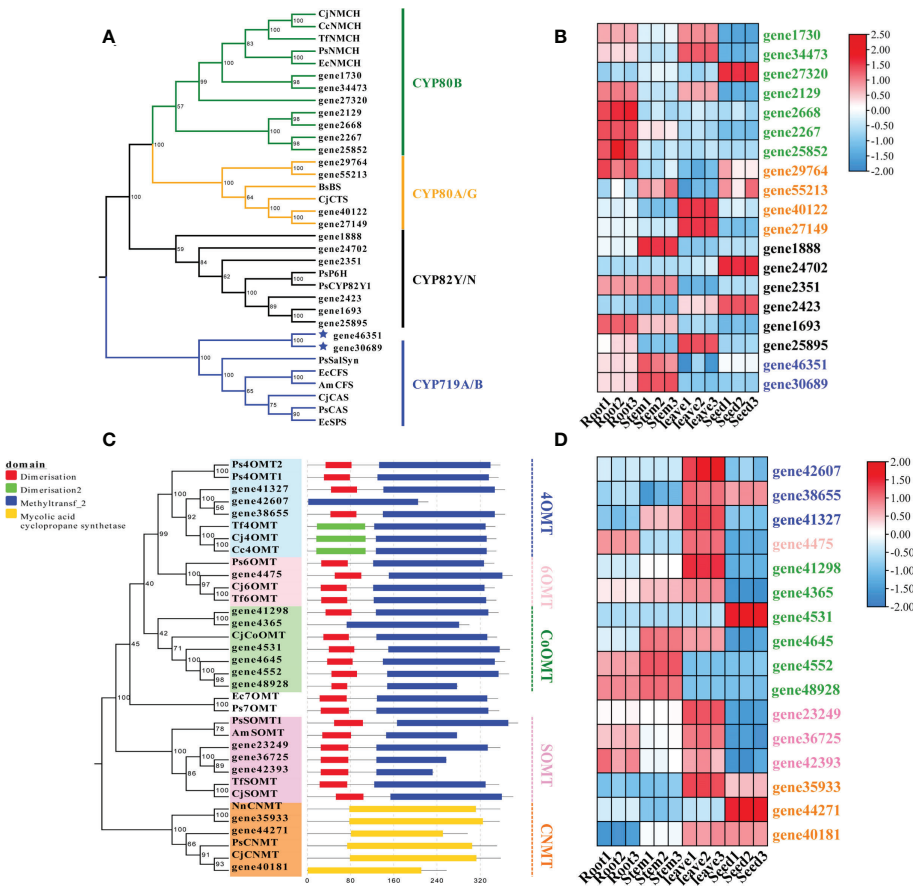


FIGURE 6
Phylogenetic analysis and expression patterns of CYPs and methyltransferases in *S. acutum*. **(A)** Phylogenetic tree of the candidate CYPs with known CYPs involved in BIA biosynthesis. **(B)** Expression levels of the screened CYP-encoding genes in the root, stem, leaf, and seed of *S. acutum*. **(C)** Phylogenetic tree of the candidate methyltransferases for BIA biosynthesis. **(D)** Expression levels of genes encoding the methyltransferases for BIA biosynthesis in the root, stem, leaf, and seed of *S. acutum*..

pathways of BIAs from *S. acutum* remain elusive due to the lack of genomic or transcriptomic information.

With the development of high-throughput sequencing technology, transcriptome sequencing has become a powerful tool for studying the regulation of gene expression and biosynthetic pathways for specific metabolites (He et al., 2018; Xu et al., 2021). However, due to the limited read length of second-generation sequencing (SGS), the quality of transcripts obtained by this technology is often unsatisfactory. In contrast, the third-generation sequencing technology represented by PacBio utilizes SMRT sequencing technology to obtain high-quality, full-length transcripts, which efficiently improves the quality of the transcriptome data (He et al., 2018; McCombie et al., 2019). However, the high error rate of SMRT sequencing is not negligible. The combination of NGS and SMRT can generate full-length transcriptome data with high accuracy (Zhang et al., 2019). In this study, we constructed twelve high-quality RNA libraries of *S. acutum*, including the root, stem, leaf, and seed tissues. The full-length transcriptome data was obtained by combining the SMRT and NGS sequencing techniques. In the absence of a reference genome of *S. acutum*, the high-quality full-length transcriptome data can significantly facilitate the characterization of the secondary metabolite biosynthetic pathways in *S. acutum*.

With the transcriptome data in hand, we annotated the functions of the unigenes using the NR, GO, KOG, and KEGG databases. More than 86% of the unigenes could be annotated in at least one public database, and around one-third of the unigenes were co-annotated in all five databases. In the NR database, 25.88% and 21.25% of the annotated unigenes matched the sequences of *M. cordata* and *N. nucifera*, respectively. It is well-known that *M. cordata* and *N. nucifera* both accumulate plenty of BIAs. Detailed characterization of the matched unigenes among these three species would promote our understanding of the regulation and biosynthesis of BIAs in plants.

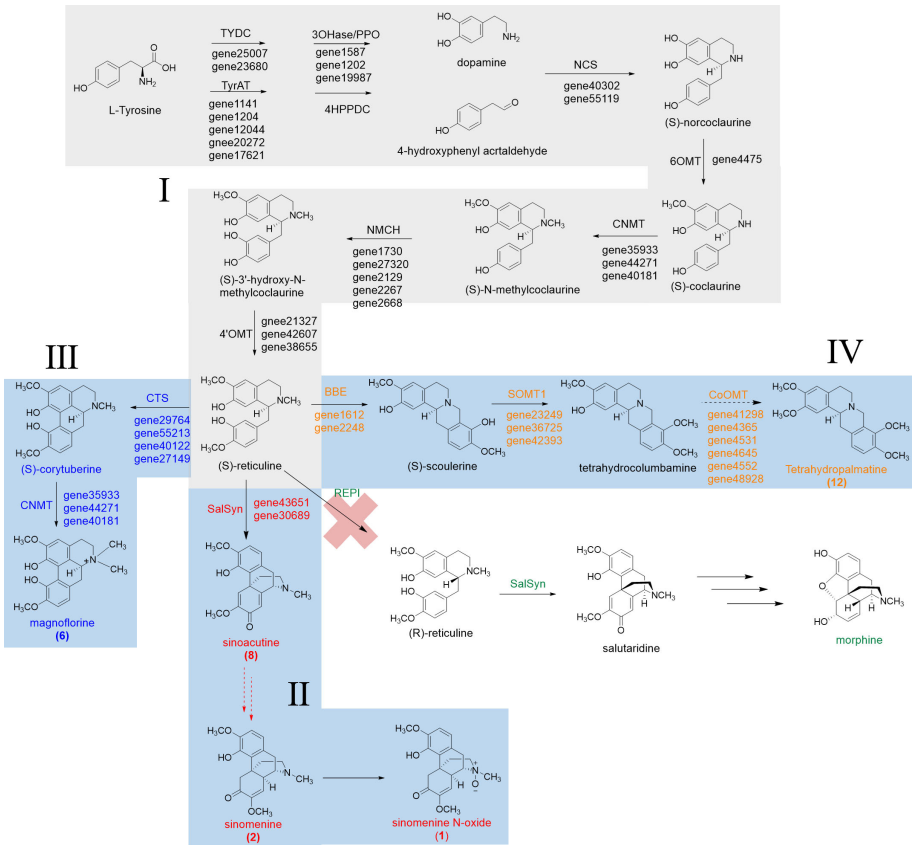


FIGURE 7
Proposed biosynthetic pathways for sinomenine, magnoflorine, and tetrahydropalmatine in *S. acutum*.

4.2 The DEG and WGCNA analysis facilitate the screening of candidate genes for BIA biosynthesis

Although the function annotation using public protein databases can assign the possible function of unigenes from the transcriptome data, screening of the candidate genes involved in BIA biosynthesis still needs more information to differentiate the actual enzymes from their respective homologs. WGCNA is now widely used in the mining of genomic data. The WGCNA algorithm assumes that the gene network follows the scale-free topology rule and generates the gene co-expression matrix. Unlike the traditional gene-to-phenotype mode, which relates individual genes to phenotype, WGCNA focuses on the relationship between a few simplified modules and the traits. The WGCNA analysis can efficiently alleviate multiple inherent problems associated with microarray data analysis (Fuller et al., 2007). The genes in the same module possess a high co-expression similarity, while genes in different modules have a low co-expression similarity.

To reduce the scale of the network, we selected seven representative alkaloids from *S. acutum*, i.e., menisperine, sinomenine, tetrahydropalmatine, sinomenine *N*-oxide, sinoacutine, magnoflorine, and acutumine, to correlate these metabolites with specific modules. In this study, seven modules were detected based on the co-expression network. We found that the pink, purple, and dark red modules correlate much more with the accumulation of the seven BIAs than the other four. We also observed that the pink module had obvious adjacency with the purple and dark red modules, while the purple and dark red modules showed no adjacency to each other. It is reasonable to propose that the unigenes in the pink, purple, and dark red module have a higher possibility of participating in the biosynthesis of BIAs in *S. acutum*.

Analysis of the expression patterns of genes in different tissues can further help to reduce the number of candidate genes. Gene expression profiles were compared among the root, stem, leaf, and seed tissues. Since the root of *S. acutum* has the highest alkaloid contents and the leaf has the lowest

contents, we compared the gene expression patterns in these two tissues and observed 8,270 DEGs. Further analysis of unigenes enriched in KEGG pathway terms revealed a total of 47 unigenes that mapped to the "isoquinoline biosynthetic pathway" category (ko00950). Proteins encoded by these 47 genes were subjected to more detailed function annotation and expression analysis. It could be concluded from the process that combining DEG and WGCNA analysis is an efficient approach to mining the potential genes involved in specific secondary metabolites.

4.3 Proposal of the BIA biosynthetic pathways in *S. acutum*

BIAs are mainly isolated from plants of the Ranunculales order, such as the Papaveraceae, Ranunculaceae, Berberidaceae, and Menispermaceae families (Hagel and Facchini, 2013). The common biosynthetic pathway for isoquinoline alkaloids has been well-studied in *P. somniferum*, *C. japonica*, *N. nucifera*, and *M. cordata* (Beaudoin and Facchini, 2014). Production of many characterized BIAs follows the common route, which starts from the condensation of dopamine and 4-hydroxyphenylacetaldehyde to (*S*)-norcoclaurine and ends at the formation of (*S*)-reticuline. (*S*)-reticuline is a key branch point intermediate to the generation of BIAs. It can undergo different reactions, such as aromatic ring hydroxylation, C–C or C–O coupling, and *O*- or *N*-methylation, to yield the vast structural diversity of BIAs. In this study, we screened 24 genes that encode proteins participating in the biosynthetic route from dopamine and 4-hydroxyphenylacetaldehyde to (*S*)-reticuline. Enzymes of the CYP450 family and methyltransferase family are the key players in converting (*S*)-reticuline to different BIAs. Candidate proteins involved in these steps were mined using the Pfam annotation and phylogenetic analysis using the characterized orthologs as references, including the CYP80, CYP82, and CYP719 subfamilies, as well as orthologs from the 4′-OMT, 6-OMT, CoOMT, SOMT, and CNMT subclasses. With the collected candidate enzymes in hand, we could propose the biosynthetic pathways for some of the BIAs, i.e., sinomenine, sinoacutine, magnoflorine, and tetrahydropalmatine.

Two points are worth mentioning from this study: a) although the root has much higher concentrations for most of the characterized BIAs herein, plenty of the screened genes were not expressed at the highest levels in the root. This phenomenon could be attributed to the spatial and temporal-specific production of some BIAs or their biosynthetic precursors in tissues other than the root. After the biosynthesis, the final or intermediate BIAs could be transported from the different tissues to the root. b) REPI can convert (*S*)-reticuline to (*R*)-reticuline, which is an essential step in the biosynthesis of some pharmaceutically important morphinan BIAs, such as morphine and codeine. However, this enzyme was not annotated in the transcriptome of *S. acutum*. Therefore, conversion of the (*S*)-reticuline intermediate to (*R*)-reticuline does not occur in this plant. This result is consistent with the observation that no (*R*)-reticuline-derived BIAs have been isolated from *S. acutum*.

5 Conclusions

Sinomenium acutum is an important medicinal plant with a long history of application. It can produce more than 50 BIAs. However, the biosynthetic pathways of BIAs in *S. acutum* remain elusive. In the current study, we analyzed the contents of 12 BIAs in four different tissues of *S. acutum*. We obtained its high-quality full-length transcriptome data by combining the NGS and SMRT sequencing. Annotation of the transcripts resulted in 60,675 unigenes. The candidate genes responsible for BIA biosynthesis were mined by the WGCNA and DEG analysis and the KEGG pathway enrichment analysis. Based on the functions of the screened candidate genes, we were able to propose the biosynthetic pathways for some of the BIAs in *S. acutum*, including sinomenine, sinoacutine, magnoflorine, and tetrahydropalmatine. Our work lays the foundation for the characterization of the enzymes involved in BIA biosynthesis and will benefit the heterologous production of high-value BIAs in microbial cell factories.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA843226.

Author contributions

YY: Methodology, data analysis, and writing. YS: Material collection, data analysis, and writing. ZW: Data analysis and writing. MY: Methodology and data analysis. RS: Material collection and data analysis. LX: Material collection. XH: Material collection and conceptualization. CW: Conceptualization and data analysis. XY: Conceptualization, funding acquisition, data analysis, project administration, and writing-reviewing. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Scientific and Technological Research Program of the Tianjin Municipal Education Commission (No. 2019ZD10).

Conflict of interest

Author Ying Sun was employed by the company WuXi App Tec (Tianjin) Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commerical or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1086335/full#supplementary-material

References

Beaudoin, G., and Facchini, P. (2014). Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta* 240 (1), 19–32. doi: 10.1007/s00425-014-2056-8

Bhambhani, S., Kondhare, K., and Giri, A. (2021). Diversity in chemical structures and biological properties of plant alkaloids. *Molecules* 26 (11), 3374. doi: 10.3390/molecules26113374

Bolger, A., Lohse, M., and Usadel, B. (2014). Trimomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* 49 (W1), W317–W325. doi: 10.1093/nar/gkab447

Catania, T., Li, Y., Winzer, T., Harvey, D., Meade, F., Caridi, A., et al. (2022). A functionally conserved STORR gene fusion in papaver species that diverged 16.8 million years ago. *Nat. Commun.* 13 (1), 3150. doi: 10.1038/s41467-022-30856-w

Conesa, A., Götz, S., García-Gómez, J., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18), 3674–3676. doi: 10.1093/bioinformatics/bti610

Dang, T., Onoyovwi, A., Farrow, S., and Facchini, P. (2012). Biochemical genomics for gene discovery in benzylisoquinoline alkaloid biosynthesis in opium poppy and related species. *Methods Enzymol.* 515, 231–266. doi: 10.1016/b978-0-12-394290-6.00011-2

Dastmalchi, M., Park, M., Morris, J., and Facchini, P. (2018). Family portraits: the enzymes behind benzylisoquinoline alkaloid diversity. *Phytochem. Rev.* 17, 279–277. doi: 10.1007/s11101-017-9519-z

Fuller, T., Ghazalpour, A., Aten, J., Drake, T., Lusis, A., and Horvath, S. (2007). Weighted gene co-expression analysis strategies applied to mouse weight. *Mamm. Genome* 18 (6-7), 463–472. doi: 10.1007/s00335-007-9043-3

Gesell, A., Rolf, M., Ziegler, J., Díaz Chávez, M., Huang, F., and Kutchan, T. (2009). CYP719B1 is salutaridine synthase, the c-c phenol-coupling enzyme of morphine biosynthesis in opium poppy. *J. Biol. Chem.* 284 (36), 24432–24442. doi: 10.1074/jbc.M109.033373

Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29 (7), 644–652. doi: 10.1038/nbt.1883

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., et al. (2018). The opium poppy genome and morphinan production. *Science* 362 (6412), 343–347. doi: 10.1126/science.aat4096

Hagel, J., and Facchini, P. (2013). Benzylisoquinoline alkaloid metabolism: a century of discovery and a brave new world. *Plant Cell Physiol.* 54 (5), 647–672. doi: 10.1093/pcp/pct020

He, S., Liang, Y., Cong, K., Chen, G., Zhao, X., Zhao, Q., et al. (2018). Identification and characterization of genes involved in benzylisoquinoline alkaloid biosynthesis in coptis species. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00731

Jiang, W., Fan, W., Gao, T., Li, T., Yin, Z., Guo, H., et al. (2020). Analgesic mechanism of sinomenine against chronic pain. *Pain Res. Manage.* 2020, 1876862. doi: 10.1155/2020/1876862

Jiang, Z., Wang, L., Pang, H., Guo, Y., Xiao, P., Chu, C., et al. (2019). Rapid profiling of alkaloid analogues in sinomenii caulis by an integrated characterization strategy and quantitative analysis. *J. Pharm. Biomed. Anal.* 174, 376–385. doi: 10.1016/j.jpba.2019.06.011

Jin, H., Wang, X., Wang, H., Wang, Y., Lin, L., Ding, J., et al. (2008). Morphinane alkaloid dimers from sinomenium acutum. *J. Nat. Prod.* 71 (1), 127–129. doi: 10.1021/np0704654

Klopfenstein, D., Zhang, L., Pedersen, B., Ramírez, F., Vesztrocy, A., Naldi, A., et al. (2018) GOATOOLS: A Python library for gene ontology analyses. *Sci. Rep.* 8 (1), 10872. doi: 10.1038/s41598-018-28948-z

Li, B., and Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinf.* 12, 323. doi: 10.1186/1471-2105-12-323

Listos, J., Łupina, M., Talarek, S., Mazur, A., Orzelska-Górka, J., and Kotlińska, J. (2019). The mechanisms involved in morphine addiction: an overview. *Int. J. Mol. Sci.* 20 (17), 4302. doi: 10.3390/ijms20174302

Liu, L., Chu, X., Tian, C., Xia, M., Zhang, L., Jiang, J., et al. (2021). Chemo proling and simultaneous analysis of different combinations of sinomenii caulis and ramulus cinnamomi using UHPLC-Q-TOF-MS, GC-MS and HPLC methods. *J. Chromatogr. Sci.* 59 (7), 606–617. doi: 10.1093/chromsci/bmab048

Liu, X., Liu, Y., Huang, P., Ma, Y., Qing, Z., Tang, Q., et al. (2017). The genome of medicinal plant macleaya cordata provides new insights into benzylisoquinoline alkaloids metabolism. *Mol. Plant* 10 (7), 975–989. doi: 10.1016/j.molp.2017.05.007

Liu, Y., Liu, C., Tan, T., Li, S., Tang, S., and Chen, X. (2019). Sinomenine sensitizes human gastric cancer cells to cisplatin through negative regulation of PI3K/AKT/Wnt signaling pathway. *Anticancer Drugs* 30 (10), 983–990. doi: 10.1097/CAD.0000000000000834

Liu, H., Zeng, K., Cao, N., Zhao, M., Jiang, Y., and Tu, P. (2018). Alkaloids from the stems and rhizomes of sinomenium acutum from the qinling mountains, China. *Phytochemistry* 156, 241–249. doi: 10.1016/j.phytochem.2018.09.009

McCombie, W., McPherson, J., and Mardis, E. (2019). Next-generation sequencing technologies. *Cold Spring Harb. Perspect. Med.* 9 (11), a036798. doi: 10.1101/cshperspect.a036798

Narcross, L., Fossati, E., Bourgeois, L., Dueber, J., and Martin, V. (2016). Microbial factories for the production of benzylisoquinoline alkaloids. *Trends Biotechnol.* 34 (3), 228–241. doi: 10.1016/j.tibtech.2015.12.005

Sato, F., and Kumagai, H. (2013). Microbial production of isoquinoline alkaloids as plant secondary metabolites based on metabolic engineering research. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 89 (5), 165–182. doi: 10.2183/pjab.89.165

Wang, C., Peng, D., Zhu, J., Zhao, D., Shi, Y., Zhang, S., et al. (2019). Transcriptome analysis of polygonatum cyrtonema hua: identification of genes involved in polysaccharide biosynthesis. *Plant Methods* 15, 65. doi: 10.1186/s13007-019-0441-9

Wang, C., Peng, D., Zhu, J., Zhao, D., Shi, Y., Zhang, S., et al. (2019). Transcriptome analysis of polygonatum cyrtonema hua: identification of genes involved in polysaccharide biosynthesis. *Plant Methods* 15, 65. doi: 10.1186/s13007-019-0441-9

Weber, C., and Opatz, T. (2019). Bisbenzylisoquinoline Alkaloids. *Alkaloids Chem. Biol.* 15, 81, 1–114. doi: 10.1016/bs.alkal.2018.07.001

Xu, W., Chen, S., Wang, X., Wu, H., Tahara, K., Tanaka, S., et al. (2021a). Effects of sinomenine on the proliferation, cytokine production, and regulatory T-cell

frequency in peripheral blood mononuclear cells of rheumatoid arthritis patients. *Drug Dev. Res.* 82 (2), 251–258. doi: 10.1002/ddr.21748

Xu, D., Lin, H., Tang, Y., Huang, L., Xu, J., Nian, S., et al. (2021b). Integration of full-length transcriptomics and targeted metabolomics to identify benzylisoquinoline alkaloid biosynthetic genes in corydalis yanhusuo. *Hortic. Res.* 8 (1), 16. doi: 10.1038/s41438-020-00450-6

Yuan, M., Zhao, B., Jia, H., Zhang, C., and Zuo, X. (2021). Sinomenine ameliorates cardiac hypertrophy by activating Nrf2/ARE signaling pathway. *Bioengineered* 12 (2), 12778–12788. doi: 10.1080/21655979.2021.2000195

Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D., et al. (2019). PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. *Plant J.* 97 (2), 296–305. doi: 10.1111/tpj.14120

Zhang, M., Wang, X., Shi, J., and Yu, J. (2021). Sinomenine in cardio-cerebrovascular diseases: potential therapeutic effects and pharmacological evidences. *Front. Cardiovasc. Med.* 8. doi: 10.3389/fcvm.2021.749113

Zhang, S., Wang, G., Zuo, T., Zhang, X., Xu, R., Zhu, W., et al. (2020a). Comparative transcriptome analysis of rhizome nodes and internodes in panax. japonicus var. major reveals candidate genes involved in the biosynthesis of triterpenoid saponins. *Genomics* 112 (2), 1112–1119. doi: 10.1016/j.ygeno.2019.06.025

Zhang, Y., Kang, Y., Xie, H., Wang, Y., Li, Y., and Huang, J. (2020b). Comparative transcriptome analysis reveals candidate genes involved in isoquinoline alkaloid biosynthesis in stephania tetrandra. *Planta Med.* 86 (17), 1258–1268. doi: 10.1055/a-1209-3407

Zhao, X., Peng, C., Zhang, H., and Qin, L. (2012). Sinomenium acutum: a review of chemistry, pharmacology, pharmacokinetics, and clinical use. *Pharm. Biol.* 50 (8), 1053–1061. doi: 10.3109/13880209.2012.656847

Zhong, F., Huang, L., Qi, L., Ma, Y., and Yan, Z. (2020). Full-length transcriptome analysis of coptis deltoidea and identification of putative genes involved in benzylisoquinoline alkaloids biosynthesis based on combined sequencing platforms. *Plant Mol. Biol.* 102 (4-5), 477–499. doi: 10.1007/s11103-019-00959-y

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fpls.2023.1104303

OPEN ACCESS

EDITED BY
Li Wang,
Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

REVIEWED BY
Mei Yang,
Wuhan Botanical Garden, Chinese Academy of Sciences (CAS), China
Aalt-Jan Van Dijk,
Wageningen University and Research, Netherlands
Ancheng Huang,
Southern University of Science and Technology, China

*CORRESPONDENCE
Agnieszka Zmienko
akisiel@ibch.poznan.pl

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 21 November 2022
ACCEPTED 11 January 2023
PUBLISHED 26 January 2023

CITATION
Marszalek-Zenczak M, Satyr A, Wojciechowski P, Zenczak M, Sobieszczanska P, Brzezinski K, Iefimenko T, Figlerowicz M and Zmienko A (2023) Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members. *Front. Plant Sci.* 14:1104303. doi: 10.3389/fpls.2023.1104303

Analysis of Arabidopsis non-reference accessions reveals high diversity of metabolic gene clusters and discovers new candidate cluster members

Malgorzata Marszalek-Zenczak[1], Anastasiia Satyr[1], Pawel Wojciechowski[1,2], Michal Zenczak[1], Paula Sobieszczanska[1], Krzysztof Brzezinski[1], Tetiana Iefimenko[3], Marek Figlerowicz[1] and Agnieszka Zmienko[1]*

[1]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland, [2]Institute of Computing Science, Faculty of Computing and Telecommunications, Poznan University of Technology, Poznan, Poland, [3]Department of Biology, National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Metabolic gene clusters (MGCs) are groups of genes involved in a common biosynthetic pathway. They are frequently formed in dynamic chromosomal regions, which may lead to intraspecies variation and cause phenotypic diversity. We examined copy number variations (CNVs) in four *Arabidopsis thaliana* MGCs in over one thousand accessions with experimental and bioinformatic approaches. Tirucalladienol and marneral gene clusters showed little variation, and the latter was fixed in the population. Thalianol and especially arabidiol/baruol gene clusters displayed substantial diversity. The compact version of the thalianol gene cluster was predominant and more conserved than the noncontiguous version. In the arabidiol/baruol cluster, we found a large genomic insertion containing divergent duplicates of the *CYP705A2* and *BARS1* genes. The *BARS1* paralog, which we named *BARS2*, encoded a novel oxidosqualene synthase. The expression of the entire arabidiol/baruol gene cluster was altered in the accessions with the duplication. Moreover, they presented different root growth dynamics and were associated with warmer climates compared to the reference-like accessions. In the entire genome, paired genes encoding terpene synthases and cytochrome P450 oxidases were more variable than their nonpaired counterparts. Our study highlights the role of dynamically evolving MGCs in plant adaptation and phenotypic diversity.

KEYWORDS

copy number variation, biosynthetic gene cluster, secondary metabolism, oxidosqualene cyclase, triterpenes, cytochrome P450

Introduction

Plants are able to produce a variety of low molecular weight organic compounds, which enhance their ability to compete and survive in nature. Secondary metabolites are not essential for plant growth and development. However, they are often multifunctional and may act both as plant growth regulators and be engaged in primary metabolism or plant protection (Isah, 2019; Erb and Kliebenstein, 2020). The ability to produce particular types of compounds is usually restricted to individual species or genera. Therefore, these compounds are enormously diverse and have a wide range of biological activities. In plants, genes involved in a common metabolic pathway are typically dispersed across the genome. In contrast, functionally related genes that encode the enzymes involved in specialized metabolite biosynthesis in bacteria and fungi are frequently coexpressed and organized in so-called operons (Boycheva et al., 2014; Nützmann et al., 2018). Similar gene organization units called biosynthetic gene clusters or metabolic gene clusters (MGCs) have recently been found in numerous plant species. MGCs have typically been defined as a group of three or more genes that i) encode a minimum of three different types of biosynthetic enzymes, ii) are involved in the consecutive steps of a specific metabolic pathway and iii) are localized in adjacent positions in the genome or are interspersed by a limited number of intervening (i.e., not functionally related) genes (Nützmann and Osbourn, 2014; Kautsar et al., 2017). A typical MGC contains a "signature" enzyme gene involved in the major (usually first) step of a biosynthetic pathway. In this step, the metabolite scaffold is generated that determines the class of the pathway products (e.g., terpenes or alkaloids). This scaffold is further modified by "tailoring" enzymes encoded by other clustered genes, e.g., cytochrome P450 oxidases (CYPs), acyltransferases or alcohol dehydrogenases. The contribution of other enzymes encoded by peripheral genes (i.e., located outside the MGC), and the connection network between different metabolite biosynthesis pathways may result in additional diversification of the biosynthetic products (Huang et al., 2019). Currently, there are over 30 known MGCs in plants from various phylogenetic clades, and new MGCs are being discovered. Their sizes range from 35 kb to several hundred kb. However, clusters of functionally related nonhomologous genes are still considered unusual in plant genomes.

In *Arabidopsis thaliana* (hereafter Arabidopsis), four MGCs have been discovered thus far (Supplemental Table S1). They are involved in the metabolism of specialized triterpenes: thalianol, marneral, tirucalladienol, arabidiol and baruol. Triterpenes constitute a large and diverse group of natural compounds derived from 2,3-oxidosqualene cyclization in a reaction catalyzed by oxidosqualene cyclases (OSCs) (Thimmappa et al., 2014). Out of 13 OSC genes known in the Arabidopsis genome, five (*THAS1*, *MRN1*, *PEN3*, *PEN1*, *BARS1*) are located within MGCs and encode the "signature" enzymes of the MGCs (Field and Osbourn, 2008; Field et al., 2011; Boutanaev et al., 2015). The thalianol gene cluster contains five members involved in thalianol production and in its conversion to another triterpene, thalianin (Fazio et al., 2004; Field and Osbourn, 2008; Huang et al., 2019). In the reference genome, this MGC is ~45 kb in size. The thalianol synthase gene *THAS1* as well as *CYP708A2*, *CYP705A5* and *AT5G47980* (BAHD acyltransferase) genes are tightly clustered together, with only one noncoding transcribed locus (*AT5G07035*) between them. The fifth member, acyltransferase *AT5G47950*, is separated from the rest of the cluster by *RABA4C* and *AT5G47970* intervening genes. The marneral gene cluster is ~35 kb in size and is the most compact plant MGC described to date. It is made up of three members: the marneral synthase gene *MRN1*, the marneral oxidase gene *CYP71A16* and the gene *CYP705A12*, whose function is unknown (Xiong et al., 2006; Field et al., 2011; Go et al., 2012). Additionally, there are three noncoding transcribed loci (*AT5G00580*, *AT5G06325* and *AT5G06335*) located between *CYP701A16* and *MRN1*. The tirucalladienol gene cluster is ~47 kb in size and includes five members: tirucalla-7,24-dien-3β-ol synthase gene *PEN3*, an uncharacterized acyltransferase gene *SCPL1*, which was identified based on its coexpression with *PEN3*, *CYP716A1*, which is involved in the hydroxylation of tirucalla-7,24-dien-3β-ol, as well as *AT5G36130* and *CYP716A2* (Morlacchi et al., 2009; Boutanaev et al., 2015; Wisecaver et al., 2017). The contiguity of this MGC is interrupted by four intervening genes (*CCB3*, *AT5G36125*, *HCF109* and *AT5G36160*) and the noncoding locus *AT5G05325*. The arabidiol/baruol gene cluster is most complex and has an estimated size of 83 kb. It encompasses two closely located OSCs, *PEN1* and *BARS1*, sharing 91% similarity at the amino acid level. *BARS1* encodes a multifunctional cyclase that produces baruol as its main product (Lodeiro et al., 2007). *PEN1* encodes arabidiol synthase and is adjacent to *CYP705A1*, which is involved in arabidiol degradation upon jasmonic acid treatment (Xiang et al., 2006; Castillo et al., 2013; Sohrabi et al., 2015). The role of the remaining genes in the arabidiol/baruol gene cluster (*CYP702A2*, *CYP702A3*, *CYP705A2*, *CYP705A3*, *CYP705A4*, *CYP702A5*, *CYP702A6* as well as acyltransferases *AT4G15390* and *BIA1*) has not been determined; however, they displayed coexpression with either *PEN1* or *BARS1* (Wada et al., 2012; Wisecaver et al., 2017). There are few intervening loci in the arabidiol/baruol gene cluster, including a protein-coding gene *CSLB06*, two pseudogenes *CYP702A4P* and *CYP702A7P* and one novel transcribed region *AT4G06325*.

Plant MGCs are thought to have arisen by duplication and subsequent neo- or subfunctionalization of genes involved in primary metabolism, which might have been followed by the recruitment of additional genes to the newly forming biosynthetic pathway (Nützmann and Osbourn, 2014). MGCs are frequently located within dynamic chromosomal regions, e.g., subtelomeric regions, centromeric regions or regions rich in transposable elements (TEs), where the possibility of bringing together the beneficial sets of genes by structural rearrangements may be higher than in the rest of the genome, thus promoting MGC formation (Field et al., 2011). However, the same factors may also contribute to further genetic modifications and alteration of the plant metabolic profile, thus making such MGCs "evolutionary hotspots". To verify this scenario, we evaluated the intraspecific diversity of Arabidopsis MGCs and examined whether this diversity is associated with trait variation. Here, we present a detailed picture of MGC copy number variations (CNVs), describe the discovery of novel, nonreference genes in the arabidiol/baruol gene cluster and reveal the links between the variation in MGC structure and plant adaptation to different natural environments.

Results

MGCs differ in levels of copy number polymorphism

We started our analysis by aligning each MGC with the common CNVs in the Arabidopsis genome, which were identified previously (Zmienko et al., 2020). As expected, each MGC had a substantial overlap with the variable regions: 100% for the thalianol gene cluster, 79.6% for the tirucalladienol gene cluster, 53.1% for the arabidiol/baruol gene cluster, and 52.8% for the marneral gene cluster (Figure 1A). However, the potential impact of CNVs on the clustered genes differed among the MGCs (Supplemental Figure S1; Supplemental Table S2). In the thalianol gene cluster, most CNVs were grouped in the region spanning *AT5G47980*, *CYP705A5*, *CYP708A2* and *THAS1*, while *AT5G47950* was covered only by the largest variant CNV_18592 (241 kb in size), which encompassed the entire cluster. In the arabidiol/baruol gene cluster, the CNVs (0.6 kb to 21 kb in size) were grouped into three distinct regions separated by invariable segments. The first variable region overlapped with *CYP702A2* and *CYP702A3*. The second variable region overlapped with *CYP705A2*, *CYP705A3* and *BARS1*. The CNVs in the third variable region were mostly intergenic and overlapped with only two genes, *CYP702A5* and *CYP702A6*. *CYP705A1*, *PEN1*, *CYP705A4*, *AT4G15390* and *BIA1* were not covered by any common CNV. In the tirucalladienol gene cluster, the CNVs accumulated in the 5' part of the cluster, and none of them overlapped with *SCPL1*. Notably, upstream of the tirucalladienol gene cluster, a region genetically divergent from the surrounding genomic segments, called a hotspot of rearrangements, was previously described (Jiao and Schneeberger, 2020). Smaller hotspots of rearrangements were also found between *CYP716A1* and *AT5G36130* in the same MGC as well as in one variable segment of the arabidiol/baruol gene cluster. It was demonstrated that the hotspots of rearrangements are highly variable in the Arabidopsis population, which was in agreement with the observed increased CNV rate in these genomic regions. The CNV arrangement in the marneral gene cluster was strikingly different from that in any other MGC in that all variants were intergenic and did not overlap with the marneral cluster genes.

For each MGC, there were CNVs that overlapped only part of the cluster. This indicated that in some accessions, gene deletions/duplications might have altered MGC composition and consequently affected the entire biosynthetic pathway. To evaluate this possibility, we retrieved copy number data for 31 genes (clustered

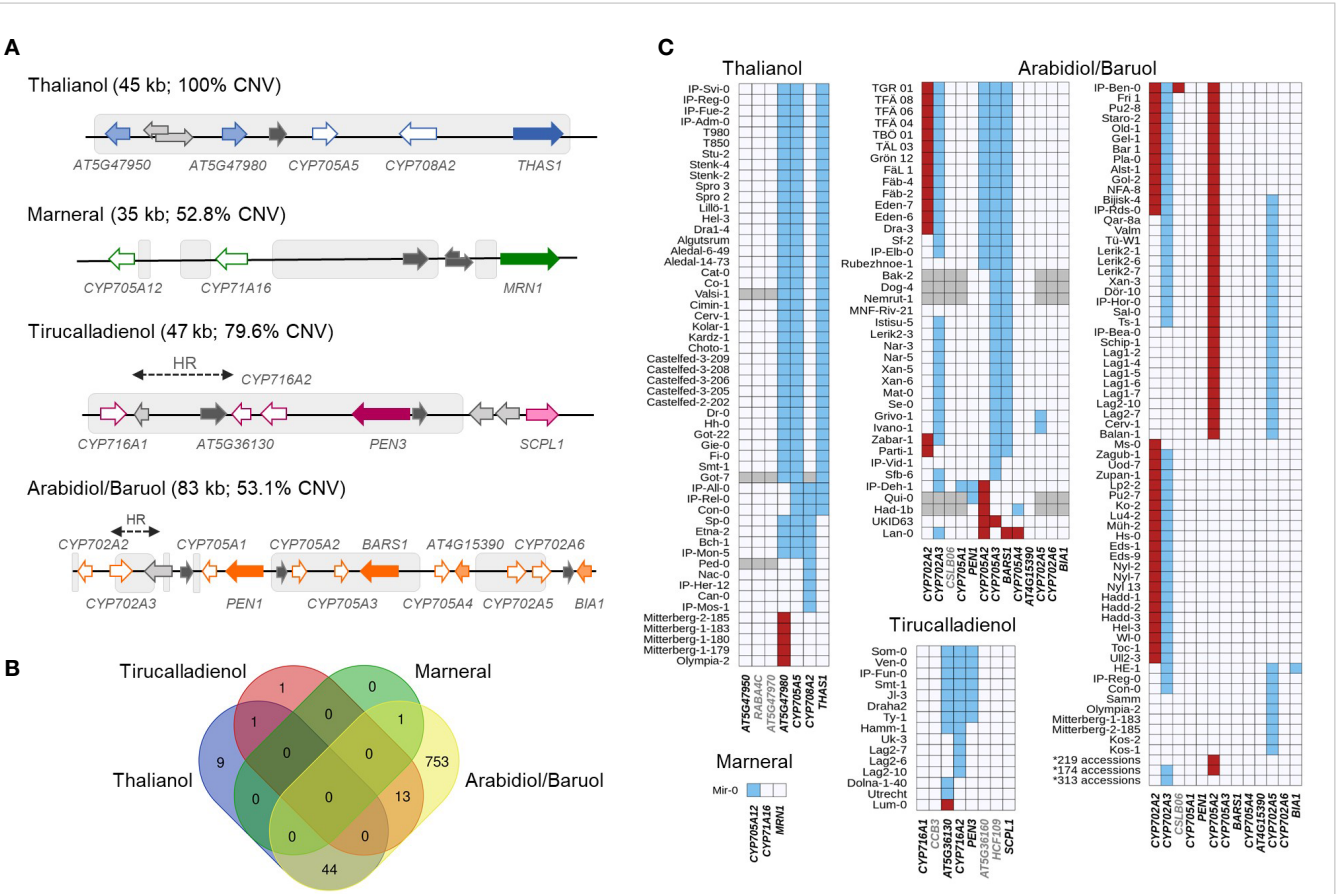


FIGURE 1
Copy number variation of Arabidopsis metabolic gene clusters. **(A)** MGC overlap with CNV regions. Colored arrows with white filling denote CYPs. Arrows with dark color filling denote OSCs. Arrows with light color filling denote other types of MGC genes. Intervening genes are in light grey. Noncoding genes are in dark grey. Grey boxes indicate overlap with CNV regions. HR – hotspot of rearrangements; **(B)** Number and overlap among the accessions with detected gene copy numbers in each of four MGCs; **(C)** Patterns of gene copy number variation in each MGC. Red – gain; blue – loss, grey – no assignment. Names of the genes considered as MGC members are in black; names of the intervening genes are in grey. Source data for histograms are in Supplemental Table S6.

and intervening genes in all MGCs), each from 1,056 accessions (RD dataset; Supplemental Table S3), and supplemented them with multiplex ligation-dependent amplification assays for 232 accessions (MLPA dataset; Supplemental Table S4) and droplet digital PCR-based genotyping assays for 20 accessions (ddPCR dataset; Supplemental Table S5). We defined the thresholds for detecting duplications and deletions for each data type. Next, we assigned the copy number status of each gene in each accession ("REF", "LOSS" or "GAIN") by combining all three datasets (Supplemental Table S6). Out of the genotypes assigned with two or three approaches, 98.8% were fully concordant, and most of the remaining discrepancies could be resolved manually (Supplemental Figures S2-S4; Supplemental Table S7). The combined genotyping data for 1,152 accessions were further used to assess and compare MGC variation at the gene level.

Only 28.6% of the assayed accessions had no gene gains or losses in any MGC (Figure 1B). This included 65% of accessions from the German genetic group and 39% of accessions from the Central Europe group. In contrast, the vast majority (at least 90%) of accessions from groups known to be genetically distant from the reference genome (North Sweden, Spain, Italy-Balkan-Caucasus, and Relict groups) displayed gene CNV in at least one MGC. We note that the real number of invariable accessions could be even lower since for 96 accessions, some MGC genes were not genotyped. Altogether, 19 genes were affected: four in the thalianol cluster, one in the marneral cluster, three in the tirucalladienol cluster and 11 in the arabidiol/baruol cluster (Figure 1C). The latter was also most variable in terms of the number of accessions carrying CNVs and the diversity of CNV patterns. For two genes, we detected only copy gains, and for 11, we detected only losses, while six genes were multiallelic (with both gains and losses). As expected, these genes resided in the previously defined variable regions. Remarkably, we did not observe complete loss or gain of the entire MGC in any accession. In the next step, we inspected in more detail the level of diversity of each MGC.

The compact version of the thalianol gene cluster is predominant and more conserved than the reference-like noncontiguous version

A survey with a combination of RD, MLPA and ddPCR approaches revealed 54 accessions with copy number changes in the thalianol gene cluster, which followed five distinct patterns, and *AT5G47950* was the only invariant gene in all accessions (Figure 2A). The most common (variant A) was the deletion of a region encompassing *AT5G47980* and *CYP705A5*, combined with the deletion of *THAS1*. We detected this variant in 37 accessions from six countries: Sweden (13), Italy (8), Germany (6), Spain (5), Bulgaria (3) and Portugal (2). We also confirmed the existence of two previously reported rare variants (Liu et al., 2020a). One of them (variant B) was a large deletion spanning *AT5G47980, CYP705A5* and *CYP708A2*. We found this variant in two accessions from Germany (Bch-1, Sp-0), in one from Italy (Etna-2) and in one from Spain (IP-Mon-5). The other one (variant C) was a deletion of a single gene, *CYP708A2*, which we found in five accessions, mainly Relicts, originating from Spain (Can-0, Ped-0, IP-Her-12 and Nac-0) and Portugal (IP-Mos-1). We also found a new type of deletion (variant D) in two Spanish Relicts (IP-Rel-0 and Con-0) and one non-Relict (IP-All-0). The deletion spanned *CYP705A5*, *CYP708A2* and *THAS1* (Supplemental Figure S5). The last variant (variant E) was a duplication of the acyltransferase gene *AT5G47980*, which was found in four accessions from Italy (Mitterberg-1-179, Mitterberg-

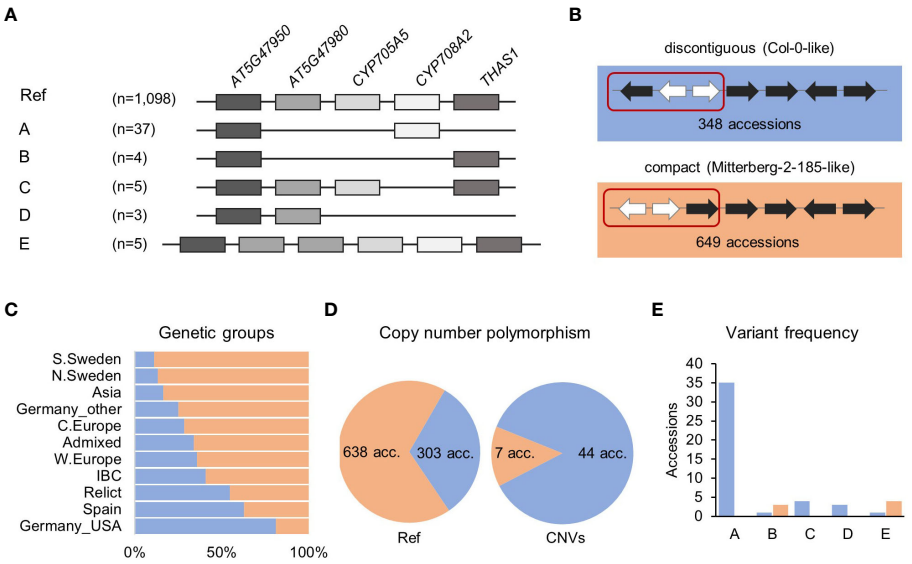


FIGURE 2
Structural variation of thalianol gene cluster. **(A)** Five types of CNVs that change the number of thalianol cluster genes. The position of intervening genes is ignored and they are not shown. Gene orientation is disregarded. **(B)** Two versions of thalianol gene cluster organization. Clustered genes are in black; interfering genes are in white. **(C)** The frequency of the two thalianol gene cluster versions (discontiguous and compact) among the genetic groups. **(D)** Rate of copy number polymorphism within discontiguous and compact clusters. **(E)** Frequency of variants presented in **(A)** among the accessions with different cluster organizations. The number of presented accessions in panels is 1,152 for **(A)** – genotyping, 997 for **(B, C)** – inversion detection and 992 for **(D, E)** – the intersection of the above.

1-180, Mitterberg-1-183, Mitterberg-2-185) and one from Greece (Olympia-2). The presence of a tandem duplication ~3 kb in size in Mitterberg-2-185 was confirmed by sequence analysis of its *de novo* genomic assembly (Supplemental Figure S6). The duplication spanned the entire *AT5G47980* and its flanks (0.5 kb upstream and 0.7 kb downstream) and differed from its copy only by two mismatches and a 1-bp gap. The predicted protein products of both gene copies were identical and shorter than the reference acyltransferase (404 aa versus 443 aa), but they possessed complete transferase domains (pfam02458).

In the Mitterberg-2-185 assembly, we also detected a chromosomal inversion (with respect to the reference genome orientation) spanning *AT5G47950* and two intervening genes, *RABA4C* and *AT5G47970*. This resulted in a more compact cluster organization compared to the reference (Figure 2B). Similar inversions were previously detected in 17 other accessions (out of 22 analyzed), which indicated that the compact version of the thalianol gene cluster might be predominant in Arabidopsis (Liu et al., 2020a). To verify this possibility, we set up a bioinformatic pipeline for detecting genomic inversions based on paired-end genomic read analysis in 997 accessions. We correctly detected inversions in 12 out of 15 previously analyzed accessions, which indicated the good sensitivity of our method. Altogether, we found inversions, 12.8 kb to 15.4 kb in size, spanning the *AT5G47950*, *RBAA4C* and A*T5G47970* genes in 649 accessions (65%), which fully confirmed our predictions (Supplemental Table S8). The compact version of the thalianol gene cluster was dominant in the South and North Sweden genetic groups as well as in the Asia group (83.6% to 88.9%), while the discontiguous version was mainly observed among the U.S.A. accessions and was also slightly more abundant in the Spain genetic group (Figure 2C). There was a similar frequency of discontiguous and compact versions among the Relicts (12 and 10 accessions, respectively). Interestingly, the CNV frequency substantially differed between the accessions with different cluster organization (Figures 2D, E). The compact cluster was more conserved; copy number changes (variants B and E) affected only 1.1% of the accessions in this group. The remaining variants, including deletions spanning the *THAS1* signature gene, were found exclusively among the accessions with the reference-like cluster type. Altogether, 12.7% of accessions with discontiguous clusters were affected by CNVs.

Marneral and tirucalladienol gene clusters display little structural variation

Analysis of RD and MLPA data confirmed exceptionally low variability of marneral cluster genes. One private variant, which we detected in Mir-0 and confirmed by Sanger sequencing, was 1.2 kb in size and spanned the first exon of the *CYP705A12* gene, which resulted in the truncation of its predicted protein product (Supplemental Figure S7). Apart from that, we did not detect any common gene duplications or deletions within this MGC. Likewise, we observed low variation in the tirucalladienol gene cluster. In 15 accessions (1.4%), deletions or duplications occurred in the region spanning the *AT5G36130*, *CYP716A2* and *PEN3* genes and affected one, two or all of them. Differences between the countries indicated that these structural variants were of local origin (Supplemental Figure 8). Sequence analysis of *de novo* genomic assemblies for Ty-1 and Dolna-1-40 confirmed the predicted deletion patterns in these accessions. It should be noted that, according to a recent study, *AT5G36130* and *CYP716A2* gene models are misannotated, and they jointly encode a single protein of the CYP716A subfamily with cytochrome oxidase activity (Yasumoto et al., 2016) (Supplemental Figure S9). Therefore, a full-length gene was absent from all 15 accessions with CNVs in the tirucalladienol gene cluster (Figure 1C).

Intraspecies variation in the arabidiol/baruol gene cluster reveals a novel OSC gene

The arabidiol/baruol gene cluster was the most heterogeneous of all the MGCs. Consistent with the segmental CNV coverage, there were apparent differences in the variation frequency between the genes. At the cluster's 5' end, *CYP702A2* was duplicated in 50 accessions, and *CYP702A3* was deleted in 564 accessions, including approximately 70% of all analyzed accessions from Sweden and Spain. In contrast, genes located at the 3' end of the cluster showed little variation. There were *CYP702A5* deletions in 35 accessions, *CYP705A4* deletions in two accessions, and *BIA1* deletion in one accession, while *CYP702A6* and *AT4G15390* were invariable in copy number.

The two OSCs, *PEN1* and *BARS1*, were located in segments with opposite variation levels. *PEN1* and the neighboring gene *CYP705A1*, both implicated in the arabidiol biosynthesis pathway, were stable in copy number, except for three accessions with full or partial gene deletions: the Qui-0 and IP-Deh-1 accessions from Spain and the Kyoto accession from Japan. In the latter, we confirmed partial deletion of both genes by analysis of its *de novo* genomic assembly (Jiao and Schneeberger, 2020). In contrast, *BARS1*, *CYP705A2* and *CYP705A3* were all deleted in several accessions originating from Sweden. We also observed smaller deletions or duplications in this genomic segment, of which the most remarkable was the duplication of *CYP705A2*, detected in 433 (37.6%) accessions. Since the genotypic data for *CYP705A2* and *BARS1* were noisy and indicated more variation than could be revealed by our standard genotyping, we manually inspected short read genomic data that mapped in this region (examples are presented in Supplemental Figure S10). In most accessions, *BARS1* lacked the largest intron, where the *ATREP11* TE (RC/Helitron superfamily) is annotated, which might explain the lower RD values for *BARS1* compared to other genes (see Supplemental Figure S3). Surprisingly, we also observed a mix of reads mapping to *CYP705A2* and *BARS1* loci with and without mismatches in a large number of accessions. Thus, we called SNPs in the coding sequences of both genes to obtain more information on their diversity. Numerous heterozygous SNPs were called in both genes in the above accessions. Because Arabidopsis is a self-pollinating species and therefore highly homozygous, we hypothesized that the reads with mismatches originated from duplicated loci, which showed similarity to *CYP705A2* and *BARS1* and mapped to the reference gene models, resulting in heterozygous SNP calls. In support of this hypothesis, we detected heterozygous SNPs at the *CYP705A2* locus in 90.6% of accessions with this gene's duplication but only in 10.7% of accessions without changes in its

copy number (Wilcoxon rank sum test with continuity correction, *p* value $<2.2\times10^{-16}$; Supplemental Figure S11A). Additionally, heterozygous SNPs at the *BARS1* locus were present in the same accessions (Pearson's correlation coefficient r = 0.86; Supplemental Figure S11B), although we found only one duplication of *BARS1* with our genotyping methods. We concluded that the sequence differences between *BARS1* and its duplicate prevented its detection by RD or MLPA assays. We also observed low but nonzero read coverage and homozygous SNPs at both loci in some accessions with intermediate RD values for *CYP705A2* (RD_{mean} = 1.5) and *BARS1* (RD_{mean} = 0.6) and with the clear loss of *CYP705A3* (RD_{mean} = 0). In agreement with the gene duplication scenario, this could be explained by the presence of *CYP705A2* and *BARS1* duplicates but absence of the entire region spanning the reference genes *CYP705A2*, *CYP705A3* and *BARS1*.

To identify the cryptic *BARS1* duplication, we analyzed genomic assemblies of seven accessions: An-1, Cvi-0, Kyoto, Ler-0, C24, Eri-1 and Sha (Jiao and Schneeberger, 2020), four of which were also genotyped in our study (Figure 3A). We reannotated the entire arabidiol/baruol cluster region in each accession and compared it with the reference (Supplemental Table S9). In six accessions, *BARS1* lacked the largest intron, as indicated earlier by short read data (Supplemental Figure S12). In the Cvi-0, Eri-1 and Ler-0 accessions,

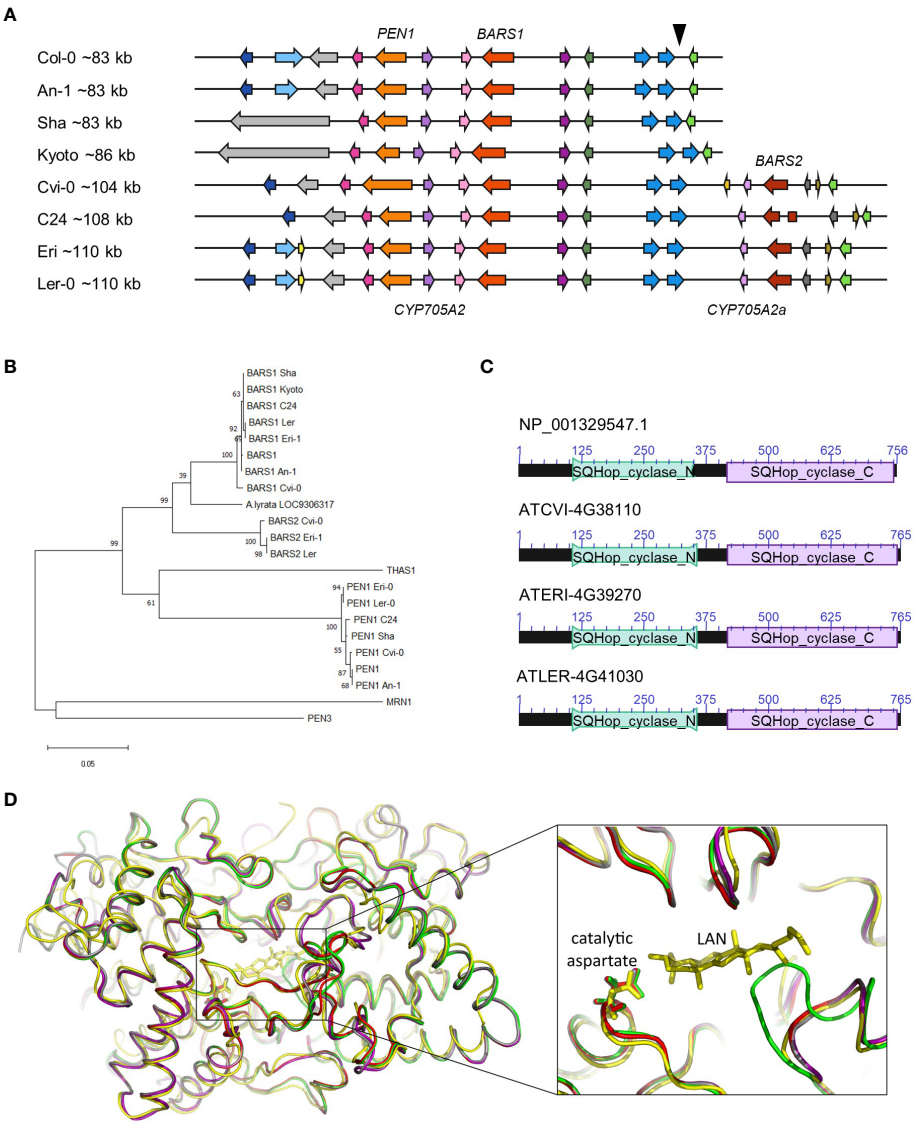


FIGURE 3
BARS2 is a *BARS1* duplicate absent from the reference genome and encodes oxidosqualene synthase. **(A)** Organization of arabidiol/baruol gene cluster in Col-0 and seven nonreference accessions. The genomic insertion including *CYP705A2a* and *BARS2* genes is marked with a triangle above the reference cluster. **(B)** Phylogeny of amino acid sequences of clade II OSCs residing in clusters. BARS1 ortholog from *A.lyrata* (LOC9306317) is included. The maximum likelihood tree was generated using the MEGA11 package with Jones-Taylor (JTT) substitution matrix and uniform rates among sites. Values along branches are frequencies obtained from 1000 bootstrap replications. **(C)** Conserved protein domains encoded in *BARS1* (Col-0) and *BARS2* (Cvi-0, Eri-1, Ler-0) genes. SQHop_cyclase_N - squalene-hopene cyclase N-terminal domain (Pfam 13249). SQHop_cyclase_C - squalene-hopene cyclase C-terminal domain (pfam13243) **(D)** 3D models of baruol synthase proteins encoded by *BARS1* and *BARS2*, predicted by ColabFold software, superposed with the crystal structure of human oxidosqualene cyclase in a complex with lanosterol (LAN). The enlargement box highlights the positions of the catalytic aspartate residue in the predicted models. Colors mark superposed models: green (Col-0 BARS1 isoform NP_193272.1), red (Col-0 BARS1 isoform NP_001329547.1), purple (Cvi-0 BARS1 ATCVI-4G38020), grey (Cvi-0 BARS2 ATCVI-4G38110) and yellow (human OSC PDB ID: 1W6K).

we identified a nonreference gene encoding a protein with ~91% identity to baruol synthase 1 (Supplemental Figure S13). In C24, it was also present but interrupted by ATCOPIA52 retrotransposon insertion, resulting in two shorter ORFs. Based on phylogenetic analysis, we concluded that the identified gene was indeed a *BARS1* duplicate, and we named it *BARS2* (Figure 3B). The differences in the exons of the *BARS1* and *BARS2* sequences matched the heterozygous SNP positions very well (Supplemental Figure S14). Their introns were much more divergent, which likely affected RD genotyping. Likewise, the probe targeting the *BARS1* locus was located in a highly divergent region, which prevented us from detecting this duplication with MLPA.

The proteins encoded by *BARS2* in Cvi-0, Eri-1 and Ler-0 possessed both N-terminal and C-terminal squalene-hopene cyclase domains, typical for OSCs (Figure 3C). We performed three-dimensional (3D) modeling of two reference (Col-0) isoforms of baruol synthase 1 (the product of *BARS1*) and its counterpart from Relict Cvi-0 as well as putative baruol synthase 2 (the product of *BARS*2) from Cvi-0 using ColabFold software. Next, we superposed these models with the experimental crystal structure of human OSC, available in a complex with its reaction product lanosterol (Thoma et al., 2004; Jumper et al., 2021) (Supplemental Information). All structures were highly similar, and we were able to identify potential substrate-binding cavities in the plant enzymes (Figure 3D; Supplemental Table S10). Notably, the catalytic aspartate residue D455 present in the human cyclase had its counterparts in the plant OSCs: D493 in the reference isoform NP_193272.1 and D490 in the remaining proteins (Supplemental Data 1-5). Together, our data indicated that *BARS2* encoded a novel, thus far uncharacterized OSC. As expected, we also found *CYP705A2* duplication in the C24, Cvi-0, Eri-1 and Ler-0 assemblies, and we named it *CYP705A2a*. It had 84% identity with *CYP705A2* at the nucleotide level and 88% similarity at the protein level (Supplemental Figure S15). *CYP705A2a* and *BARS2* were adjacent to each other and located on the minus strand of the large genomic sequence insertion between *CYP702A6* and *BIA* genes (Figure 3A), next to an ~5 kb long interspersed nuclear element 1 (LINE-1) retrotransposon and some shorter, undefined ORFs. The presence of the insertion increased the size of the entire arabidiol/baruol gene cluster by 21-27 kb.

Structural diversity of the arabidiol/baruol gene cluster is associated with the climatic gradient and root growth variation

In the next step, we used the results from the SNP analysis to evaluate the presence/absence variation of both reference (*CYP705A2* + *BARS1*) and nonreference (*CYP705A2a* + *BARS2*) gene pairs in the Arabidopsis population (Supplemental Table S11). The group with only the reference gene pair present was the largest (PP-AA; 628 accessions). Nearly one-third of the population had both gene pairs (PP-PP; 326 accessions). We also separated two smaller groups with the local range of occurrence. The first one, with only the nonreference gene pair, was found in Azerbaijan, Spain, Bulgaria, Russia, Serbia and the U.S.A. (AA-PP; 14 accessions). The last group, where we did not detect any of these genes, was mostly observed at the Bothnian Bay coast collection site in North Sweden (AA-AA; 15 accessions). For 73 accessions, the data were inconclusive. The accuracy of group assignments was validated by sequence analysis of *de novo* genomic assemblies for An-1, Kyoto, Mitterberg-2-185 and Kn-0 (PP-AA group) as well as Cvi-0, Ler-0, Dolna-1-40 and Ty-1 (PP-PP group). Additionally, the results of PCR amplification with gene-specific primers and genomic DNA template for a subset of 36 accessions from all four groups confirmed the differences between them (Supplemental Figure S16). We could not detect *BARS2*-specific products in many samples from the AA-PP group; however, we did detect the band for *CYP705A2a*. We suppose that the *BARS2* sequence might further diverge in this minor group.

The accessions with the nonreference gene pair (AA-PP; PP-PP) dominated among Relicts (81%) and among the Spain (60%) and Italy/Balkan/Caucasus (89.6%) genetic groups but constituted the minority at the northern and eastern margins of the species range (North Sweden 18.6%, South Sweden 16%, Asia 9.4%; Figure 4B). They were also mostly absent among U.S.A. accessions. The widespread presence of *CYP705A2a* and *BARS2* genes in Relicts suggested that the duplication event preceded the recent massive species migration, which took place in the postglacial period and shaped the current Arabidopsis population structure (Lee et al., 2017). We next visualized the four groups in principal component analysis (PCA) plots generated with genome-wide biallelic SNPs (1001 Genomes Consortium, 2016; Zmienko et al., 2020). At a low linkage disequilibrium parameter, where the contribution of the ancestral alleles to PCA was highest, there was a clear convergence of the PC1 and PC2 components with the presence/absence of gene duplication (Figure 4C; Supplemental Figure S17). This suggested that the presence/absence of the genomic insertion containing *CYP705A2a* and *BARS2* genes had some impact on the current geographic distribution of the Arabidopsis accessions. We then evaluated the accessions' latitudes of origin and found that accessions with the nonreference gene pair originated from significantly lower latitudes compared to the remaining accessions (one-way rank-based analysis of variance, ANOVA, p value<0.001, followed by Dunn's test with BH correction, p value<0.001) (Figure 4D). This difference was noticeable even within individual countries and was significant for Germany, Spain and Italy (Supplemental Figure S18). We observed the reverse trend in Russia, where PP-AA accessions were in great excess (88%), and in France; however, we also noticed that PP-AA accessions outnumbered PP-PP accessions in the Pyrenees, Alps and Tian Shan mountain ranges (Supplemental Information). This result suggested that there was an association between arabidiol/baruol gene cluster variation and environmental conditions; therefore, we decided to investigate this in the next step. Since climate is a substantial selection factor, we also checked for phenotypic variability between the most abundant PP-AA and PP-PP groups. To this end, we performed two-group comparisons of 516 phenotypic and climatic variables retrieved from the Arapheno database (Seren et al., 2017; Togninalli et al., 2020) and focused on those that significantly differed between both groups (Wilcoxon rank sum test with continuity correction, p value <0.05) (Supplemental Table S12). Notably, we observed differences in 88 climatic variables (Exposito-Alonso et al., 2019), especially maximal and minimal temperature conditions, precipitation and evapotranspiration (Figure 5A). Apart from the climate data, 40 diverse phenotypes varied significantly

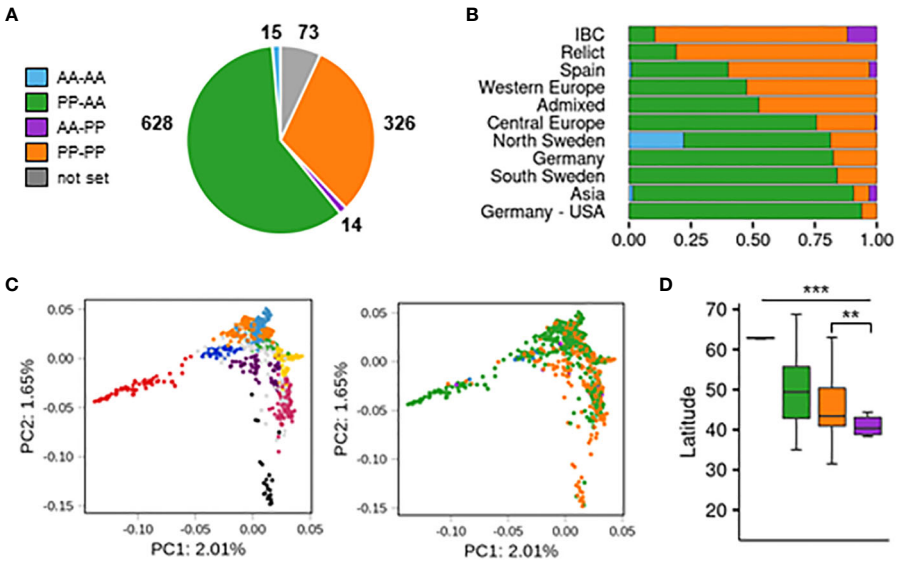


FIGURE 4
Population-scale diversity of *CYP705A2* and *BARS1* duplication status. **(A)** The sizes of four groups differing by the presence (PP)/absence (AA) of *CYP705A2-BARS1* and *CYP705A2-BARS2* gene pairs. **(B)** Group distribution among the genetic groups. U.S. accessions from the German group were separated from the remaining accessions. **(C)** Principal component analysis (PCA) plots, generated at linkage disequilibrium LD = 0.3. The first two components are presented. Accessions are colored according to their genetic group (left) or CYP-BARS status (right). U.S. accessions were not included in the analysis, in order to better visualize the remaining groups. PCA plots with other LD parameters are in Supplemental Figure S17. **(D)** Latitudes of accessions' sites of origin, grouped by CYP-BARS status. One-way rank-based analysis of variance ANOVA, p value<0.001, followed by Dunn's test with BH correction, $^{**}p$ value<0.05 (PP-PP vs AA-PP); $^{***}p$ value < 0.001 (all the other pairwise comparisons).

between both groups. Although some of these differences, e.g., flowering-related phenotypes, might be influenced by another genetic factor, independent from the arabidiol/baruol gene cluster structure (Li et al., 2010), we paid special attention to root growth-related phenotypes, since all Arabidopsis MGCs are considered to have root-specific expression (Huang et al., 2019). We observed significant differences between the PP-AA and PP-PP groups in root growth dynamics, which was analyzed during the first week after germination by Bouain et al. (2018). More specifically, the roots of PP-PP accessions elongated slower than those of PP-AA accessions (Figure 5B). Additionally, PP-PP accessions showed a significantly lower rate of root organogenesis from explants under one of three growth conditions tested in another study (Lardon et al., 2020) (Figure 5C). We next applied a linear mixed model in a genome-wide association study on the same 516 phenotypes to independently evaluate the significance of our observations after correction for the population structure and multiple testing. We used a genome-wide matrix of over 250 thousand biallelic SNPs supplemented with SNP-like encoded information about the gene duplication status (only PP-AA and PP-PP groups were analyzed). Although the association of *CYP705A2a* and *BARS2* presence/absence variation was not statistically significant for any variable we tested, we again obtained the lowest p values for the climatic data and root organogenesis phenotypes (Figure 5D, Supplemental Table S12). We then checked for the genetic interactions between the thalianol and arabidiol/baruol clusters to exclude the possibility that they affected our results, since the distribution of discontiguous and compact versions of the thalianol gene cluster was also strongly associated with latitude (Supplemental Figure S19). However, structural variation of the arabidiol/baruol gene cluster better explained the geographical distribution of the accessions. Moreover, variation in thalianol gene cluster organization did not affect the expression of the thalianol biosynthesis genes and had little impact on root growth phenotypic variation (Supplemental Figure S20).

In the reference accession Col-0, all genes in the arabidiol/baruol cluster were expressed at low levels and were active almost exclusively in roots (Supplemental Figure S21). In search of the possible links between arabidiol/baruol gene cluster structure and phenotypic variation, we investigated *CYP705A2*, *BARS1*, *CYP705A2a* and *BARS2* expression profiles in Col-0 and Cvi-0. We used RNA-Seq data from roots, shoots and leaves, which we retrieved from the studies where these accessions were grown in parallel under standard conditions (Kawakatsu et al., 2016; van Veen et al., 2016). We mapped the data to the respective (Col-0 or Cvi-0) annotated genome and compared the gene expression profiles (Figure 5E; Supplemental Table S13). In both accessions, the arabidiol/baruol gene cluster was silenced in shoots, except for the low activity of acyltransferase gene *AT4G15390*, detected in Cvi-0. Additionally, in both accessions, the clusters were active in roots, and the expression of *AT4G15390* was much stronger than that of the remaining genes. In Cvi-0, genes located in the genomic insertion (*CYP705A2a*, *BARS2* and *ATCVI-4G38100*, the latter encoding a protein with partial similarity to acyltransferase) were also expressed, although at a lower level, compared to the rest of the cluster. Surprisingly, in leaves of Cvi-0, but not Col-0, we also detected transcriptional activity within the arabidiol/baruol gene cluster. Most clustered genes were expressed at lower levels than in Cvi-0 roots, and the transcripts of *CYP705A2*, *CYP705A3* and *BARS1* were barely detectable. However, *ATCVI-4G38100*, *CYP705A2a* and *BARS2* had similar expression in leaves and roots. Taking these observations into account, it should not be

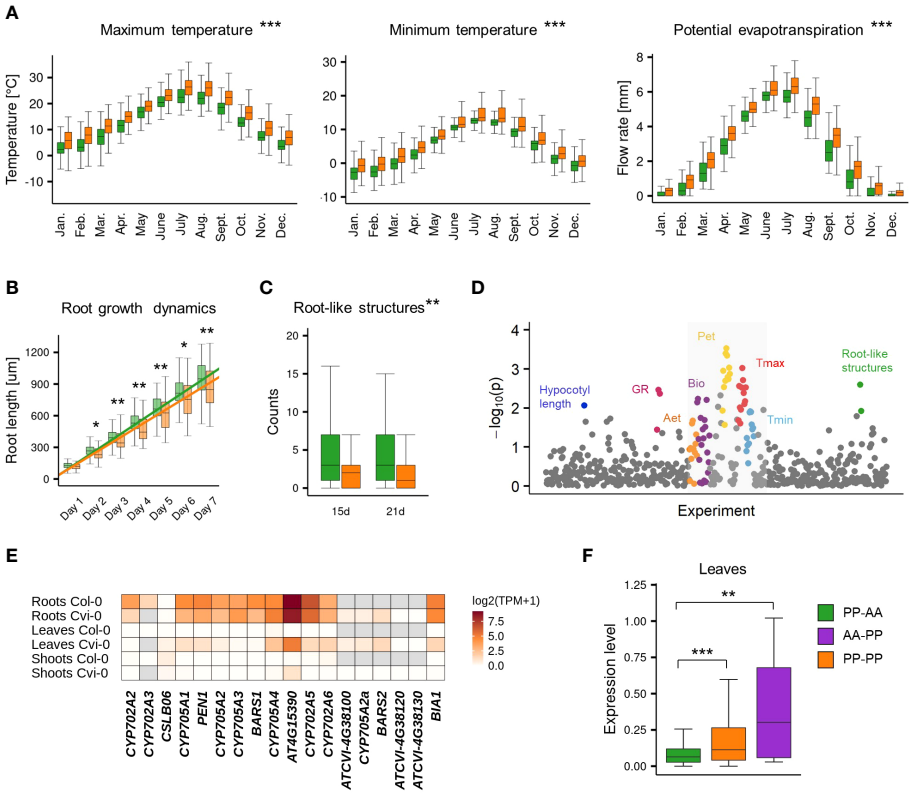


FIGURE 5
Phenotypic variation of PP-AA and PP-PP groups. **(A-C)** Two-group comparisons of climatic **(A)**, root growth dynamics **(B)** and root organogenesis **(C)** data between PP-AA (green) and PP-PP (orange) accessions. Stars denote the significance of Wilcoxon rank sum test with continuity correction, *p.value<0.1, **p.value<0.05, ***p.value<0.001. **(D)** Results of a genome-wide association study for PP-AA/PP-PP allelic variation. Study with climatic data is in the grey box **(E)** Tissue specificity of arabidiol/baruol gene cluster expression in Col-0 and Cvi-0. **(F)** Population-level differences in gene expression in leaves among the PP-AA, PP-PP and AA-PP groups. Expression levels are shown as $\log_2$(TPM+1). Stars denote the significance of one-way rank-based analysis of variance ANOVA, p.value<0.001, followed by Dunn's test with BH correction, **p.value<0.05, ***p.value<0.001. Source data are available in the Arapheno database (plots **A-C**), Supplemental Table S12 (plot **D**) and Supplemental Table S13 (plots **E-F**).

excluded that the metabolic products of arabidiol/baruol gene cluster activity in the roots and leaves of the Cvi-0 accession are not identical.

Since the PP-PP group represented a substantial fraction of the Arabidopsis population, we wanted to check whether the gene expression profile, which we observed in leaves of Cvi-0, was ubiquitous among the accessions from this group. To this end, we analyzed RNA-Seq data for 552 accessions mapped against the reference genome (Kawakatsu et al., 2016), and we compared the *BARS1* expression level between the AA-PP, PP-PP and PP-AA groups. It was significantly higher in accessions with the *CYP705A2a* + *BARS2* gene pair than in the PP-AA group (one-way rank-based analysis of variance, ANOVA, p value<0.001, followed by Dunn's test with BH correction, p value <0.05) (Figure 5F), in agreement with our predictions that *BARS2* was expressed in the leaves of these accessions and that reads from *BARS2* transcripts mapped to the *BARS1* locus, elevating its measured expression level. We also remapped the raw RNA-Seq reads from the Ty-1 and Cdm-1 accessions (PP-PP group), as well as from the Kn-0 and Sha (PP-AA group) accessions to their respective genomic assemblies and separately measured the expression levels of *BARS1* and *BARS2*. As expected, *BARS2* was expressed in the leaves of PP-PP accessions, while *BARS1* was not (Supplemental Figure S21).

Paired terpenoid synthase and cytochrome P450 genes are more variable than nonpaired genes

In many plant genomes, genes encoding terpenoid synthases (TSs, including the OSCs analyzed in our study) are positioned in the vicinity of CYPs more often than expected by chance. Therefore, they frequently exist as TS-CYP pairs (Boutanaev et al., 2015). TS-CYP pairs located in MGCs had similar (either high or low) copy number diversity and were frequently duplicated or deleted together. We wanted to check whether this observation could be extended to other TS-CYP pairs in the Arabidopsis genome. Therefore, we created a comprehensive list of 48 TSs and 242 CYPs based on trusted sources (Paquette et al., 2000; Bak et al., 2011; Nelson and Werck-Reichhart, 2011; Boutanaev et al., 2015). We then retrieved information about each gene's copy number diversity among 1,056 accessions (Supplemental Tables S14-S15). For 13 TSs, including *THAS1* and *BARS1*, we observed gains or losses in at least 1% of accessions. Only 33 CYPs showed such variability, and they represented three clans: CYP71 (26 variable genes out of 151), CYP85 (6 variable genes out of 29) and CYP72 (1 variable gene out of 19). The remaining clans showed very low variability. Next, for each TS, we selected all CYPs

within +/- 30-kb distance, which produced 38 pairs between 18 TSs and 27 CYPs, including pairs in thalianol, marneral, tirucalladienol and arabidiol/baruol gene clusters, as well as other putative secondary metabolism clusters, listed in the plantiSMASH resource (Kautsar et al., 2017). Subsequent group comparisons revealed that TSs and CYPs occurring in pairs were more variable than their nonpaired counterparts (Wilcoxon rank sum test with continuity correction, p value<0.01 for TSs, p value<0.001 for CYPs).

Discussion

According to our current understanding of the MGC formation phenomenon, nonrandom gene clustering in eukaryotes is linked with highly dynamic chromosomal regions. Numerous studies have highlighted that structural variations are the main genetic drivers of metabolic profile diversity and MGC evolution in plants (Fan et al., 2020; Li et al., 2020; Liu et al., 2020a; Liu et al., 2020b; Zhan et al., 2020; Katz et al., 2021). These studies suggested that plant MGCs are dynamically evolving and that the genetic mechanisms that originally led to their formation may be captured at the intraspecies genetic variation level. Similar conclusions were drawn from a previous study of the filamentous fungus *Aspergillus fumigatus*, in which secondary metabolic pathway genes were commonly organized into clusters (Lind et al., 2017). During evolution, new biochemical pathways are tuned and tested by many rounds of natural selection. The analysis of intraspecies MGC variants, which are more recent than the variants found in interspecies comparisons, may provide important insight into the formation of clustered gene architectures and plant metabolic diversity in a small evolutionary time frame. Accordingly, in our study we established that the mechanisms driving gene duplications and deletions contributed to the formation of Arabidopsis MGC in their present form and that they are still involved in shaping their structures. The dynamics of these mechanisms is e.g. marked by the observed extensive variation of the thalianol gene cluster and the arabidiol/baruol gene cluster.

The four MGCs in Arabidopsis are implicated in the biosynthesis of structurally diverse triterpenes and are dated after the α whole-genome duplication event, which occurred in the Brassicaceae lineage ~23-43 Mya (Field et al., 2011). These MGCs are assembled around the gene(s) encoding clade II OSCs. It has been shown that in various Brassicaceae genomes, clade II OSCs are often colocalized with genes from the CYP705, CYP708 and CYP702 clans and with genes from the acyltransferase IIIa subfamily (Liu et al., 2020b). Bioinformatic studies have also revealed that TSs and CYPs are paired in plant genomes more frequently than expected (Boutanaev et al, 2015). We found that in Arabidopsis, the physical proximity of CYPs and TSs was associated with increased CNV rates for these genes compared to the nonpaired ones. This might suggest that the occurrence of such a specific gene mix, combined with the structural instability of its genomic neighborhood, boosted the potential to produce novel metabolic pathways. The four Arabidopsis MGCs had different levels of variation (Figure 1), which generally reflected the phylogeny of clade II OSCs contained in these clusters (Figure 3C). Of them, MRN1 is most divergent in amino acid sequence. It is also mono-functional, i.e., catalyzes the formation of one specific product – marneral (Xiong et al., 2006). Functional studies have indicated a critical role of marneral synthase in Arabidopsis development (Go et al., 2012). Consistent with these findings, *MRN1* was the only clustered OSC gene, which was not affected by deletions or duplications, in any accession. Additionally, the neighboring CYPs were stable in copy number. Our results indicate that the marneral gene cluster is fixed in the Arabidopsis genome.

The arabidiol/baruol gene cluster was the most variable MGC. It comprises few gene subfamilies but is significantly expanded compared to the sister species *A. lyrata*, which is suggestive of recent duplications. For example, *PEN1* and *BARS1* have only one ortholog in *A. lyrata*, *LOC9306317*. Accordingly, we observed an exceptionally high rate of intraspecific gene gains and losses within this MGC. The segmentation of the arabidiol/baruol gene cluster into variable and invariable gene blocks may result from the ongoing process of selection-driven fixation of the arabidiol subcluster. The products of *PEN1* and *CYP705A1* are involved in the response to jasmonic acid treatment and infection with the root-rot pathogen *Pythium irregulare* (Sohrabi et al., 2015). Moreover, arabidiol may be further converted to arabidin in the pathway involving acyltransferase encoded by *AT5G47950*, which is located in the thalianol gene cluster (Huang et al., 2019) and which was also invariable in copy number in the present study. The fixation of genes involved in arabidin biosynthesis may indicate the biological significance of this pathway. CRISPR mutants with a disrupted *AT5G47950* gene has been shown to have significantly shorter roots than wild-type plants, and arabidin did not accumulate in these roots (Bai et al., 2021). Interestingly, *A. lyrata* is able to convert apo-arabidiol (the product of arabidiol degradation) into downstream compounds, despite the lack of arabidiol synthase (Sohrabi et al., 2017). This indicates that there may be modularity of the biosynthetic pathways in plants. This modularity might facilitate the assembly of a biosynthesis network and lead to an increase in the repertoire of secondary metabolites produced by the plant. Understanding the complexity of this network may be supported by in-depth analysis of MGC intraspecies variation.

The initial diversity of 2,3-oxidosqualene cyclization products generated by the plant is determined by OSC diversity. Here, we report the discovery of the *BARS2* gene, which was found in numerous accessions but was absent from Col-0; hence, it was absent from the reference genome (Figure 3A). Our data indicated that *BARS2* encodes a functional clade II OSC (Figures 3C, D). Notably, baruol synthase 1 encoded by its closest paralog, *BARS1*, has the lowest product specificity among plant OSCs (Lodeiro et al., 2007; Ghosh, 2016). Why some OSCs are highly multifunctional is not well understood. It has been suggested that they are undergoing evolution toward increased product specificity. It has been demonstrated that only two amino acid changes in cycloartenol synthase lead to its conversion into an accurate lanosterol synthase (Lodeiro et al., 2005). Biochemical characterization of baruol synthase 2 and its comparison with baruol synthase 1 may help reveal the role of particular amino acids in acquiring specificity for given products.

According to our data, the *BARS2* and *CYP705A2a* gene pair may be present in nearly one-third of the Arabidopsis population (Figure 4A), and their presence/absence variation is associated with the climatic gradient and root growth dynamics (Figures 5A-D). In Col-0, MGCs are embedded in local hotspots of three-dimensional chromatin interactions. Their activation in roots and repression in leaves is combined with the distinct chromatin condensation states

and nuclear repositioning of MGC regions between these tissues (Nützmann et al., 2020). Loss of the histone mark H3K27me3 in the *clf/swn* mutant resulted in the loss of interactive domains associated with the thalianol, marneral and arabidiol/baruol cluster regions, indicating that different transcriptional states of these MGCs are strictly regulated by the switches in their conformation. Curiously, in accessions with *CYP705A2a* and *BARS2*, we observed some transcriptional activity of arabidiol/baruol cluster genes in leaves (Figures 5E, F). The presence of an ~25-kb insertion in the arabidiol/baruol gene cluster may alter its structure and affect the epigenetic regulation of its activity. Thus, variation at the epigenetic and transcriptional level might lead to phenotypic differences, which could in turn contribute to local adaptation and eventually affect the global distribution of Arabidopsis accessions. However, additional studies are needed to assess whether the association between *BARS2* and *CYP705A2a* presence/absence variation and the global distribution of Arabidopsis accessions may be linked to the expression of these two genes or to the differences in transcriptional activity of the entire cluster (Wegel et al., 2009; Yu et al., 2016; Roulé et al., 2022).

The thalianol gene cluster was the second most variable MGC in our analysis. The first evidence for its structural diversity comes from the study of Liu et al. (2020a), who found large deletions affecting thalianol biosynthesis genes in ~2% of the studied accessions. Since our approach was specifically focused on CNV analysis and was duplication-aware, we were able to detect over two times more CNVs in a similar population (4.7%), with 49 accessions carrying gene deletions and five accessions with gene duplications (Figure 2A). Apart from the identification of two new variants – one large deletion and a duplication – we validated earlier assumptions that the nonreference compact version of the thalianol gene cluster is predominant in Arabidopsis (Figure 2B). Moreover, it is also better conserved than the discontiguous version (Figures 2D, E). It remains to be investigated whether tighter clustering of the thalianol gene cluster may be advantageous in certain environmental conditions or whether it is just less prone to structural variation due to physical constraints.

Triterpenes are high-molecular-weight nonvolatile compounds that are likely to act locally. However, they may be further processed and generate various breakdown products, both volatile and nonvolatile, which may be biologically active (Sohrabi et al., 2015; Sohrabi et al., 2017). Compounds of plant origin may also be metabolized by plant-associated microbiota. A recent study demonstrated that various combinations of thalianin, thalianyl fatty acid esters and arabidin attracted or repelled various microbial communities present in the soil and participated in the plant's active selection of root microbiota (Huang et al., 2019). In fact, a small but significant effect of Arabidopsis genotype on the root microbiome has been demonstrated previously (Bulgarelli et al., 2012; Lundberg et al., 2012). In a recent study by Karasov et al. (2022), bacterial communities that colonized the leaves of 267 local Arabidopsis populations, assessed at various localizations in Europe, formed two distinct groups strongly associated with the latititude. Specifically, a significant latitudinal cline was observed for the strains of the *Sphingomonas* genus, which is commonly associated with Arabidopsis (Bodenhausen et al., 2013). Various *Sphingomonas* species possess a range of biodegradative and biosynthetic capabilities (Mohn et al., 1999; Asaf et al., 2020). *Sphingomonas* is implicated in promoting Arabidopsis growth, increasing drought resistance and protecting plants against the leaf-pathogenic *Pseudomonas syringae* (Innerebner et al., 2011; Luo et al., 2019). Notably, in the study by Karasov et al. (2022), the host plant genotype alone could explain 52% to 68% of the observed variance in the phyllosphere microbiota. Moreover, the microbiome type was strongly associated with the dryness index of the local environment based on recent precipitation and temperature data. We propose that the genetic diversity of terpenoid metabolism pathways in Arabidopsis may be interdependent on the diversity of soil bacterial communities present in various environments, and this relationship might play a role in Arabidopsis adaptation to climate-driven selective pressures. Further exploration of MGC diversity may help us understand these biotic interactions.

Currently, the bioinformatic identification of new MGC candidates is mainly based on the combination of physical gene grouping and coexpression analyses. The accuracy and sensitivity of such approaches strongly depend on the abundance of data from various tissues, time points, and environmental conditions (Wisecaver et al., 2017). We suggest that the analysis of intraspecies genetic and transcriptomic variation may provide a valuable addition to MGC studies. The genome of one individual may not be representative enough to reveal the entire complexity of a given pathway, not to mention the metabolic diversity of the entire species (Kawakatsu et al., 2016; Shirai et al., 2017; Zmienko et al., 2020; Katz et al., 2021). With the rapid increase in the number of near-to-complete assemblies of individuals' genomes facilitated by the development of third-generation sequencing technologies, we are now entering the era of intense exploration of the impressive plasticity of plant metabolic pathways.

Materials and methods

Plant material and DNA samples

Arabidopsis seeds were obtained from The Nottingham Arabidopsis Stock Centre. The seeds were surface-sterilized, vernalized for 3 days, and grown on Jiffy pellets in ARASYSTEM containers (BETATECH) in a growth chamber (Percival Scientific). A light intensity of 175 mmol m^{-2} s^{-1} with proportional blue, red, and the far red light was provided by a combination of fluorescent lamps (Philips) and GroLEDs red/far red LED Strips (CLF PlantClimatics). Plants were grown for 3 weeks under a 16-h light (22°C)/8-h dark (18°C) cycle, at 70% RH, nourished with half-strength Murashige & Skoog medium (Serva). Genomic DNA for MLPA and ddPCR assays was extracted from 100 mg leaves with a DNeasy Plant Mini Kit (Qiagen), according to manufacturer's protocol, which included RNase A treatment step.

RD assays

To determine the boundaries of each MGC, the relevant literature and gene coexpression datasets were surveyed (Fazio et al., 2004; Xiong et al., 2006; Xiang et al., 2006; Lodeiro et al., 2007; Field and Osbourn, 2008; Morlacchi et al., 2009; Field et al., 2011; Go et al.,

2012; Thimmappa et al., 2014; Sohrabi et al., 2015; Yasumoto et al., 2016; Wisecaver et al., 2017). TAIR10 genome version and Araport 11 annotations (Cheng et al., 2017) were used as a reference in all analyses. Short read sequencing data from Arabidopsis 1001 Genomes Project (1001 Genomes Consortium, 2016) were downloaded from National Center for Biotechnology Information Sequence Read Archive repository (PRJNA273563), processed and mapped to the reference genome as described in (Zmienko et al., 2020). The gene copy number estimates based on read-depth analysis of short reads (RD dataset) were generated previously and are available at http://athcnv.ibch.poznan.pl. Accessions BRR57 (ID 504), KBS-Mac-68 (ID 1739), KBS-Mac-74 (ID 1741) and Ull2-5 (ID 6974), which we previously identified as harboring unusually high level of duplications, were removed from the analysis.

MLPA assays

MLPA probes were designed according to a procedure designed previously and presented in detail in (Samelak-Czajka et al., 2017). Probe genomic target coordinates are listed in Supplemental Table S16. The MLPA assays were performed using 5 ng of DNA template with the SALSA MLPA reagent kit FAM (MRC-Holland). The MLPA products were separated by capillary electrophoresis in an ABI Prism 3130XL analyzer at the Molecular Biology Techniques Facility in the Department of Biology at Adam Mickiewicz University, Poznan, Poland. Raw electropherograms were quality-checked and quantified with GeneMarker v.2.4.2 (SoftGenetics), with peak intensity and internal control probe normalization options enabled. Data were further processed in Excel (Microsoft). To allow easy comparison of RD and MLPA values, the MLPA results were normalized to a median of all samples' intensities and then multiplied by 2, separately for each gene/MLPA probe.

ddPCR assays

Genomic DNA samples were digested with XbaI (Promega). DNA template (2.5 ng) was mixed with 1× EvaGreen ddPCR Supermix (Bio-Rad), 200 nM gene-specific primers (Supplemental Table S17) and 70 µl of Droplet Generation Oil (Bio-Rad), then partitioned into approximately 18,000 droplets in a QX200 Droplet Generator (Bio-Rad), and amplified in a C1000 Touch Thermal Cycler (Bio-Rad), with the following cycling conditions: 1× (95°C for 5 min), 40× (95°C for 30 s, 57°C for 30 s, 72°C for 45 s), 1× (4°C for 5 min, 90°C for 5 min), with 2°C/s ramp rate. Immediately following end-point amplification, the fluorescence intensity of the individual droplets was measured using the QX200 Droplet Reader (Bio-Rad). Positive and negative droplet populations were automatically detected by QuantaSoft droplet reader software (Bio-Rad). For each accession and each gene, the template CNs [copies/µl PCR] were calculated using Poisson statistics, background-corrected based on the no-template control sample and normalized against the data for previously verified non-variable control gene *DCL1*.

PCR assays

Genomic DNA samples (5 ng) were used as templates in 20 µl reactions performed with PrimeSTAR GXL DNA Polymerase (TaKaRa), according to the manufacturer's instructions, in a three-step PCR. Amplicons (10 ul) were analyzed on 1% agarose with 1kb Gene Ruler DNA ladder (Fermentas). Primer sequences are listed in Supplemental Table S17. Primer pairs for *BARS1-BARS2* and *CYP705A2-CYP705A2a* were designed in corresponding genomic regions, that assured primer divergence between the paralogs. However, primers designed for *CYP705A2* produced unspecific bands of ~5kb in many samples. Therefore, this gene was excluded from the analysis.

Genotype assignments

For MLPA dataset, genotypes were assigned to each gene and each accession based on normalized MLPA values of ≤1 for LOSS genotype and >3 for GAIN genotype. The remaining cases were assigned REF genotype. For RD dataset, the respective RD thresholds were ≤1 for LOSS genotype and >3.4 for GAIN genotype, except for *BARS1*, for which both thresholds were lowered by 0.2. The remaining cases were assigned REF genotype. For ddPCR, genes with normalized CN=0 were assigned LOSS genotype and genes with normalized CN=2 were assigned REF genotype. The RD, MLPA and ddPCR datasets were then combined using the following procedure. For genes and accessions covered by multiple datasets, the final genotype was assigned based on all data. Discordant genotype assignments (21 out of 1,784 covered by multiple datasets) were manually investigated and 19 of them were resolved (Supplemental Figure S4; Supplemental Table S7). Out of the remaining 32,000, which were assayed with one method only, the genotype was manually corrected in 13 cases with values very close to the arbitrary threshold, based on population data distribution. Final genotype assignments for each gene and each accession are listed in Supplemental Table S6.

Sanger sequencing

The genomic DNA of Mir-0 accession (ID 8337) was used as a template (2 ng) for amplification using PrimeSTAR® GXL DNA Polymerase (TaKaRa), in a 40-µl PCR reaction with 0.3 µM primers OP009 and OP010, according to general manufacturer instructions. The amplified product, of ~8 kb in length, was purified with DNA Clean & Concentrator (ZYMO Research) and checked by gel electrophoresis and analysis on NanoDrop™ 2000 Spectrophotometer. The purified product (110 ng) was mixed with 1 ul of sequencing primer Mar02_R and sequenced on ABI Prism 3130XL analyzer at the Molecular Biology Techniques Facility in the Department of Biology at Adam Mickiewicz University, Poznan, Poland. Sequencing files were analyzed with Chromas Lite v. 2.6.6. (Technelysium) software.

De novo genomic assemblies generation, annotation and analysis

Mitterberg-2-185 and Dolna-1-40 genomic sequences were extracted, sequenced on 1 MinION flowcell (*Oxford Nanopore Technologies*) each and assembled *de novo* with Canu. Genomic sequences of interest (corresponding to thalianol gene cluster for Mitterberg-2-185 and tirucalladienol gene cluster for Dolna-1-40) were then retrieved with megablast (blast-2.10.0+ package) using TAIR10 reference genomic sequence as a query. The remaining *de novo* assemblies were retrieved from the following public databases. The PacBio-based genomic assemblies, gene annotations and orthogroups for An-1, C24, Cvi-0, Eri-1, Kyoto, Ler-0 and Sha accessions, as well as the reference genome coordinates of the hotspots of rearrangements, were downloaded from Arabidopsis 1001 Genomes Project Data Center (MPIPZJiao2020) or retrieved from the corresponding paper (Jiao and Schneeberger, 2020). Assembled genomic sequences of Ty-1 (PRJEB37258), Cdm-0 (PRJEB40125) and Kn-0 (PRJEB37260) accessions were retrieved from NCBI/Assembly database (Sayers et al., 2022). Gene prediction was performed with Augustus v.3.3.3 (Stanke and Morgenstern, 2005) with the following settings: "Species *Arabidopsis thaliana*", "both strands", "few alternative transcripts" or "none alternative transcripts", "predict only complete genes". These parameters were first optimized by gene prediction in the corresponding TAIR 10 genomic sequence and comparison with Araport 11 annotation. For previously annotated assemblies, we added information about the newly predicted genes to existing annotations. The protein sequences of *de novo* predicted genes and the information about their best blast hit in the reference genome are available in Supplemental Information. The search for conserved domain organization was performed with the online NCBI search tool against Pfam v.33.1 databases. Protein sequence alignment was done with Multalin or EMBL online tools (Corpet, 1988; Madeira et al., 2019). TEs were annotated with RepeatMasker software version 4.1.2 (http://www.repeatmasker.org), using homology-based method with TAIR10-transposable-elements reference library.

Identification of chromosomal inversions

The BreakDancerMax program from the BreakDancer package v.1.3.6 (Chen et al., 2009) was used to detect inversions in each of 997 samples with paired-end data and unimodal insert size distribution. Variants were called separately for each accession and each chromosome. Only calls with lengths within the range 0.5 kbp – 50 kbp and with the Confidence Score >35 were retained. Since BreakDancerMax output included numerous overlapping calls for individual accessions, we first minimized its redundancy. From the overlapping regions, we kept one variant with i) the highest Confidence Score, and ii) the highest number of supporting reads. If two or more overlapping variants had the same score and the number of supporting reads number, maximized coordinates of these variants were used. This step was carried out in two iterations, considering the 50% reciprocal overlap of the variants. Then, the inversions that overlapped with the thalianol gene cluster were selected from each genome-wide dataset.

SNP calling at *CYP705A2* and *BARS1* genes

Variants (SNPs and short indels) were called with DeepVariant v.1.3.0 in WGS mode and merged with GLnexus (Yun et al., 2021). Analysis was performed for *CYP705A2* and *BARS1* genomic loci. The results were further filtered to include only biallelic variants, that were located in the exons of each gene (for *BARS1*, exon intersections from two transcript models were used). The number of heterozygous positions was then calculated for each accession and each gene. The same procedure was repeated by taking into account only biallelic variants with at least 1% frequency, which resulted in nearly identical results. Both types of analysis led to the selection of the same set of accessions with duplication at both loci.

Genome-wide SNP analysis

Variants for 983 accessions with known *CYP705A2* + *BARS1* and *CYP705A2a* + *BARS2* pair status were downloaded from the 1001 Genomes Project Data Center (1001genomes_snp-short-indel_only_ACGTN_v3.1.vcf.snpeff file) (1001 Genomes Consortium, 2016). Data preprocessing was performed using PLINK v.1.90b3w (https://www.cog-genomics.org/plink/1.9/; Chang et al., 2015). Variants with missing call rates exceeding value 0.5 and variants with minor allele frequency below 3% were filtered out. The LD parameter for linkage disequilibrium-based filtration was set as follows: indep-pairwise 200'kb' 25 0.3. For PCA analysis with EIGENSOFT v.7.2.1 (Price et al., 2006; Patterson et al., 2006) at least 130,000 SNPs were used. PCA for a wide LD range between 0.3 - 0.9 was then calculated in a similar manner. U.S.A accessions which only recently separated geographically from the rest of the population (Lee et al., 2017) were excluded, to ensure better visibility of the remaining accessions. The ggplot2 package was used for data visualization in R v4.0.4 (https://www.r-project.org; Wickham, 2016).

Genome-wide association study and phenotype analysis

The entire set of 516 phenotypes from 26 studies was downloaded from the Arapheno database on 26 April 2022 (Seren et al., 2017; Togninalli et al., 2020). The above genome-wide SNP dataset, to which we added a biallelic variant representing PP-AA or PP-PP group assignment, was used. The IBS kinship matrix was calculated on 954 accessions. Association analysis was performed for each phenotype using a mixed model correcting for population structure using Efficient Mixed-Model Association eXpedited, version emmax-beta-07Mar2010 (Kang et al., 2008). Input file generation and analysis of the results were performed with PLINK v.1.90b3w and R v4.0.4.

Analysis of RNA-Seq data

Processed RNA-seq data from leaves for 728 accessions (552 in common with our study) mapped to the reference transcriptome (Kawakatsu et al., 2016) were downloaded from NCBI/SRA (PRJNA319904), normalized and used to compare *BARS1* expression

levels between PP-AA, PP-PP and AA-PP groups. Additionally, raw RNA-Seq reads from leaves were downloaded from the same source for accessions-specific mapping and analysis of Cdm-0, Col-0, Cvi-0, Kn-0, Ty-1 and Sha accessions. Raw RNA-Seq reads from roots and shoots of Col-0 and Cvi-0 accessions were retrieved from BioProject PRJEB14092 (van Veen et al., 2016). SRA Toolkit v2.8.2. (https://github.com/ncbi/sra-tools) and FastQC v0.11.4 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used for downloading the raw reads and for the quality analysis. For Cdm-0, Kn-0 and Ty-1 genomes.gtf files were generated based on Augustus results, that included the annotations for the genes of interest (provided as Supplemental Information). Raw reads were mapped to the respective genomes using the STAR aligner version 2.7.8a (Dobin et al., 2013). STAR indices were generated with parameters: "–runThreadN 24 –sjdbOverhang 99 –genomeSAindexNbases 12". The following parameters were used for the mapping step: "–runThreadN 24 –quantMode GeneCounts –outFilterMultimapNmax 1 –outSAMtype BAM SortedByCoordinate –outSAMunmapped Within". Bioinfokit v1.0.8 https://zenodo.org/record/3964972#.Yyw6oRzP1hE) was used to convert.gff3 to.gtf files. Transcripts per million (TPM) values and fragments per kilobase exon per million reads (FPKM) with total exon length for each gene were computed in R v4.0.4.

Analysis of TS-CYP pairs

A list of Arabidopsis CYP genes was created by collecting information from previous studies and acknowledged website resources (Arabidopsis Cytochromes P450; Paquette et al., 2000; Ehlting et al., 2008; Nelson, 2009; Bak et al., 2011; Nelson and Werck-Reichhart, 2011; Boutanaev et al., 2015) (http://www.p450.kvl.dk/p450.shtml). Genes marked in Araport 11 as pseudogenes were excluded from the further analysis. Genes were assigned to clans and families according to the information from the above resources. A list of TS genes was created based on a previous study (Boutanaev et al., 2015) and restricted to genes with valid Araport 11 locus. Genotypes were assigned based on criteria defined for RD dataset: (CN =< 1 as losses, CN >=3.4 as gains, the remaining genotypes were classified as unchanged). Genes from thalianol, tirucalladienol, arabidiol/baruol and marneral gene clusters were already genotyped. Gene coordinates were downloaded from Araport 11. All CYP genes positioned at a distance +/- 30 kb from TS gene borders were classified as paired with a given TS gene. Information about predicted secondary metabolism clusters was retrieved from plantiSMASH resource (Kautsar et al., 2017).

Prediction and analysis of BARS1 and BARS2 3D protein structures

The three-dimensional structures of the reference baruol synthase 1 proteins NP_193272.1, NP_001329547.1, as well as Cvi-0 proteins encoded by *ATCVI-4G38020* (*BARS1*) and *ATCVI-4G38110* (*BARS2*), were predicted from their amino acid sequences using the AlphaFold2 code through the ColabFold software (Jumper et al., 2021; Mirdita et al., 2022). The modeling studies were performed for a single amino acid chain. A crystal structure of human OSC in a complex with lanosterol (ID 1W6K) was retrieved from the Protein Data Bank (Thoma et al., 2004; Berman et al., 2007). The SSM algorithm implemented in COOT was used for superpositions of protein models (Krissinel and Henrick, 2004; Emsley et al., 2010) (Supplemental Information).

Data availability statement

Publicly available datasets were analyzed in this study. Sequence data can be found at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA273563/, https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31147/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37258/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB40125/; https://www.ncbi.nlm.nih.gov/bioproject/PRJEB37260/; https://www.ncbi.nlm.nih.gov/bioproject/PRJNA319904/; and https://www.ncbi.nlm.nih.gov/bioproject/PRJEB14092/). Genomic variants can be found in the 1,001 Genomes Project resources (https://1001genomes.org/data/GMI-MPI/releases/v3.1/1001genomes_snpeff_v3.1/). All phenotyping data and the associated metadata can be found in the AraPheno database (https://arapheno.1001genomes.org/static/database.zip). Individual phenotypes with their DOI identifiers can be additionally accessed and downloaded from https://arapheno.1001genomes.org/phenotypes/. The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization: AZ. Methodology: MM-Z, PW, and AZ. Investigation: MM-Z, AS, PW, PS, KB, and TI. Software: MM-Z, AS, PW, and MZ. Visualization: MM-Z, KB, and AZ. Formal analysis: MM-Z. Writing – original draft: MM-Z, and AZ. Writing – review and editing: MM-Z, KB, MF, MZ, and AZ. Supervision: MF, and AZ. Project administration: AZ. Funding acquisition: MF, and AZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Science Centre (Poland) grants 2014/13/B/NZ2/03837 to MF and 2017/26/D/NZ2/01079 to AZ. TI obtained funding from the support program for Ukrainian researchers under the Agreement between the Polish Academy of Sciences and the U.S. National Academy of Sciences. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgments

We thank Piotr Kozłowski for fruitful discussions and comments on the manuscript. Computations were supported in part by PLGrid Infrastructure.

Conflict of interest

The 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1104303/full#supplementary-material

SUPPLEMENTAL FILE 1
Supplemental Tables S1-S17.

SUPPLEMENTAL FILE 2
Supplemental information and Supplemental Figures S1-S21.

SUPPLEMENTARY DATA SHEET 15
Superposed 3D models of BARS1, BARS2 and human oxidosqualene cyclase proteins.

References

1001 Genomes Consortium (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 1–11. doi: 10.1016/j.cell.2016.05.063

Asaf, S., Numan, M., Khan, A. L., and Al-Harrasi, A. (2020). *Sphingomonas*: from diversity and genomics to functional role in environmental remediation and plant growth. *Crit. Rev. Biotechnol.* 40, 138–152. doi: 10.1080/07388551.2019.1709793

Bai, Y., Fernández-Calvo, P., Ritter, A., Huang, A. C., Morales-Herrera, S., Bicalho, K. U., et al. (2021). Modulation of arabidopsis root growth by specialized triterpenes. *New Phytol.* 230, 228–243. doi: 10.1111/nph.17144

Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., et al. (2011). Cytochromes p450. *Arabiopsis Book* 9, e0144. doi: 10.1199/tab.0144

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 35, D301–D303. doi: 10.1093/nar/gkl971

Bodenhausen, N., Horton, M. W., and Bergelson, J. (2013). Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PloS One* 8, e56329. doi: 10.1371/journal.pone.0056329

Bouain, N., Satbhai, S. B., Korte, A., Saenchai, C., Desbrosses, G., Berthomieu, P., et al. (2018). Natural allelic variation of the AZI1 gene controls root growth under zinc-limiting condition. *PloS Genet.* 14, e1007304. doi: 10.1371/journal.pgen.1007304

Boutanaev, A. M., Moses, T., Zi, J., Nelson, D. R., Mugford, S. T., Peters, R. J., et al. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.* 112, E81–E88. doi: 10.1073/pnas.1419547112

Boycheva, S., Daviet, L., Wolfender, J. L., and Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends Plant Sci.* 19, 447–459. doi: 10.1016/j.tplants.2014.01.013

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., et al. (2012). Revealing structure and assembly cues for arabidopsis root-inhabiting bacterial microbiota. *Nature* 488, 91–95. doi: 10.1038/nature11336

Castillo, D. A., Kolesnikova, M. D., and Matsuda, S. P. (2013). An effective strategy for exploring unknown metabolic pathways by genome mining. *J. Am. Chem. Soc* 135, 5885–5894. doi: 10.1021/ja401535g

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., and Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J.* 89, 789–804. doi: 10.1111/tpj.13415

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–681. doi: 10.1038/nmeth.1363

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16, 10881–10890. doi: 10.1093/nar/16.22.10881

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Ehlting, J., Sauveplane, V., Olry, A., Ginglinger, J. F., Provart, N. J., and Werck-Reichhart, D. (2008). An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in *Arabidopsis thaliana*. *BMC Plant Biol.* 8, 47. doi: 10.1186/1471-2229-8-47

Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010). Features and development of coot. *Acta Crystallogr. D. Biol. Crystallogr.* 66, 486–501. doi: 10.1107/S0907444910007493

Erb, M., and Kliebenstein, D. J. (2020). Plant secondary metabolites as defenses, regulators, and primary metabolites: The blurred functional trichotomy. *Plant Physiol.* 184, 39–52. doi: 10.1104/pp.20.00433

Exposito-Alonso, M.500 Genomes Field Experiment Team, , Burbano, H. A., Bossdorf, O., Nielsen, R., Weigel, D. (2019). Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature* 573, 126–129. doi: 10.1038/s41586-019-1520-9

Fan, P., Wang, P., Lou, Y. R., Leong, B. J., Moore, B. M., Schenck, C. A., et al. (2020). Evolution of a plant gene cluster in *Solanaceae* and emergence of metabolic diversity. *Elife* 9, e56717. doi: 10.7554/eLife.56717.sa2

Fazio, G. C., Xu, R., and Matsuda, S. P. T. (2004). Genome mining to identify new plant triterpenoids. *J. Am. Chem. Soc* 126, 5678–5679. doi: 10.1021/ja0318784

Field, B., Fiston-Lavier, A. S., Kemen, A., Geisler, K., Quesneville, H., and Osbourn, A. E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci.* 108, 16116–16121. doi: 10.1073/pnas.1109273108

Field, B., and Osbourn, A. E. (2008). Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science* 320, 543–547. doi: 10.1126/science.1154990

Ghosh, S. (2016). Biosynthesis of structurally diverse triterpenes in plants: the role of oxidosqualene cyclase. *Proc. Indian Natl. Sci. Acad.* 82, 1189–1210. doi: 10.16943/ptinsa/2016/48578

Go, Y. S., Lee, S. B., Kim, H. J., Kim, J., Park, H. Y., Kim, J. K., et al. (2012). Identification of marneral synthase, which is critical for growth and development in arabidopsis. *Plant J.* 72, 791–804. doi: 10.1111/j.1365-313X.2012.05120.x

Huang, A. C., Jiang, T., Liu, Y. X., Bai, Y. C. Y., Reed, J., Qu, B., et al. (2019). A specialized metabolic network selectively modulates arabidopsis root microbiota. *Science* 364, eaau6389. doi: 10.1126/science.aau6389

Innerebner, G., Knief, C., and Vorholt, J. A. (2011). Protection of *Arabidopsis thaliana* against leaf-pathogenic *Pseudomonas syringae* by *Sphingomonas* strains in a controlled model system. *Appl. Environ. Microbiol.* 77, 3202–3210. doi: 10.1128/AEM.00133-11

Isah, T. (2019). Stress and defense responses in plant secondary metabolites production. *Biol. Res.* 52, 39. doi: 10.1186/s40659-019-0246-3

Jiao, W. B., and Schneeberger, K. (2020). Chromosome-level assemblies of multiple arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11, 989. doi: 10.1038/s41467-020-14779-y

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi: 10.1038/s41586-021-03819-2

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Karasov, T. L., Neumann, M., Shirsekar, G., Monroe, G.PATHODOPSIS Team, , Weigel, D., et al. (2022) (Accessed November 2, 2022).

Katz, E., Li, J. J., Jaegle, B., Ashkenazy, H., Abrahams, S. R., Bagaza, C., et al. (2021). Genetic variation, environment and demography intersect to shape arabidopsis defense metabolite variation across Europe. *Elife* 10, e67784. doi: 10.7554/eLife.67784.sa2

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., and Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* 45, W55–W63. doi: 10.1093/nar/gkx305

Kawakatsu, T., Huang, S. S. C., Jupe, F., Sasaki, E., Schmitz, R. J., Urich, M. A., et al. (2016). Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 166, 492–505. doi: 10.1016/j.cell.2016.06.044

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D. Biol. Crystallogr.* 60, 2256–2268. doi: 10.1107/S0907444904026460

Lardon, R., Wijnker, E., Keurentjes, J., and Geelen, D. (2020). The genetic framework of shoot regeneration in arabidopsis comprises master regulators and conditional fine-tuning factors. *Commun. Biol.* 3, 549. doi: 10.1038/s42003-020-01274-9

Lee, C. R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., et al. (2017). On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nat. Commun.* 8, 14458. doi: 10.1038/ncomms14458

Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 107, 21199–21204. doi: 10.1073/pnas.1007431107

Lind, A. L., Wisecaver, J. H., Lameirasm, C., Wiemann, P., Palmer, J. M., Keller, N. P., et al. (2017). Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PloS Biol.* 15, e2003583. doi: 10.1371/journal.pbio.2003583

Li, Q., Ramasamy, S., Singh, P., Hagel, J. M., Dunemann, S. M., Chen, X., et al. (2020). Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. *Nat. Commun.* 11, 1190. doi: 10.1038/s41467-020-15040-2

Liu, Z., Cheema, J., Vigouroux, M., Hill, L., Reed, J., Paajanen, P., et al. (2020a). Formation and diversification of a paradigm biosynthetic gene cluster in plants. *Nat. Commun.* 11, 5354. doi: 10.1038/s41467-020-19153-6

Liu, Z., Suarez Duran, H. G., Harnvanichvech, Y., Stephenson, M. J., Schranz, M. E., Nelson, D., et al. (2020b). Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the brassicaceae. *New Phytol.* 227, 1109–1123. doi: 10.1111/nph.16338

Lodeiro, S., Schulz-Gasch, T., and Matsuda, S. P. T. (2005). Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. *J. Am. Chem. Soc* 127, 14132–14133. doi: 10.1021/ja053791j

Lodeiro, S., Xiong, Q., Wilson, W. K., Kolesnikova, M. D., Onak, C. S., and Matsuda, S. P. T. (2007). An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis. *J. Am. Chem. Soc* 129, 11213–11222. doi: 10.1021/ja073133u

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488, 86–90. doi: 10.1038/nature11237

Luo, Y., Wang, F., Huang, Y., Zhou, M., Gao, J., Yan, T., et al. (2019). *Sphingomonas* sp. Cra20 increases plant growth rate and alters rhizosphere microbial community structure of *Arabidopsis thaliana* under drought stress. *Front. Microbiol.* 10, 1221. doi: 10.3389/fmicb.2019.01221

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nat. Methods* 19, 679–682. doi: 10.1038/s41592-022-01488-1

Mohn, W. W., Yu, Z., Moore, E. R., and Muttray, A. F. (1999). Lessons learned from *Sphingomonas* species that degrade abietane triterpenoids. *J. Ind. Microbiol. Biotechnol.* 23, 374–379. doi: 10.1038/sj.jim.2900731

Morlacchi, P., Wilson, W. K., Xiong, Q., Bhaduri, A., Sttivend, D., Kolesnikova, M. D., et al. (2009). Product profile of PEN3: The last unexamined oxidosqualene cyclase in *Arabidopsis thaliana*. *Org. Lett.* 11, 2627–2630. doi: 10.1021/ol9005745

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65. doi: 10.1186/1479-7364-4-1-59

Nelson, D., and Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *Plant J.* 66, 194–211. doi: 10.1111/j.1365-313X.2011.04529.x

Nützmann, H. W., Doerr, D., Ramírez-Colmenero, A., Sotelo-Fonseca, J. E., Wegel, E., Di Stefano, M., et al. (2020). Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc. Natl. Acad. Sci. U. S. A.* 117, 13800–13809. doi: 10.1073/pnas.1920474117

Nützmann, H. W., and Osbourn, A. (2014). Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.* 26, 91–99. doi: 10.1016/j.copbio.2013.10.009

Nützmann, H. W., Scazzocchio, C., and Osbourn, A. (2018). Metabolic gene clusters in eukaryotes. *Annu. Rev. Genet.* 52, 159–183. doi: 10.1146/annurev-genet-120417-031237

Paquette, S. M., Bak, S., and Feyereisen, R. (2000). Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* 19, 307–317. doi: 10.1089/10445490050021221

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PloS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Roulé, T., Christ, A., Hussain, N., Huang, Y., Hartmann, C., Benhamed, M., et al. (2022). The lncRNA MARS modulates the epigenetic reprogramming of the marneral cluster in response. *Mol. Plant* 15, 840–856. doi: 10.1016/j.molp.2022.02.007

Samelak-Czajka, A., Marszalek-Zenczak, M., Marcinkowska-Swojak, M., Kozlowski, P., Figlerowicz, M., and Zmienko, A. (2017). MLPA-based analysis of copy number variation in plant populations. *Front. Plant Sci.* 8, 222. doi: 10.3389/fpls.2017.00222

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Seren, Ü, Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., et al. (2017). AraPheno: a public database for *Arabidopsis thaliana* phenotypes. *Nucleic Acids Res.* 45, D1054–D1059. doi: 10.1093/nar/gkw986

Shirai, K., Matsuda, F., Nakabayashi, R., Okamoto, M., Tanaka, M., Fujimoto, A., et al. (2017). A highly specific genome-wide association study integrated with transcriptome data reveals the contribution of copy number variations to specialized metabolites in *Arabidopsis thaliana* accessions. *Mol. Biol. Evol.* 34, 3111–3122. doi: 10.1093/molbev/msx234

Sohrabi, R., Ali, T., Harinantenaina Rakotondraibe, L., and Tholl, D. (2017). Formation and exudation of non-volatile products of the arabidiol triterpenoid degradation pathway in arabidopsis roots. *Plant Signal. Behav.* 12, e1265722. doi: 10.1080/15592324.2016.1265722

Sohrabi, R., Huh, J. H., Badieyan, S., Rakotondraibe, L. H., Kliebenstein, D. J., Sobrado, P., et al. (2015). In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT *via* triterpene degradation in arabidopsis roots. *Plant Cell* 27, 874–890. doi: 10.1105/tpc.114.132209

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Thimmappa, R., Geisler, K., Louveau, T., O'Maille, P., and Osbourn, A. (2014). Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.* 65, 225–257. doi: 10.1146/annurev-arplant-050312-120229

Thoma, R., Schulz-Gasch, T., D'Arcy, B., Benz, J., Aebi, J., Dehmlow, H., et al. (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* 432, 118–122. doi: 10.1038/nature02993

Togninalli, M., Seren, Ü, Freudenthal, J. A., Monroe, J. G., Meng, D., Nordborg, M., et al. (2020). AraPheno and the AraGWAS catalog 2020: a major database update including RNA-seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids Res.* 48, D1063–D1068. doi: 10.1093/nar/gkz925

van Veen, H., Vashisht, D., Akman, M., Girke, T., Mustroph, A., Reinen, E., et al. (2016). Transcriptomes of eight *Arabidopsis thaliana* accessions reveal core conserved, genotype- and organ-specific responses to flooding stress. *Plant Physiol.* 172, 668–689. doi: 10.1104/pp.16.00472

Wada, M., Takahashi, H., Altaf-Ul-Amin, M., Nakamura, K., Hirai, M. Y., Ohta, D., et al. (2012). Prediction of operon-like gene clusters in the *Arabidopsis thaliana* genome based on co-expression analysis of neighboring genes. *Gene* 503, 56–64. doi: 10.1016/j.gene.2012.04.043

Wegel, E., Koumproglou, R., Shaw, P., and Osbourn, A. (2009). Cell type-specific chromatin decondensation of a metabolic gene cluster in oats. *Plant Cell* 21, 3926–3936. doi: 10.1105/tpc.109.072124

Wickham, H. (2016). *ggplot2* (Springer Cham). 2nd ed. doi: 10.1007/978-3-319-24277-4

Wisecaver, J. H., Borowsky, A. T., Tzin, V., Jander, G., Kliebenstein, D. J., and Rokas, A. (2017). A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell* 29, 944–959. doi: 10.1105/tpc.17.00009

Xiang, T., Shibuya, M., Katsube, Y., Tsutsumi, T., Otsuka, M., Zhang, H., et al. (2006). A new triterpene synthase from *Arabidopsis thaliana* produces a tricyclic triterpene with two hydroxyl groups. *Org. Lett.* 8, 2835–2838. doi: 10.1021/ol060973p

Xiong, Q., Wilson, W. K., and Matsuda, S. P. T. (2006). An arabidopsis oxidosqualene cyclase catalyzes iridal skeleton formation by grob fragmentation. *Angew. Chem. Int. Ed. Engl.* 45, 1285–1288. doi: 10.1002/anie.200503420

Yasumoto, S., Fukushima, E. O., Seki, H., and Muranaka, T. (2016). Novel triterpene oxidizing activity of *Arabidopsis thaliana* CYP716A subfamily enzymes. *FEBS Lett.* 590, 533–540. doi: 10.1002/1873-3468.12074

Yun, T., Li, H., Chang, P. C., Lin, M. F., Carroll, A., and McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* 36, 5582–5589. doi: 10.1093/bioinformatics/btaa1081

Yu, N., Nützmann, H. W., MacDonald, J. T., Moore, B., Field, B., Berriri, S., et al. (2016). Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.* 44, 2255–2265. doi: 10.1093/nar/gkw100

Zhan, C., Lei, L., Liu, Z., Zhou, S., Yang, C., Zhu, X., et al. (2020). Selection of a subspecies-specific diterpene gene cluster implicated in rice disease resistance. *Nat. Plants* 6, 1447–1454. doi: 10.1038/s41477-020-00816-7

Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozlowski, P., et al. (2020). AthCNV: A map of DNA copy number variations in the arabidopsis genome. *Plant Cell* 32, 1797–1819. doi: 10.1105/tpc.19.00640

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 February 2023
DOI 10.3389/fpls.2022.1095644

OPEN ACCESS

EDITED BY
Pengxiang Fan,
Zhejiang University, China

REVIEWED BY
Lianxuan Shi,
Northeast Normal University, China
Simona Proietti,
Research Institute on Terrestrial Ecosystems (CNR), Italy

*CORRESPONDENCE
Changming Wang
wcm@swfu.edu.cn
Rui Shi
shirui@swfu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 11 November 2022
ACCEPTED 30 December 2022
PUBLISHED 01 February 2023

CITATION
Wang Y, Xia J, Wang Z, Ying Z, Xiong Z, Wang C and Shi R (2023) Combined analysis of multi-omics reveals the potential mechanism of flower color and aroma formation in *Macadamia integrifolia*. *Front. Plant Sci.* 13:1095644. doi: 10.3389/fpls.2022.1095644

Combined analysis of multi-omics reveals the potential mechanism of flower color and aroma formation in *Macadamia integrifolia*

Yonggui Wang[1†], Jing Xia[1†], Zile Wang[2†], Zhiping Ying[1], Zhi Xiong[1], Changming Wang[1*] and Rui Shi[1*]

[1]Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, International Ecological Forestry Research Center of Kunming, Southwest Forestry University, Kunming, China, [2]Yunnan Agricultural University College of Plant Protection, Kunming, China

Introduction: *Macadamia integrifolia* Maiden & Betche is a domesticated high-value nut crop. The development of nut flower affects the fruit setting rate, yield and quality of nuts. Therefore, in this experiment, two varieties with different flower color, flowering time, flowering quantity and nut yield (single fruit weight) were selected as the research objects.

Methods: Transcriptome (RNA-Seq) and metabolome (LC-MS/MS, GC-MS) analyses were performed to study the regulatory mechanisms of nut flower development, color and aroma.

Results: The results indicated that plant hormone signal transduction, starch sucrose metabolism, phenylpropanoid metabolism, flavonoid biosynthesis, and anthocyanin biosynthesis pathways were related to nut flower development and flower color formation. In the early stage of flowering, most of the differentially expressed genes (DEGs) are involved in the IAA signal transduction pathway, while in the later stage, the brassinolide signal pathway is mainly involved. In starch and sugar metabolism, DEGs are mainly involved in regulating and hydrolyzing stored starch into small molecular sugars in flower tissues. In the phenylpropanoid biosynthesis pathway, DEGs are mainly related to the color and aroma (volatile organic compounds, VOCs) formation of nut flowers. Four color formation metabolites (anthocyanins) in nut flowers were found by LC-MS/MS detection. In addition, the VOCs showed no significant difference between red nut flowers (R) and white nut flowers (W), which was mainly reflected in the aroma formation stage (flowering time). And 12 common differentially accumulation metabolites (DAMs) were detected by GC-MS and LC-MS/MS. At the same time, the DEGs, *AAT*, *LOX* and *PAL* genes, were also identified to regulate key metabolite synthesis during nut flower development. These genes were further verified by qRT-PCR.

Conclusion: Our results provide insights to clarify the molecular mechanism of color and aroma formation during *M. integrifolia* flower development that pave the way for nut quality and yield breeding.

KEYWORDS

Macadamia integrifolia, LC-MS/MS, flower color, transcriptome, GC-MS, aroma

1 Introduction

Macadamia integrifolia Maiden & Betche is a nut crop with high nutritional value and healthcare functions. Macadamia nuts are rich in unsaturated fatty acids, such as palmitoleic acid (POA). The proportion of unsaturated fatty acids is 85.74%, with oleic acid accounting for 58.60%, arachidonic acid accounting for 15.99%, and POA accounting for 11.15% (Du et al., 2010). It is also rich in protein and vitamins (such as vitamins B1 and B2 and nicotinic acid), which contain about 9% protein and are made up of 18 amino acids. Among the 18 amino acids, glutamic acid and arginine account for a high proportion, including eight essential amino acids, accounting for 28.88% of the total amino acids (Du et al., 2010). Long-term consumption of macadamia nuts can alleviate the occurrence of heart disease (Duxbury, 1995; Garg et al., 2003). Macadamia nut is listed as the most expensive nut because of its higher economic benefits and is considered as the queen of nuts (Stephensonra and Macadamia, 1994). In 2020, China's *M. integrifolia* planting area of 30 million square meters, ranking first in the world, ranks second in output worldwide. However, the fruit setting rate of nuts is low, generally only 0.1%–0.3% (Urata, 1954; Trueman et al., 1994; Olesen et al., 2011; Mcfadyen et al., 2012), affecting the effective market supply. At present, reports on increasing the yield of *M. integrifolia* are mainly on disease control (Li et al., 2022; Trueman et al., 2022; Qi et al., 2022) and crop cultivation (A'Ida et al., 2021; Russell et al., 2018). In recent years, the research on the *M. integrifolia* breeding system has increased, but there is no report on the molecular mechanism of nut flower development possibly affecting the fruit quality and yield. A low fruit setting rate has become a key factor restricting the yield and industrial development of *M. integrifolia*. There is an urgent need to provide scientific and technical support to solve the problem of low fruit setting rates (Maguire et al., 2004; Herbert et al., 2019; Kämper et al., 2019).

The formation of plants' flower fragrance volatiles is catalyzed by related enzymes, and there are some differences in the effects of different types of compounds. A previous study found that terpenes and phenylpropanes are the main substances that emit signals for attracting pollinators (Schiestl, 2010). They act as the primary medium between plants and pollinators. Aliphatic compounds, on the other hand, play a defensive role between plants and herbivores. The regulatory genes of the phenylpropanoid pathway are mostly from the MYB family, which is involved in the regulation of the secondary metabolism of flowers and the synthesis of various aroma types (Schiestl, 2010; Xie et al., 2016). *Petunia axillaris* is one of the model plants used to study the flower aroma of plants, and phenylpropanoid compounds are the main source of the flower fragrance of *P. axillaris*. *PhODO1* is the first transcription factor found to promote phenylpropanoid biosynthesis during the odor formation of *P. axillaris* (Xie et al., 2016). The researchers also found that plants usually emit the scent of flowers during periods when pollinators are more active (Hoballah et al., 2005). Like many organisms, there is a corresponding biological clock in plants that enables them to regulate their growth and development rhythms and affect plant metabolites. Many studies have shown that this biological clock plays an important role in the regulation of plant volatiles. However, the effect of flowering time, flower color, and flower fragrance on the yield of macadamia nut has not been reported, and we think this is a subject worthy of study.

In this experiment, two nut varieties with different flower colors, flowering times, flowering quantities, and yield (single fruit weight) were selected as the research subjects. The dynamic changes of the gene expression and metabolite accumulation in nut flowers (red and white nut flowers) at different flowering stages were analyzed using RNA sequencing (RNA-seq), liquid chromatography–tandem mass spectrometry (LC-MS/MS), and gas chromatography–mass spectrometry (GC-MS). This study investigates the molecular mechanism of compound synthesis and related gene regulation in nut flowers during flower development. The results of this study provide a new perspective for the further study of the main metabolites affecting the color and aroma of macadamia nut flowers, which can provide a theoretical reference for the study of nut flower resistance and optimize the quality and yield of nuts. It will also provide a reference for future breeding and cultivation.

2 Materials and methods

2.1 Plant materials and treatments

The plant materials were collected from three key developmental stages of nut flowers in the "695" (red flower) and "660" (white flower) varieties, which were divided into six groups (i.e., R1, R2, and R3 and W1, W2, and W3). All nut flower samples were collected from Lincang, Yunnan Province, from February to March 2021. The altitude is 850 m (99.259340 E, 24.018357 N). R1 and W1 were at the bud stage (30 days), R2 and W2 were at the half-blooming stage (50 days), and R3 and W3 were at the blooming stage (70 days). The "695" variety (R) has a large number of red flowers that bloom late every year, a medium-sized nut fruit, and is suitable for tropical and subtropical regions (800–1,300 m altitude). The inflorescences of the

"660" variety (W) are white, the flowers are short (about 11 cm), and the nuts are small, making them suitable for tropical and subtropical areas (altitudes of 600–1,300 m). All of the samples were whole flowers, which were immediately frozen in liquid nitrogen and stored at −80°C until use. Three biological replicates were utilized at each time point for LC-MS/MS, GC-MS, and RNA-seq analyses. Each repeat includes at least six flowers collected from two to three macadamia nut trees and mixed in equal proportions.

2.2 Transcriptome sequencing and DEG analysis

Total RNA was extracted from *M. integrifolia* flowers using the Fast Pure Plant Total RNA Isolation Kit (Vazyme, Nanjing, China). The RNA library was prepared and sequenced in the six *M. integrifolia* flower groups: R1, R2, R3, W1, W2, and W3. The first and second strands were synthesized with random oligonucleotides, SuperScript II, DNA polymerase I, and ribonuclease H, and then the 18 libraries were sequenced on the Illumina sequencing platform (HiSeqTM 2500 or HiSeq X Ten; Illumina, San Diego, CA, USA). After sequencing, the raw reads were filtered and the low-quality reads removed to obtain clean reads. All clean reads were assembled into transcripts by Trinity (Grabherr et al., 2011). The transcripts were hierarchically clustered using the Corset program, and the longest cluster sequence was obtained for subsequent analysis (Davidson and Oshlack, 2014). DESeq2 was used to analyze the input read count data, and the screening thresholds were set as follows: $p_{adj} < 0.05$ and $|\log2FoldChange| \geq 1$. An independent statistical hypothesis test was carried out on a large number of genes. Finally, the differentially expressed genes (DEGs) in the transcriptome profile were obtained. The function of the unigenes was annotated using Gene Ontology (GO) terms (http://www.geneontology.org) and analyzed with the Blast2GO program. The Blastall software was used to annotate the GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases.

2.3 GC-MS and volatile organic compound analysis

After vacuum freeze drying, the samples from the different nut flowering periods were ground to powder. For each sample, 500 mg powder was placed in a headspace bottle with saturated NaCl solution and 10 µl (50 µg/ml) internal standard solution. After absorption of the supernatant, the sample was filtered with a microporous membrane (0.22 µm pore size) and then stored in the injection bottle for later GC-MS detection by fully automatic headspace solid-phase microextraction (HS-SPME). In this experiment, the NIST database was used to identify the volatile organic compounds (VOCs) in nut flowers. Multivariate statistical analysis of the VOCs, including principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA), was performed to reveal the differences in the composition of the VOCs in each comparison group. The variable importance in projection (VIP) value of the first principal component of the PLS-DA model was used and was combined with the *p*-value of the *t*-test to determine the differentially accumulated metabolites (DAMs). The selection criteria were VIP ≥ 1.0, $|\log2(\text{fold change})| \geq 1$, and $p < 0.05$.

2.4 Metabolite profiling by UPLC-MS/MS

The freeze-dried flower samples of R and W at three different developmental stages were weighed and ground with zirconia beads at 30 Hz using a mixer mill (MM 400; Retsch, Haan, Germany) for 1.5 min. Each sample (100 mg) was placed into a centrifuge tube (5 ml), extracted by adding 1,500 µl of 1:1 methanol/water, adsorbed using a CNWBOND Carbon-GCB SPE cartridge (250 mg, 3 ml; ANPEL, Shanghai, China), and then filtered (SCAA-104, 0.22 µm pore size; ANPEL, Shanghai, China). The extract was then fed into the LC–electrospray ionization–MS/MS (LC-ESI-MS/MS) system. The high-performance liquid chromatography (HPLC) system used was the Shimadzu Shim-pack UFLC CBM30A system (www.shimadzu.com.cn/). Based on the method of Liu et al. (2021) with some modifications, the chromatographic column was an ACQUITY UPLC HSS T3 C18 (pore size, 1.8 µm; length, 2.1 mm × 100 mm). For the solvent system, phase A was ultrapure water (adding 0.1% formic acid) and phase B was acetonitrile (adding 0.1% formic acid). The experiment was carried out under an injection volume of 5 µl, a flow rate of 0.4 ml/min, and a column temperature of 40°C. The Thermo QE Focus high-resolution mass spectrometer in information correlation acquisition mode was used to collect high-resolution mass spectrometry data. Three ions with strength greater than 5,000 were selected for each cycle. The mass spectral data were analyzed using MAPS software (Kuhl et al., 2012).

2.5 Combined analysis of RNA-seq, LC-MS/MS, and GC-MS

Pearson's correlation coefficient analysis was conducted between the significant DEGs [with fragments per kilobase per million mapped reads (FPKM) values from the RNA-seq profile] and DAMs (as well as VOCs) in the six nut flower groups, which included certain primary and secondary metabolites. The association between the DEGs, DAMs, and VOCs was analyzed in this study. A correlation coefficient less than 0 represents a negative correlation, while a correlation coefficient greater than 0 denotes a positive correlation. Subsequently, the log2-transformed datasets were loaded in the "cor" package from R software. The top 50 and top 100 DAMs, VOCs, and DEGs were then identified. Finally, the biological significance of modern metabolites was analyzed based on their metabolic pathways and other functions. The DEGs, DAMs, and VOCs were selected using $R^2 > 0.9$ as the filter criterion, and network diagrams were constructed between the metabolites and the DEGs. Cytoscape software (v3.6) was used for network diagram visualization to represent the relationship between the metabolome and the transcriptome (LC-MS and GC-MS).

2.6 Quantitative real-time PCR analysis

Sixteen genes related to metabolite synthesis in flowers were selected for real-time quantitative polymerase chain reaction (qRT-

PCR) analysis. These 16 coding genes are involved in nut flower development and color formation and included anthocyanidin 5,3-*O*-glucosyltransferase, flavanone 3-hydroxylase, lipoxygenase, UDP-glycosyltransferase, naringenin, F-box protein, salicylic acid-binding protein, and jasmonic acid-amido synthetase, among others. The primers were designed with Primer 5.0, and their sequences are listed in Table 1. RNA was extracted from nut flowers to synthesize the complementary DNA (cDNA). According to the manufacturer's instructions, qRT-PCR was performed using the ChamQ Universal SYBR q-PCR Master Mix (Vazyme, Nanjing, China). The standard curve was prepared according to the method described by Fan et al. (2017) with *GAPDH* as the internal reference gene. In this study, all genes underwent three biological repeats (with each biological repeat containing three technical repeats). The gene expression level was

TABLE 1 Primers selected for real-time quantitative polymerase chain reaction (qRT-PCR) analysis.

Gene ID		Sequence (5′–3′)	Product size (bp)
Cluster-25626.107700	Forward	TTTGGCACAGTAGGGGCCTT	152
	Reverse	AGTCCCCTCTCCCTTGTACTG	
Cluster-25626.128800	Forward	ATTCACGGGCAAACAACAGG	129
	Reverse	AAACCATCACTCAACGCACA	
Cluster-25626.104601	Forward	GCTTTACTGCCCGAAGGGTT	181
	Reverse	AGAGGGGCCAACTCACCATA	
Cluster-25626.128311	Forward	TGGCAAACGGTCTTCGCTAT	139
	Reverse	TTCCCGACTCAGCACCTCTA	
Cluster-25626.136861	Forward	CAGATGCCGGAGGTCTTACAT	137
	Reverse	ACGGCCTCGATTGCTCTTG	
Cluster-25626.96173	Forward	TCACACCGATCCAGGCACTA	130
	Reverse	GCCAAGGTTGACGACGAAAG	
Cluster-25626.117468	Forward	ACCACAGGAAGAGAAGGAAGC	116
	Reverse	GGAACAAGAAATCCACCCAGC	
Cluster-25626.145312	Forward	GTATCGCCGCCTCTACATGG	168
	Reverse	CTGCCCAAGACCAACCTCTC	
Cluster-25626.70089	Forward	AATCCTAGGTTGCCTCCCGA	200
	Reverse	AGCCTGTACCTGTATAAACCTGC	
Cluster-25626.109979	Forward	GCAATGGATGGTGGTTGTGTG	108
	Reverse	CTGGTTTGGCTGGACACGA	
Cluster-25626.101660	Forward	GTGCAGGTCAAGGACAATGG	200
	Reverse	GTGAGCCACAAGCAAAACCT	
Cluster-25626.109071	Forward	GAGGTCAGAACACAAGGCCA	129
	Reverse	TACAGCAGAATGGTGGCAGG	
Cluster-25626.115681	Forward	CCGGTTATTGAAGCCGCATT	172
	Reverse	CAAAGCCTCCTCGTCAAACC	
Cluster-25626.120871	Forward	CCAGACATTATCACCAACGCAC	181
	Reverse	ACTATTGCTCCTGTTGGGGC	
Cluster-25626.112538	Forward	TTGGGCAAAGGGACAATCCAA	166
	Reverse	AGCAAGGTGTATGGGACCAA	
Cluster-25626.112534	Forward	CTAGATGGCTTCACCGTCGAT	164
	Reverse	GGCTTCAAGGTCCCATCCTC	
GAPDH	Forward	CTTCAACATCATCCCTAGCAGC	102
	Reverse	GTGGGAACACGGAAGGACA	

calculated using the $2^{-\Delta\Delta Ct}$ method. The normalized relative gene expression level and the FPKM value of the RNA-seq data were calculated as the log2(fold change). R software package 3.1.3 was used to analyze the correlation between the RNA-seq and qRT-PCR data.

3 Results

3.1 RNA sequencing, assembly, and quality assessment

To study the gene expression and related regulatory pathways of *M. integrifolia* flowers during their critical developmental stages, transcriptome sequencing was performed on 18 samples from two different flower color varieties of *M. integrifolia*. A total of 124.33 Gb of clean data was obtained through sequencing quality control, and the clean data of each sample reached 6 Gb. The percentage of the Q30 bases was >92% (Table 2). This result indicates that the assembled genome had high accuracy, contiguity, and completeness. The longest cluster sequence obtained by Corset hierarchical clustering and the spliced unigene N50 was 1,476 bp (Table 2). Before the follow-up analysis, we first evaluated the correlation between each sample. The results showed that the correlation of each repetition in nut flowers was greater than 0.9 (Figure 1A). As expected, the replicates of each nut flower pattern were clustered together, indicating small variations among replicates.

The PCA showed that the first principal component could explain 29.89% of the total variance and distinguished the samples based on R (R1, R1, and R3) and W (W1, W2, and W3). The second principal component (PC2) explained 15.04% of the total variance and separated the different flowering stages (Figure 1B). The results suggested that there are significant differences among the different nut varieties (Figure 1). This may be due to the difference in the genetic background of each variety, or could be caused by the difference in the sampling period. In addition, the different flowering stages of the same nut variety were significantly separated on PC2, and the differences increased with flower development (Figure 1B). Subsequently, function annotation was performed on these unigenes, which used the NR, Swiss-Prot, KEGG, GO, and PFAM databases. The annotated unigenes (a total of 287,486) were analyzed for significant enrichment in the different comparison groups.

3.2 Analysis of the DEGs in *M. integrifolia* flowers at different stages

DESeq2 was used to analyze the DEGs from each nut flower group at different developmental stages. The screening criteria for DEGs were |log2FoldChange| ≥1 and false discovery rate (FDR) <0.05. Under these conditions, the results of the DEGs in each group are shown in Supplementary Figure S1, which quantified the correlation between the PCA and the inter-group samples. The results indicated that regardless of the R or W varieties, the first and third flower development stages had the most number of DEGs (Supplementary Figure S1). In the R variety, the numbers of DEGs in the R2 *vs.* R1 and R2 *vs.* R3 comparison groups were comparable.

TABLE 2 Genome sequencing, assembly, and quality assessment in nut flowers.

Type	Number	Mean length	N50	N90	Total bases
Transcript	373,304	854	1,363	347	318,627,966
Unigene	287,486	1,027	1,476	471	295,315,014

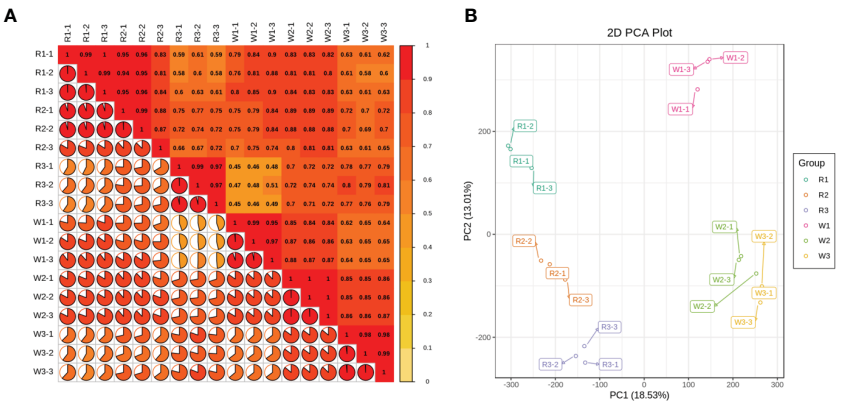


FIGURE 1
(A) Sample-to-sample correlation analysis between each nut flower at different developmental stages. The *numbers in the box* represent the correlation coefficient between each sample; more depth of the *color bar on the right* indicates a higher correlation. The pie chart and the *color* of each sample represent the proportions of the correlation coefficients. **(B)** Principal component analysis (PCA) for each group of nut flowers. *R1*, *R2*, and *R3* represent the three stages of red *Macadamia integrifolia* flowers, while *W1*, *W2*, and *W3* represent the three stages of white *M. integrifolia* flowers. Each color and developmental stage of the nut flower samples had three biological repeats, denoted as *-1*, *-2*, and *-3*.

In the W variety, this change seemed to be concentrated in the early stage of flower development (W1 *vs.* W2). Therefore, we next conducted an in-depth analysis of these DEGs.

GO functional analysis was conducted in the R1 *vs.* R2 and R2 *vs.* R3 groups. The 50 significantly enriched terms were identified in the flowers of the two nut varieties (Supplementary Figure S2). The results showed that, compared with the DEGs in the R2 *vs.* R3 group, those in the R1 *vs.* R2 group were specifically enriched in photosynthesis, light harvesting in photosystem I, and energy and primary metabolism that involved DEGs, such as NAD activity, amino acid transport, and fatty acid metabolism (Supplementary Figure S2A). However, in the late stage of R flower development (R2 *vs.* R3), the DEGs were significantly enriched in secondary regulatory molecular functional terms such as naringenin-chalcone synthase, quercetin 3-*O*-glucosyltransferase, and quercetin 7-*O*-glucosyltransferase (Supplementary Figure S2B). At the same time, we also analyzed the GO enrichment of the DEGs at the different stages of W flower development (Supplementary Figures S2C, D), which was similar to R. The findings suggested that, during the early stage of nut flower development, the DEGs were primarily associated with the accumulation of primary substances, such as energy metabolism. However, in the later stage, the DEGs mainly regulated the accumulation of the secondary functional components.

3.3 Pathway analysis for nut flower color formation

KEGG enrichment analysis was performed in each flowering stage to further examine the color formation and related pathways of *M. integrifolia* flowers. The results showed that plant hormone signal transduction, starch sucrose metabolism, phenylpropane metabolism, and flavonoid biosynthesis pathways were significantly enriched in the different groups of both R and W nut flowers (Figures 2A–D). Therefore, we speculated that these pathways are related to *M.*

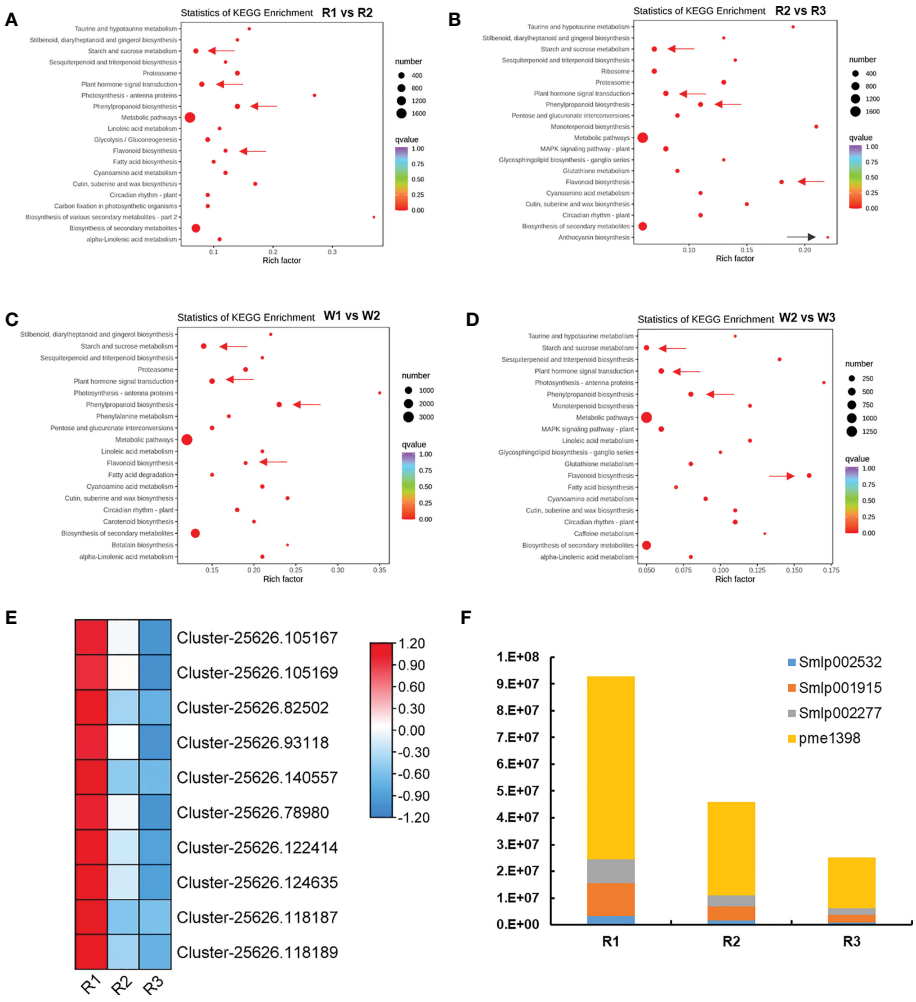


FIGURE 2
Analysis of the differentially expressed genes (DEGs) and flower color formation during nut flower development (*R* and *W* refer to red and white nut flowers, respectively). **(A, B)** Top 20 of the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment of the DEGs in the R1 *vs.* R2 **(A)** and R2 *vs.* R3 **(B)** nut flower comparison groups. **(C, D)** Top 20 of the KEGG enrichment of the DEGs in the W1 *vs.* W2 **(C)** and W2 *vs.* W3 **(D)** nut flower comparison groups. **(E)** Expression changes of the DEGs in the anthocyanin biosynthesis pathway in R nut flowers. **(F)** Changes of the anthocyanin differentially accumulated metabolites (DAMs) in R with the development of nut flowers. The *size of the red dots* in the figure represents the number of enriched DEGs, and the redder the color, the more significant the enrichment. The *arrows next to the red dots* indicate the avenues that this article focused on. *Each square in the heat map* represents a gene, and the *color bars* represent changes in the gene expression, with *red* indicating upregulated gene expression and *blue* downregulated gene expression.

integrifolia flower development and flower color formation. In addition, in the R2 *vs.* R3 comparison group (Figure 2B), the DEGs were also significantly enriched in the anthocyanin biosynthesis pathway (ko00942), which may be the main factor for the R (red) flower color formation.

Therefore, we analyzed the changes of the DEGs in the anthocyanin biosynthesis pathway (Figure 2E). A total of 10 genes were identified. In the R1 stage, these genes were highly expressed, including anthocyanidin synthase anthocyanidin (*ANS*) (*Cluster-25626.105167*, *Cluster-25626.105169*, *Cluster-25626.82502*, and *Cluster-25626.93118*); anthocyanidin reductase (*ANR*) (*Cluster-25626.140557*, *Cluster-25626.78980*, *Cluster-25626.122414*, and *Cluster-25626. 124635*); and leucoanthocyanidin reductase (*LAR*) (*Cluster-25626.118187* and *Cluster-25626.118189*). The results suggested that these genes are related to the color formation of *M. integrifolia* flowers. In order to accurately identify the key genes related to anthocyanin accumulation and to further elucidate the potential mechanism and chemical basis of the coloration and nutritional quality between R and W flowers, metabolomic analysis was conducted through LC-MS/MS.

Throughout the nut flower opening period, eight key anthocyanins (DAMs) showed a distinctive and specific accumulation pattern, distribution, and fading. These DAMs included cypermethrin-3-arabinoside, deltamethrin-3-arabinoside, petunin-3-*O*-arabinoside, cypermethrin-3-*O*-galactoside, cypermethrin-3-*O*-glucoside, paeoniflorin-3-*O*-glucoside, deltamethrin-3-*O*-galactoside, and trifolin-3-*O*-glucoside (*Additional file 1*). Our analysis revealed that the anthocyanin content was closely related to the color of *M. integrifolia* flower, which is consistent with a previous study (Ma et al., 2022). In the metabolic dynamic changes of the flower color along with flowering, two anthocyanins—delphinidin-3-*O*-rutinoside and cyanidin-3-*O*-rutinoside—played a key role in color formation. In the analysis of the accumulation content of these metabolites, four DAMs were not detected (non-existent) in W, while the accumulation of these DAMs was significantly high in R, with the accumulation content decreasing significantly with the development of flowers (Figure 2F). Therefore, based on the phenotypic and transcriptome data of *M. integrifolia* flowers, four possible nut flower chromogenic substances were identified: delphindin-3-*O*-glucoside (Pme1398), petunidin-3-*O*-arabinoside (Smlp002277), delphinidin-3-*O*-arabinoside (Smlp001915), and cyanide-3-*O*-arabinoside (Smlp002532) (Figure 2F). The related DEGs for DAM synthesis were verified by qRT-PCR. These DEGs and DAMs were critical for the color formation of *M. integrifolia* flowers.

3.4 Effects of the plant hormone pathway on *M. integrifolia* flower development

Plant hormones play an important role in the regulation of plant flowering. Further analysis showed that the DEGs were significantly enriched in the plant hormone signal transduction pathway during the process of nut flower development. Therefore, the expression patterns of the plant hormone regulation-related genes were analyzed in R and W. Of these genes, 205 DEGs in R1 *vs.* R2, 213 DEGs in R2 *vs.* R3, 395 DEGs in W1 *vs.* W2, and 166 DEGs in W2 *vs.* W3 were screened (Figure 3A). A total of 137 DEGs were shared in the R1 *vs.* R2 and W1 *vs.* W2 comparison groups. After removing the DEGs whose expression level was less than 10, a total of 77 DEGs were obtained for analysis of the expression patterns (Figure 3). The results revealed that, in both R and W, some genes were downregulated while others were upregulated with flower development (Figure 3B).

Further analysis showed that the downregulated DEGs included flowering inhibitory genes such as *MYC2* (*Cluster-25626.29606*, *Cluster-25626.151777*, *Cluster-25626.206736*, *Cluster-25626.62227*, and *Cluster-25626.190943*), which inhibit plant flowering through various pathways. In addition, some genes, including *GID1* and *GH3*, were downregulated during flower development. Among them, the upregulated DEGs were mainly related to *ARF* (auxin response factor) (*Cluster-25626.109491*, *Cluster-25626.107280*, and *Cluster-25626.129499*); *IAA* (*Cluster-25626.143080*, *Cluster-25626.100525*,

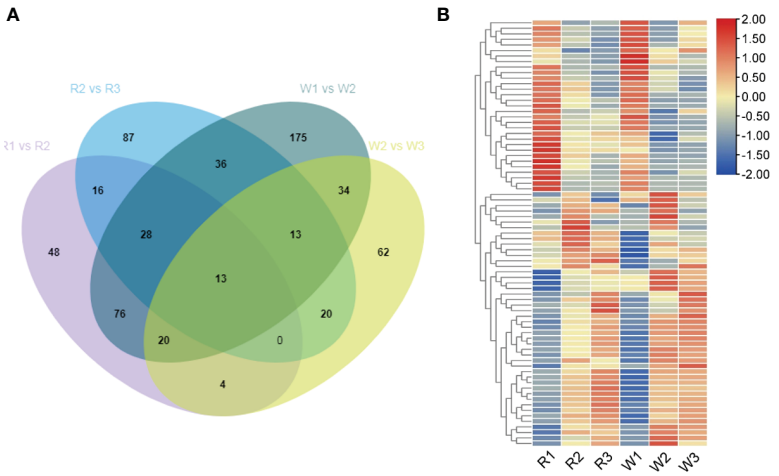


FIGURE 3
Analysis of the differentially expressed genes (DEGs) of the plant hormone pathway in red (R) and white (W) nut flowers at different development stages. **(A)** Venn diagram of the DEGs related to the plant hormone pathway in four comparison groups. **(B)** Analysis of the DEGs in the different development stages of nut flowers. In the Venn diagram, the *numbers* represent the shared or unique DEG counts in the different comparison groups. In the heat map, the *red* and *blue colors* represent upregulated and downregulated gene expression, respectively.

Cluster-25626.132270, *Cluster-25626.130747*, and *Cluster-25626.119650*); *SAUR* (*Cluster-25626.164744*, *Cluster-25626.116576*, *Cluster-25626.86188*, *Cluster-25626.106735*, *Cluster-25626.123530*, *Cluster-25626.119538*, *Cluster-25626.127234*, and *Cluster-25626.106483*); and *AUX1* (*Cluster-25626.127778*, *Cluster-25626.81269*, *Cluster-25626.106873*, *Cluster-25626.107669*, *Cluster-25626.107666*, *Cluster-25626.107665*, *Cluster-25626.72889*, and *Cluster-25626.126447*).

In addition, the DEGs related to plant hormone synthesis were involved in *M. integrifolia* flowering in the R2 *vs.* R3 and W2 *vs.* W3 stages. The results showed that these DEGs mainly included 46 common genes. Except for the 13 DEGs shared in the early stage of flower development, there were 33 DEGs in the early stage of flower development (Figure 3A). In this study, some genes were significantly enriched in the brassinolide (BR) signal transduction pathway at the later stage of nut flower development, including *BRI1* (*Cluster-25626.145185*, *Cluster-25626.70019*, and *Cluster-25626.45513*); *BAK1* (*Cluster-25626.141544*); and *BSK* (*Cluster-25626.225384* and *Cluster-25626.143495*) (Figure 3B). These results indicated that, in the early stage of flower development, most of the DEGs were involved in the IAA signal transduction pathway, which plays a very important role in the early stage of nut flower development (stages 1 and 2). The DEGs in the BR signal transduction pathway play an important role in the late anthesis development of *M. integrifolia* flowers.

3.5 Effects of starch and sucrose metabolism on *M. integrifolia* flower development

This study also found that the DEGs in nut flowers were significantly enriched in the starch and sucrose metabolism pathways. Therefore, we further analyzed the changes in the gene expression in this pathway (Figure 4). Among them, sucrose synthase (SUS), sucrose invertase (INV), and beta-glucosidase (E3.2.1.21) were highly expressed in the early stage of nut flower development, which is consistent with metabolite accumulation. However, the expression of the *TPS* and *ostB* genes was gradually upregulated with the development of flowers (Figure 4). These DEGs regulate and hydrolyze the stored starch in nut flower tissue cells for conversion into small molecular sugars, such as sucrose and fructose, for nut flower growth, which also play an important role in the regulation of *M. integrifolia* flowering.

3.6 Effects of phenylpropanoid biosynthesis on *M. integrifolia* flower development

We analyzed the gene expression changes in phenylpropanoid metabolism, including the phenylpropanoid biosynthesis, flavonoid biosynthesis, anthocyanin biosynthesis, flavonoid and flavonol biosynthesis, and isoflavone biosynthesis pathways. Previously, we have analyzed the relationship between anthocyanin synthesis and nut flower color formation. Therefore, we further analyzed the changes in the gene expression in these metabolic pathways. A total of 83 DEGs were identified in this study (Figure 5). The expression level of phenylalanine ammonia-lyase (*PAL*) (upstream of these pathways) was gradually upregulated with the development of nut flowers (Figure 5). The results showed that the phenylpropanoid biosynthesis pathway was activated during nut flower development. The downstream key genes (peroxidase) of the lignin biosynthesis pathway also basically showed upregulation, while the key genes of the flavonoid biosynthesis pathway, such as chalcone synthase (*CHS*), chalcone isomerase gene (*CHI*), flavanone 3-hydroxylase gene (*F3H*), and 4-coumarate:CoA ligase (*4CL*), were downregulated. These results suggested that, with the development of nut flowers, the gene and related metabolites in the phenylpropanoid biosynthesis pathway mainly flow to lignin biosynthesis branches, while the flavonoid biosynthesis pathway is weakened. The anthocyanin biosynthesis pathway gene and related metabolite synthesis variation was also consistent with that of the flavonoid biosynthesis pathway. In addition, the activation of the phenylalanine pathway during nut flower development may be related to flower aroma formation. Therefore, we further analyzed the VOCs in each stage of *M. integrifolia* flower development.

3.7 Analysis of the aroma in *M. integrifolia* flowers

Only a few studies have revealed the mechanism of aroma formation during nut flower development. Therefore, to elucidate

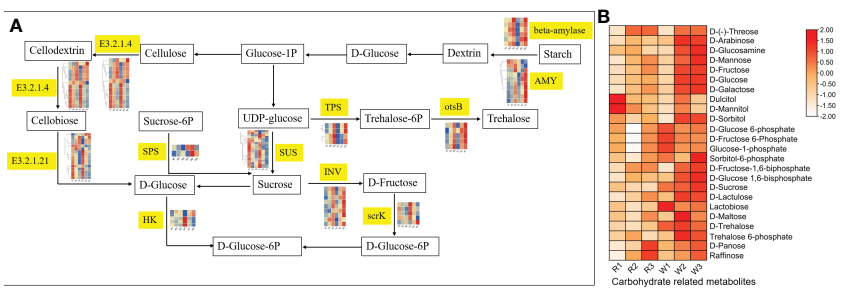


FIGURE 4
Changes of the gene expression and metabolite accumulation levels of nut flowers in the starch and sucrose metabolism pathway. **(A)** Flow diagram of the starch and sucrose metabolism pathway in different nut flower development groups. The *words in yellow background* indicate the names of the genes involved in the pathway. The small heat map corresponding to the gene name represents the gene expression in the different groups of nut flower. The *red* and *blue colors* in the small heat map indicate upregulated and downregulated gene expression, respectively. **(B)** Accumulation patterns of the carbohydrates involved in the starch and sucrose metabolic pathway in the red (R) and white (W) nut flowers. In the heat map, the *red and blue colors* indicate upregulated and downregulated metabolite accumulation, respectively.

this mechanism, GC-MS and LC-MS were used to detect the VOCs (aromas) in R and W in this study. A total of 100 VOCs were identified, including esters (18), alkanes (20), ketones (9), terpenes (12), aldehydes (9), and alcohols (11) (Figure 6A). We analyzed the correlation between the differentially accumulated VOCs and associated DEGs in this study. The VOCs accumulated in the different stages of flower development were screened based on fold change ≥2, fold change ≤0.5, VIP ≥ 1, and $p < 0.05$. The results showed that the contents of VOCs increased gradually with the development of flowers in both R and W (Supplementary Figure S3). However, interestingly, the accumulation of VOCs mainly occurred in the later stages of R, with 27 in the R1 *vs.* R2 group and 40 in the R2 *vs.* R3 group, which were significantly upregulated (Supplementary Figures S3A, B). Such VOCs may be present in the form of precursor metabolites at the early stage (SA). Therefore, they have not been detected by GC-MS. Our results showed that these aroma-related VOCs were produced only at certain stages during flowering. However, the accumulation of VOCs seemed to occur mainly in the early stage of flower development in W (40 in W1 *vs.* W2 and 20 in W2 *vs.* W3) (Supplementary Figures S3C, D). Furthermore, we conducted Venn analysis of the DEGs (upregulated) and found that most of the substances in W1 *vs.* W2 (Figure 6B) overlapped with those in R2 *vs.* R3, while only six were unique. Similarly, in W2 *vs.* W3 (Figure 6C), there were only four unique differences in the accumulation of VOCs. The results showed that there was no significant difference in the aroma composition between R and W, but the main difference was reflected in the time of aroma formation (flower development stage). The formation of flower VOCs (aroma) in W was earlier than that in R. This may be the main factor leading to the difference in yield between the two nut varieties.

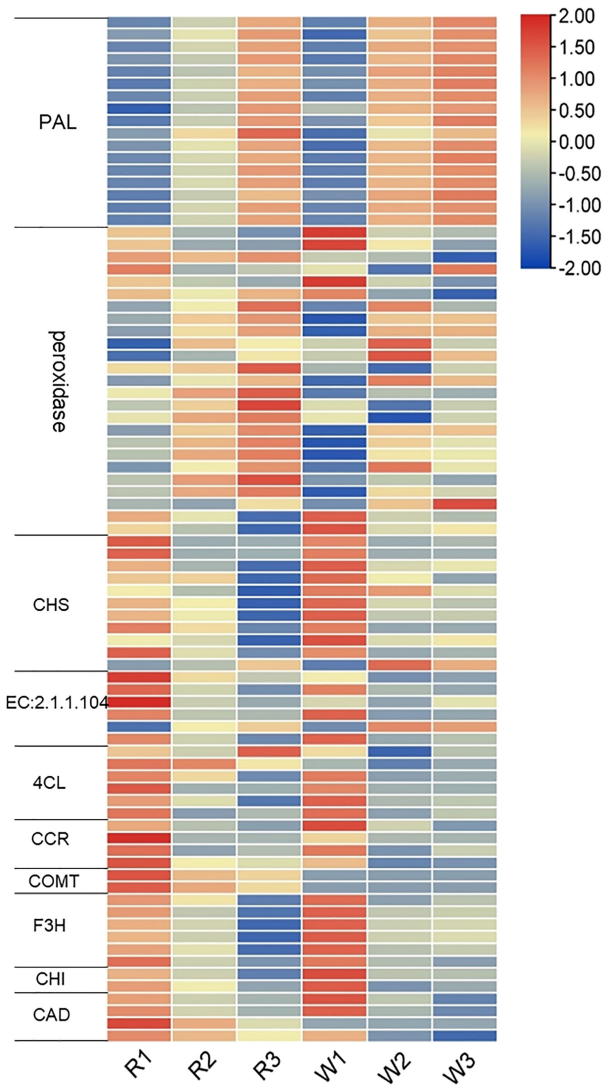


FIGURE 5
Expression analysis of the differentially expressed genes (DEGs) in the phenylpropane biosynthesis pathway at different nut flower development stages. The *red and blue colors* represent upregulated and downregulated gene expression, respectively. On the *left* are the names and classifications of the genes in the phenylpropane biosynthesis pathway. *PAL*, phenylalanine ammonia-lyase; *CHS*, chalcone synthase; *4CL*, 4-coumarate:CoA ligase; *CCR*, cinnamoyl-CoA reductase gene; *COMT*, catechol-*O*-methyltransferase; *F3H*, flavanone 3-hydroxylase gene; *CHI*, chalcone isomerase gene; *CAD*, cinnamyl-alcohol dehydrogenase.

3.8 Combined transcriptome and metabolome analysis

Ultimately, 12 different accumulative VOCs were screened out at different stages of *M. integrifolia* flower development (Table 3). These differentially accumulated VOCs were critical for the aroma formation of *M. integrifolia* flowers. The contents of these VOCs accumulated gradually with the development of flowers. In addition, the accumulation of VOCs in W was significantly higher than that in R. It was speculated that these metabolites are extremely important for the formation of the flower aroma of *M. integrifolia* (Figure 7A). Therefore, the results showed that the synthesis of VOCs in W was earlier than that in R, which is consistent with the phenotypic results of the two varieties. In this experiment, 17 genes were identified in the combined analysis of the metabolome and transcriptome, which were significantly associated with 12 DAMs, and their contents were regulated by the expression of these genes. These genes included three *AAT* genes, one *LOX* gene, and 13 *PAL* genes (Figure 7B). PAL is a key enzyme of the phenylpropanoid pathway that catalyzes the deamination of phenylalanine to *trans*-cinnamic acid, a precursor for the lignin and flavonoid biosynthetic pathways. To date, *PAL* genes have been less extensively studied in gymnosperms than in angiosperms. The key DEGs involved in DAM synthesis were verified by qRT-PCR. The results suggested that these genes are related to the synthesis of key metabolites during *M. integrifolia* flower development, thus regulating the development of the different varieties of flowers. Therefore, in this experiment, these key DEGs were used as the candidate genes for further functional verification.

4 Discussion

Global crop yields are currently trending below the anticipated food demand (Ray et al., 2013; Fróna et al., 2019). Tree crops contribute over 600 million tons of the 10,600 million tons of annual global food production, and fruit number and fruit size are the key components of tree crop yield (Nyomora et al., 1999; Garner et al., 2011; Patrick and Colyvas, 2014). Macadamia (*M. integrifolia*,

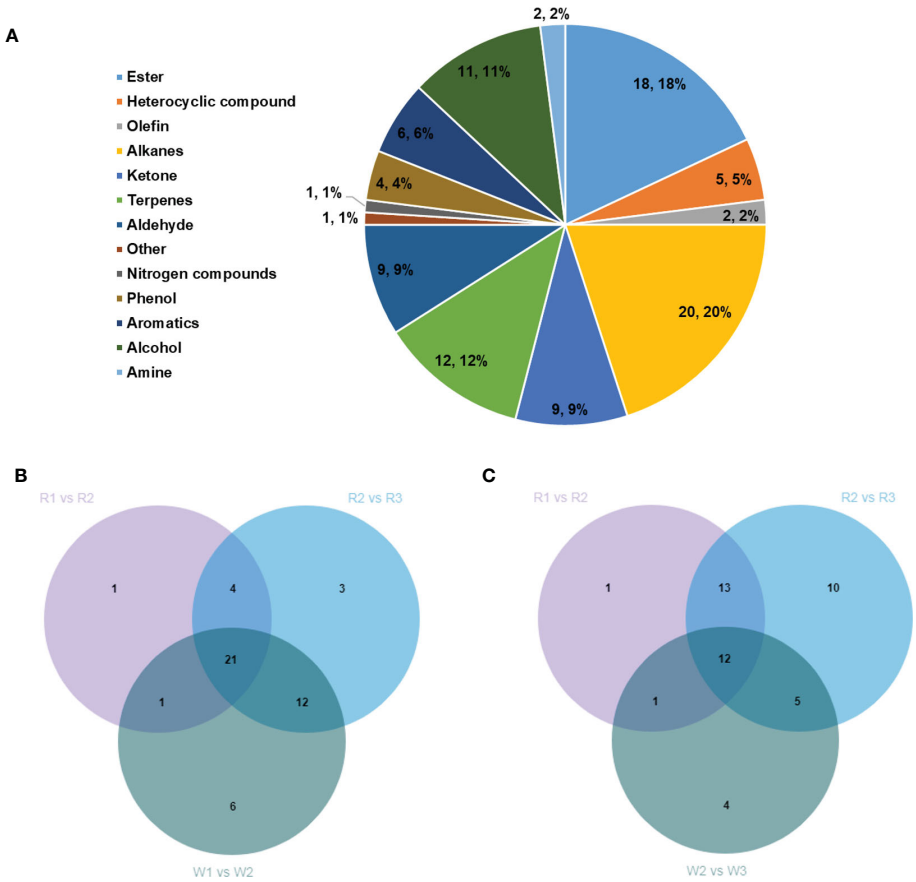


FIGURE 6
(A) Classification and proportion of a total of 100 volatile organic compounds (VOCs) detected in *Macadamia integrifolia* flowers at different development stages. **(B)** Venn diagram analysis of the upregulated VOCs in the W1 *vs.* W2 group compared with the R1 *vs.* R2 and R2 *vs.* R3 nut flower comparison groups. **(C)** Venn diagram analysis of the upregulated VOCs in the W2 *vs.* W3 group compared with the R1 *vs.* R2 and R2 *vs.* R3 nut flower comparison groups. The *percentages* in the pie chart represent the total proportions of the different classified VOCs in nut flowers. The *numbers* in the Venn diagram represent the shared or unique VOC counts in the different comparison groups. *R* and *W* denote red and white nut flowers, respectively.

TABLE 3 Twelve different accumulative volatile organic compounds (VOCs) at different stages of *Macadamia integrifolia* flower development.

Index	Compounds	Exact mass (Da)	Formula	NIST_RI	Class	CAS	RT	RI	Match factor
KMW0323	Benzyl alcohol	$1.08E^{+02}$	C_7H_8O	$1.04E^{+03}$	Alcohol	100-51-6	$1.20E^{+01}$	$1.03E^{+03}$	$9.84E^{+01}$
KMW0384	(2-Nitroethyl)benzene	$1.51E^{+02}$	$C_8H_9NO_2$	$1.19E^{+03}$	Aromatics	6125-24-2	$1.97E^{+01}$	$1.30E^{+03}$	$9.73E^{+01}$
KMW0212	Benzeneacetaldehyde	$1.20E^{+02}$	C_8H_8O	$1.05E^{+03}$	Aldehyde	122-78-1	$1.22E^{+01}$	$1.04E^{+03}$	$9.56E^{+01}$
KMW0350	Acetic acid, phenylmethyl ester	$1.50E^{+02}$	$C_9H_{10}O_2$	$1.16E^{+03}$	Ester	140-11-4	$1.58E^{+01}$	$1.16E^{+03}$	$7.82E^{+01}$
KMW0421	Methyl salicylate	$1.52E^{+02}$	$C_8H_8O_3$	$1.23E^{+03}$	Ester	119-36-8	$1.67E^{+01}$	$1.19E^{+03}$	$7.92E^{+01}$
KMW0437	Acetic acid, 2-phenylethyl ester	$1.64E^{+02}$	$C_{10}H_{12}O_2$	$1.26E^{+03}$	Ester	103-45-7	$1.84E^{+01}$	$1.25E^{+03}$	$9.71E^{+01}$
WMW0196	Lilac aldehyde C	$1.68E^{+02}$	$C_{10}H_{16}O_2$	$1.20E^{+03}$	Aldehyde	53447-48-6	$1.52E^{+01}$	$1.14E^{+03}$	$9.03E^{+01}$
NMW0029	Formic acid, phenylmethyl ester	$1.36E^{+02}$	$C_8H_8O_2$	$1.08E^{+03}$	Ester	104-57-4	$1.32E^{+01}$	$1.08E^{+03}$	$9.00E^{+01}$
NMW0249	(3E,7E)-4,8,12-Trimethyltrideca-1,3,7,11-tetraene	$2.18E^{+02}$	$C_{16}H_{26}$	$1.58E^{+03}$	Terpenes	62235-06-7	$2.67E^{+01}$	$1.57E^{+03}$	$9.53E^{+01}$
XMW1136	Lilac alcohol C	$1.70E^{+02}$	$C_{10}H_{18}O_2$	$1.22E^{+03}$	Alcohol	33081-36-6	$1.72E^{+01}$	$1.21E^{+03}$	$9.40E^{+01}$
XMW1344	Benzoic acid, 2-methylbutyl ester	$1.92E^{+02}$	$C_{12}H_{16}O_2$	$1.39E^{+03}$	Ester	1000367-91-3	$2.34E^{+01}$	$1.44E^{+03}$	$9.21E^{+01}$
D329	Myroxide	$1.52E^{+02}$	$C_{10}H_{16}O$	$1.14E^{+03}$	Terpenes	28977-57-3	$1.89E^{+01}$	$1.27E^{+03}$	$6.61E^{+01}$

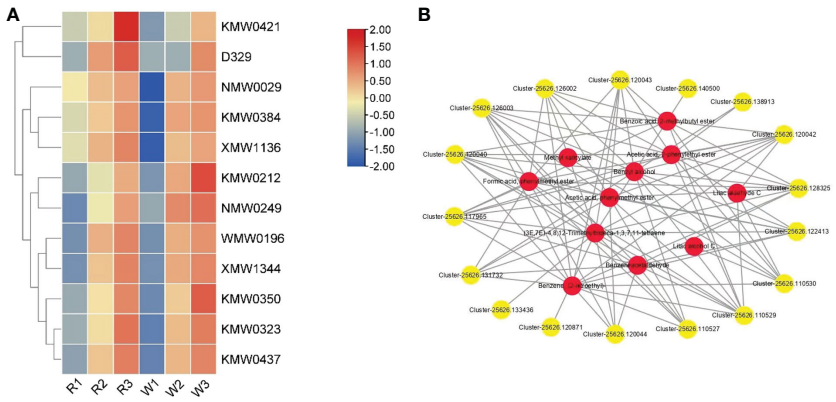


FIGURE 7
(A) Accumulation patterns of the 12 volatile organic compounds (VOCs) in *Macadamia integrifolia* flowers at the R1, R2, R3, W1, W2, and W3 stages (R and W refer to red and white nut flowers, respectively). The *red* and *blue colors* represent up-accumulated and down-accumulated VOCs, respectively. **(B)** STRING network analysis of the differentially expressed genes (DEGs) with shared differentially accumulated VOCs in nut flowers. The correlation in each gene and related VOCs was greater than 0.8. In the network diagram, *yellow dots* represent the key genes and *red dots* represent the associated VOC metabolites. The *solid lines* and *line numbers* indicate the strength of the correlation between the genes and metabolites.

Macadamia tetraphylla, and hybrids) is a subtropical nut crop that produces up to 3,500 racemes per tree annually (Moncur et al., 1985; McFadyen et al., 2011; Olesen et al., 2011). Each raceme has between 100 and 300 flowers (Trueman et al., 2022). Low and inconsistent macadamia yields are often attributed to low levels of initial fruit set, poor fruit retention, and variations in the nut and kernel size (Trueman et al., 1994; Howlett et al., 2019; Kämper et al., 2019; Trueman et al., 2022). Macadamia has a very low fruit-to-flower ratio, which is common among species of the Proteaceae family (Stephenson, 1981; Ayre and Whelan, 1989). Typically, less than 2% of macadamia flowers develop into mature fruits. The fruit setting rate, yield, and quality of the nuts are closely related to the pollination and development of nut flowers. Therefore, we carried out research on nut flowers in order to contribute to the development of the yield and quality of nuts.

With the gradual improvements in the study of flower aroma metabolic engineering, aroma-related substances can be identified in more plants, aroma synthesis pathways can be more clearly elucidated, and more related enzyme genes can be cloned in the future. However, due to the complexity of the flower aroma metabolic pathway and the specificity of plant species and genera, there are still a lot of problems in genetic engineering related to the flower aroma metabolic pathway (Jin et al., 2020). For example, the terpenoid metabolic pathway produces not only a large number of substances related to flower aroma but also some substances related to plant physiological activities, such as abscisic acid and ethylene, and sometimes some metabolic disorders (Jin et al., 2020). In addition, there are still many gaps in the study of phenylpropanes and fatty acids. Further studies are needed to determine the biosynthesis-related genes of phenylpropanes and aliphatic groups and how these genes are regulated by transcription factors. The relationship between nut flower development and flower color formation was investigated from four perspectives: plant hormone signal transduction, starch and sucrose and phenylpropane metabolism, flavonoid biosynthesis pathway, and anthocyanin biosynthesis pathway.

Macadamia flowers are protandrous and have bisexual flowers that are partially self-incompatible (Sedgley, 1983; Trueman, 2013), but the pollen movement between genotypes (cultivars) has positive effects on the yield, improving both nut retention and maximal nut weight (Howlett et al., 2019; Langdon et al., 2020). Dependence on cross-pollination varies between cultivars (Langdon et al., 2019). Many cultivars predominantly set cross-pollinated nuts, even when cross-pollen donors are not interplanted within the block (Richards et al., 2020). Honey bees (*Apis mellifera*) and stingless bees (*Tetragonula* spp., predominantly *Tetragonula carbonaria*) are pollen foraging workers that are effective pollinators of macadamia (Heard and Exley, 1994; Rhodes, 2001; Evansl et al., 2021). The VOCs (aroma) of flowers comprise the main mechanism to attracting bees and other insects to pollinate, thus affecting the fruit setting rate and the yield of nuts. In this study, 10 different types of VOCs and 12 different types of DAMs were found in the macadamia R and W nut flower varieties. The results of the present study suggested that the R nut variety has a large number of flowers and blooms later every year, which can better attract stingless bees for pollination. At the same time, the accumulation of VOCs mainly occurred in the later stage (40 metabolites were significantly accumulated) in R. However, in W, the accumulation of VOCs mainly occurred in the early stage of nut flower development. The period of aroma formation may be related to the attraction of stingless bees. These key metabolites and VOC synthesis were regulated by three *AAT* genes, one *LOX* gene, and 13 *PAL* genes during nut flower development, thus regulating the development of the different varieties of nut flowers. Our results may help to clarify the molecular mechanism of the effect of *M. integrifolia* flower development on fruit yield.

Flower aroma consists of a series of low-molecular-weight and volatile compounds. The composition and the contents of flower fragrance volatiles differ in different types of plants. According to the synthetic metabolic pathway, flower aroma volatiles can be divided into three categories: terpenes, phenylpropane compounds, and aliphatic compounds (Dudareva and Pichersky, 2000). Similarly, this study also detected the differential accumulation of these volatile metabolites in the different developmental stages of nut flowers. We focused on the synthetic pathway of phenylpropane compounds. The expression of the *PAL* genes was upregulated,

while the key genes of the flavonoid biosynthesis pathway, such as *CHS*, *CHI*, *F3H*, and *4CL*, were downregulated. PAL, 4-hydroxycinnamic acid-4-hydroxylase (C4H), and 4-coumaryl-CoA-ligase (4CL) are three key enzymes in the metabolic pathway of phenylpropane compounds (Wang and Cui, 2009). PAL is the key and the rate-limiting enzyme in the phenylpropanoid metabolic pathway. It catalyzes the first step of the phenylalanine pathway and can catalyze the deamination of phenylalanine to cinnamic acid (CA). CA produces 4-coumarate under the hydroxylation of C4H. C4H is the second step of the phenylalanine pathway that requires the joint action of NADPH and oxygen. It is highly specific to the substrate and is closely related to plant lignin (Li et al., 2007; Liang et al., 2014). On the other hand, 4CL catalyzes the synthesis of CoA esters, such as coumaric acid and CA. These are then further transformed into secondary metabolites such as lignin and flavonoids, which comprise the last step of the phenylalanine pathway and the branching point of the different products formed by phenylpropane metabolism.

Flavonoids are a large group of plant-derived compounds that share a common three-ring phenyl benzopyrone structure and are present in nature as free aglycones or glycosides (Khan et al., 2021). They can be classified into anthocyanins, flavans, flavones, flavanols, flavonols, flavanones, flavonones, and isoflavones, among others, based on the degree of oxidation of the middle pyrone ring or the substitution patterns (Di Carlo et al., 1999). In this study, anthocyanins were identified as the key flavonoids, and four possible nut flower coloring substances were identified: delphindin-3-*O*-glucoside (Pme1398), petunidin-3-*O*-arabinoside (Smlp002277), delphinidin-3-*O*-arabinoside (Smlp001915), and cyanide-3-*O*-arabinoside (Smlp002532). These flavonoids may be involved in the coloration of nut flowers and the formation of floral scent VOCs downstream, which help attract pollinators and increase the fruit setting rate and yield.

In conclusion, the transcriptome and metabolome analyses provided in-depth insights into the dynamic flower coloration of *M. integrifolia*. The results of this study provide a reference for researchers to further study nut flower development, flower aroma, and flower color and also provide insights into improving nut yield. The RNA-seq data from this study will be an essential resource for the further functional study of several traits with genome editing and also for molecular marker-assisted breeding to promote genetic studies and novel cultivar breeding for *M. integrifolia*.

Data availability statement

The data present in the study are deposited in the NCBI BioProject repository, accession number PRJNA899604.

Author contributions

YW: Literature search, analysis, investigation, resources, and writing—original draft. JX: Investigation, metabolite analysis, methodology, and writing—review and editing. ZW: Resources, bioinformatic analysis, and writing—review and editing. ZY: Investigation, and sample collection. ZX: Investigation and transcriptome analysis. CW and RS: Conceptualization, experimental design, obtaining research funds, resources, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (32260720), National Key R&D Program of China (2021YFD1000202), the earmarked fund for CARS (CARS-21), Major Science and Technology Project of Yunnan (202102AE090042 and 202002AA10007), Major Science and Technology Project of Kunming (2021JH002), High-End Foreign Experts Program (G2021039002 and 2019013), Yunnan Provincial Financial Forestry Science and Technology Promotion Demonstration Special Project in 2020 [(2020) TS09], and the International Cooperation Base.

Conflict of interest

The 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1095644/full#supplementary-material

References

A'Ida, N., Wilda Larekeng, S. H., Iswanto, I., and Arsyad, M. A. (2021). Shoot initiation for macadamia integrifolia explant with tissue culture technique. *IOP. Conf. Series.: Earth Environ. Sci.* 886 (1):1231. doi: 10.1088/1755-1315/886/1/012133

Ayre, D. J., and Whelan, R. J. (1989). Factors controlling fruit set in hermaphroditic plants: Studies with the Australian proteaceae. *Trends Ecol. Evol.* 4 (9), 267–272. doi: 10.1016/0169-5347(89)90197-3

Davidson, N. M., and Oshlack, A. (2014). Corset: Enabling differential gene expression analysis for *de novo* assembled transcriptomes. *Genome Biol.* 15 (7), 410. doi: 10.1186/s13059-014-0410-6

Di Carlo, G., Mascolo, N., Izzo, A. A., and Capasso, F. (1999). Flavonoids: old and new aspects of a class of natural therapeutic drugs. *Life Sci.* 65 (4), 337–353. doi: 10.1016/s0024-3205(99)00120-4

Du, L. Q., Zou, M. H., Zeng, H., Luo, L. F., Zhang, H. Z., and Lu, C. Z. (2010). Analysis of the nutritional components of macadamia integrifolia kernel. *Acta Nutrimenta. Sin.* 32 (1), 95–97. doi: 10.1631/jzus.B1000005

Dudareva, N., and Pichersky, E. (2000). Biochemical and molecular genetic aspects of floral scents. *Plant Physiol.* 122 (3), 627–633. doi: 10.1104/pp.122.3.627

Duxbury, D. D. (1995). Lipid scientists shake healthy macadamia nut tree. *Food Process.* 54 (6), 83.

Evansl, J., Jesson, L., Read, S. F. J., Jochym, M., and Howlett, B. G. (2021). Key factors influencing forager distribution across macadamia orchards differ among species of managed bees. *Basic. Appl. Ecol.* 53, 74–85. doi: 10.1016/j.baae.2021.03.001

Fan, H., Dong, H., Xu, C., Liu, J., Hu, B., Ye, J., et al. (2017). Pectin methylesterases contribute the pathogenic differences between races 1 and 4 of fusarium oxysporum f. sp. cubense. *Sci. Rep.* 7 (1), 13140. doi: 10.1038/s41598-017-13625-4

Fróna, D., Szenderák, J., and Harangi-Rákos, M. (2019). The challenge of feeding the world. *Sustainability* 11, 5816. doi: 10.3390/su11205816

Garg, M. L., Blake, R. J., and Wills, R. B. (2003). Macadamia nut consumption lowers plasma total and LDL cholesterol levels in hypercholesterolemic men. *J. Nutr.* 133 (4), 1060–1063. doi: 10.1093/jn/133.4.1060

Garner, L., Klein, G., Zheng, Y., Khuong, T., and Lovatt, C. J. (2011). Response of evergreen perennial tree crops to gibberellic acid is crop load-dependent: II. GA3 increases yield and fruit size of 'Hass' avocado only in the on-crop year of an alternate bearing orchard. *Sci. Horticult.* 130, 753–761. doi: 10.1016/j.scienta.2011.08.033

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29 (7), 644–652. doi: 10.1038/nbt.1883

Heard, T. A., and Exley, E. M. (1994). Diversity, abundance, and distribution of insect visitors to macadamia flowers. *Environ. Entomol.* 23, 91–100. doi: 10.1093/ee/23.1.91

Herbert, S. W., Walton, D. A., and Wallace, H. M. (2019). Pollen-parent affects fruit, nut and kernel development of macadamia. *Sci. Horticult.* 244, 406–412. doi: 10.1016/j.scienta.2018.09.027

Hoballah, M. E., Stuurman, J., Turlings, T. C., Guerin, P. M., Connétable, S., and Kuhlemeier, C. (2005). The composition and timing of flower odour emission by wild petunia axillaris coincide with the antennal perception and nocturnal activity of the pollinator manduca sexta. *Planta* 222 (1), 141–150. doi: 10.1007/s00425-005-1506-8

Howlett, B. G., Read, S. F., Alavi, M., Cutting, B. T., Nelson, W. R., Goodwin, R. M., et al. (2019). Crosspollination enhances macadamia yields, even with branch-level resource limitation. *HortScience* 54, 609–615. doi: 10.21273/HORTSCI13329-18

Jin, L., Zhang, D. S., Liu, Z. X., Shi, J. X., Xu, J., Wang, J. Z., et al. (2020). Research progress on metabolic pathway and molecular mechanism of plant flower fragrance production. *Jiangsu. Agric. Sci.* 48 (23), 51–59.

Kämper, W., Wallace, H. M., Ogbpurnrne, S. M., and Trueman, S. J. (2019). Dependence on Cross-Pollination in Macadamia and Challenges for Orchard Management. *Procedings* 36 (1), 76. doi: 10.3390/proceedings2019036076

Khan, J., Deb, P. K., Priya, S., Medina, K. D., Devi, R., Walode, S. G., et al. (2021). Dietary flavonoids: Cardioprotective potential with antioxidant effects and their pharmacokinetic, toxicological and therapeutic concerns. *Molecules* 26 (13), 4021. doi: 10.3390/molecules26134021

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., and Neumann, S. (2012). CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84 (1), 283–289. doi: 10.1021/ac202450g

Langdon, K. S., King, G. J., Baten, A., Mauleon, R., Bundock, P. C., Topp, B. L., et al. (2020). Maximising recombination across macadamia populations to generate linkage maps for genome anchoring. *Sci. Rep.* 10 (1), 5048. doi: 10.1038/s41598-020-61708-6

Langdon, K. S., King, G. J., and Nock, C. J. (2019). DNA Paternity testing indicates unexpectedly high levels of self-fertilisation in macadamia tree. *Genet. Genomes* 15, 29. doi: 10.1007/s11295-019-1336-7

Li, L., Zhao, Y., and Ma, J. L. (2007). Recent progress on key enzymes: PAL, C4H and 4CL of phenylalanine metabolism pathway. *Chin. J. Bioinf.* 4, 187–189.

Li, Y. P., Finnegan, P. M., Liu, M., Zhao, J. X., Nie, Y. L., and Tang, L. (2022). First report of neofusicoccum parvum causing leaf spot disease on macadamia integrifolia in China. *Plant Dis.* 25, 1034. doi: 10.1094/PDIS-02-22-0299-PDN

Liang, L., Han, X. M., Zhang, Z., Guo, Q. M., Xu, Y. H., Liu, J., et al. (2014). Cloning and expression analysis of cinnamate 4-hydroxylase (C4H) reductase gene from aquilaria sinensis. *China J. Chin. Mater. Med.* 39 (10), 1767–1771. doi: 10.4268/cjcmm20141004

Liu, R., Lu, J., Xing, J., Du, M., Wang, M., Zhang, L., et al. (2021). Transcriptome and metabolome analyses revealing the potential mechanism of seed germination in polygonatum cyrtonema. *Sci. Rep.* 11 (1), 12161. doi: 10.1038/s41598-021-91598-1

Ma, B., Wu, J., Shi, T. L., Yang, Y. Y., Wang, W. B., Zheng, Y., et al. (2022). Lilac (*Syringa oblata*) genome provides insights into its evolution and molecular mechanism of petal color change. *Commun. Biol.* 5 (1), 686. doi: 10.1038/s42003-022-03646-9

Maguire, L. S., O'Sullivan, S. M., Galvin, K., O'Connor, T. P., and O'Brien, N. M. (2004). Fatty acid profile, tocopherol, squalene and phytosterol content of walnuts, almonds, peanuts, hazelnuts and the macadamia nut. *Int. J. Food Sci. Nutr.* 55 (3), 171–178. doi: 10.1080/09637480410001725175

Mcfadyen, L., Cfadyen, L., Robertson, D., Sedgley, M., Kristiansen, P., and Olesen, T. (2012). Effects of the ethylene inhibitor aminoethoxyvinylglycine (AVG) on fruit abscission and yield on pruned and unpruned macadamia trees. *Sci. Horticult.* 137, 125–130. doi: 10.1016/j.scienta.2012.01.028

McFadyen, L. M., Robertson, D., Sedgley, M., Kristiansen, P., and Olesen, T. (2011). Post-pruning shoot growth increases fruit abscission and reduces stem carbohydrates and yield in macadamia. *Ann. Bot.* 107 (6), 993–1001. doi: 10.1093/aob/mcr026

Moncur, M. W., Stephenson, R. A., and Trochoulias, T. (1985). Floral development of macadamia integrifolia maiden & betche under Australian conditions. *Sci. Horticult.* 27, 87–96. doi: 10.1016/0304-4238(85)90058-5

Nyomora, A. M., Brown, P. H., and Krueger, B. (1999). Rate and time of boron application increase almond productivity and tissue boron concentration. *HortScience* 34, 242–245. doi: 10.1023/A:1008795417575

Olesen, T., Huett, D., and Smith, G. (2011). The production of flowers, fruit and leafy shoots in pruned macadamia trees. *Funct. Plant Biol.* 38 (4), 327–336. doi: 10.1071/FP11011

Patrick, J. W., and Colyvas, K. (2014). Crop yield components - photoassimilate supply- or utilisation limited-organ development? *Funct. Plant Biol.* 41 (9), 893–913. doi: 10.1071/FP14048

Qi, Y., Zhang, H., Peng, J., Zeng, F., Xie, Y., Yu, Q., et al. (2022). First report of neopestalotiopsis clavispora causing leaf spot disease on banana (Musa acuminate l.) in China. *Plant Dis.* 17, 1790. doi: 10.1094/PDIS-03-22-0455-PDN

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PloS One* 8 (6), e66428. doi: 10.1371/journal.pone.0066428

Rhodes, J. (2001). Macadamia pollination. *CalMac. News.* 2, 10–19.

Richards, T. E., Kämper, W., Trueman, S. J., Wallace, H. M., Ogbourne, S. M., Brooks, P. R., et al. (2020). Relationships between nut size, kernel quality, nutritional composition and levels of outcrossing in three macadamia cultivars. *Plants (Basel).* 9 (2), 228. doi: 10.3390/plants9020228

Russell, G., Alyssa, C., Amjad, A., and Theodore, R. (2018). Soil amendments and soil profiling impact on macadamia growth and yield performance. *HortScience* 54 (3), 519–527. doi: 10.21273/HORTSCI13572-18

Schiestl, F. P. (2010). The evolution of floral scent and insect chemical communication. *Ecol. Lett.* 13 (5), 643–656. doi: 10.1111/j.1461-0248.2010.01451.x

Sedgley, M. (1983). Pollen-tube growth in macadamia. *Sci. Hortic.* 18, 333–341. doi: 10.1016/0304-4238(83)90015-8

Stephenson, A. G. (1981). Flower and fruit abortion: Proximate causes and ultimate functions. *Annu. Rev. Ecol. Systematics.* 12, 253–279. doi: 10.1146/annurev.es.12.110181.001345

Stephensonra, T., and Macadamia, T. (1994). *Handbook of environmental physiology of fruit crops, volume II, subtropical and tropical crops* (Boca Ration: CRC Press), 147–163.

Trueman, S. J. (2013). The reproductive biology of macadamia. *Sci. Horticult.* 150, 354–359. doi: 10.1016/j.scienta.2012.11.032

Trueman, S. J., Kämper, W., Nichols, J., Ogbourne, S. M., Hawkes, D., Peters, T., et al. (2022). Pollen limitation and xenia effects in a cultivated mass-flowering tree, macadamia integrifolia (Proteaceae). *Ann. Bot.* 129 (2), 135–146. doi: 10.1093/aob/mcab112

Trueman, S. J., Rueman, S. J., and Turnbull, C. (1994). Effects of cross-pollination and flower removal on fruit set in macadamia. *Ann. Bot.* 73 (1), 23–32. doi: 10.1006/anbo.1994.1003

Urata, U. (1954). "Pollination requirements of macadamia," in *Hawaii Agricultural experiment station technical bulletin.* vol. 6 (Honolulu, HI: Hawaii Agricultural Experiment Station, University of Hawaii), 1–40. Available at: http://hdl.handle.net/10125/34466.

Wang, H., and Cui, Z. F. (2009). Regulation of shikimic acid biosynthesis pathway. *Biotechnol. Bull.* 03), 50–53.

Xie, Q., Liu, Z., Meir, S., Rogachev, I., Aharoni, A., Klee, H. J., et al. (2016). Altered metabolite accumulation in tomato fruits by coexpressing a feedback-insensitive AroG and the PhODO1 MYB-type transcription factor. *Plant Biotechnol. J.* 14 (12), 2300–2309. doi: 10.1111/pbi.12583

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 February 2023
DOI 10.3389/fpls.2023.1067613

OPEN ACCESS

EDITED BY
Li Wang,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Joseph Lynch,
West Virginia University, United States
Tao Zhao,
Northwest A&F University, China

*CORRESPONDENCE
Gaurav D. Moghe
gdm67@cornell.edu

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 12 October 2022
ACCEPTED 12 January 2023
PUBLISHED 10 February 2023

CITATION
Kruse LH, Fehr B, Chobirko JD and Moghe GD (2023) Phylogenomic analyses across land plants reveals motifs and coexpression patterns useful for functional prediction in the BAHD acyltransferase family.
Front. Plant Sci. 14:1067613.
doi: 10.3389/fpls.2023.1067613

Phylogenomic analyses across land plants reveals motifs and coexpression patterns useful for functional prediction in the BAHD acyltransferase family

Lars H. Kruse[1,2], Benjamin Fehr[3], Jason D. Chobirko[4] and Gaurav D. Moghe[1]*

[1]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, United States, [2]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada, [3]Computational Biology Department, Cornell University, Ithaca, NY, United States, [4]Molecular Biology and Genetics Department, Cornell University, Ithaca, NY, United States

The BAHD acyltransferase family is one of the largest enzyme families in flowering plants, containing dozens to hundreds of genes in individual genomes. Highly prevalent in angiosperm genomes, members of this family contribute to several pathways in primary and specialized metabolism. In this study, we performed a phylogenomic analysis of the family using 52 genomes across the plant kingdom to gain deeper insights into its functional evolution and enable function prediction. We found that BAHD expansion in land plants was associated with significant changes in various gene features. Using pre-defined BAHD clades, we identified clade expansions in different plant groups. In some groups, these expansions coincided with the prominence of metabolite classes such as anthocyanins (flowering plants) and hydroxycinnamic acid amides (monocots). Clade-wise motif-enrichment analysis revealed that some clades have novel motifs fixed on either the acceptor or the donor side, potentially reflecting historical routes of functional evolution. Co-expression analysis in rice and Arabidopsis further identified BAHDs with similar expression patterns, however, most co-expressed BAHDs belonged to different clades. Comparing BAHD paralogs, we found that gene expression diverges rapidly after duplication, suggesting that sub/neo-functionalization of duplicate genes occurs quickly *via* expression diversification. Analyzing co-expression patterns in Arabidopsis in conjunction with orthology-based substrate class predictions and metabolic pathway models led to the recovery of metabolic processes of most of the already-characterized BAHDs as well as definition of novel functional predictions for some uncharacterized BAHDs. Overall, this study provides new insights into the evolution of BAHD acyltransferases and sets up a foundation for their functional characterization.

KEYWORDS

gene family, phylogenomics, protein function, co-expression, plant metabolism

1 Introduction

The metabolic diversity of plants is immense, and this diversification is a result of frequent gene duplications in plant genomes as well as enzyme promiscuity (Moghe and Kruse, 2018; Pichersky and Raguso, 2018). Proliferation of duplicated genes *via* tandem, segmental and whole genome duplication has resulted in the emergence of enzyme families, which abound in plant metabolism. Reduced selection pressure, reduced turnover rates, and increased promiscuity are key characteristics of such families (Milo and Last, 2012). Gene duplication and divergence has also driven the creation in emergence of novel clades especially in larger families. Understanding how these clades originated and evolved is crucial to understanding how new functions emerge in enzyme families. In this study, we sought to examine the patterns of sequence and expression evolution in the different clades of the BAHD acyltransferase enzyme family (BAHDs or BAHD family) across plant evolution.

The BAHD family is one of the largest multi-functional enzyme families in plants (D'Auria, 2006; Bontpart et al., 2015; Kruse et al., 2022; Moghe et al., 2023). Members of the family perform an acylation reaction using an acyl coenzyme A donor and an acceptor with a hydroxyl or an amine group. The acyl chains transferred can be very diverse and can include aromatic groups (e.g. benzoyl, coumaroyl) as well as aliphatic chains from 2-12 carbons long, with unsaturation (e.g. tigloyl) and branching (e.g. isovaleryl). The family comprises members involved in a wide range of metabolic pathways such as those of lignins, cuticular waxes, anthocyanins and flavonoids, herbivore defense compounds, polyamines, volatile terpenes, aromatics, and many others (D'Auria, 2006; Bontpart et al., 2015; Kruse et al., 2022; Moghe et al., 2023). Thus, the BAHD family has played a critical role in adaptation of plants to terrestrial environments, abiotic stresses and biotic interactions. The wide range of decorations performed by many substrate-promiscuous members of this family leads to emergence of new structural scaffolds (e.g. monolignols, acylsugars) or altered physicochemical properties (e.g. aromatic esters, acylated anthocyanins), increasing the functional diversity of plant metabolites.

BAHDs are closely related to alcohol acyltransferases in fungal species (Moghe et al., 2023) and our previous study (Kruse et al., 2022) revealed that this family expanded in land plants from 1-2 members in algae to ~100 members in several diploid angiosperm genomes, likely *via* tandem gene duplication. Eight clades were identified in the family of which seven (clades 1-7) are present across land plants and clade 0 present only in algae. While most clades comprise enzymes restricted to using a predictable substrate class (e.g. aromatic or aliphatic alcohols, anthocyanins/flavonoids), some clades have diversified members and more lineage-specific sub-clades. Prior studies have also identified rapid functional divergence in BAHDs, even between species (Fan et al., 2016; Fan et al., 2017) and populations (Kim et al., 2012; Schilmiller et al., 2015; Landis et al., 2021) thereby revealing a substantial diversification of the BAHD family in land plants. A recent review described the mechanistic and evolutionary aspects of BAHDs in detail (Moghe et al., 2023). However, the clade-wise patterns of evolution in this family in land plants have not been studied, limiting our understanding of rapid enzyme diversification in such a large and important enzyme family.

In this study, we sought to determine the different ways by which BAHDs have diversified at the sequence, structural, and expression level during land plant evolution. We found evidence of clade-specific expansions and fixation of lineage-specific clades at different points in the evolution of plants. Discriminant analysis of clade-specific motifs revealed some clades with acceptor-side evolution vs. others with donor-side evolution. We also found that duplicated BAHDs have diversified at both expression and substrate-preference level, although some still retain functional similarity with their closest paralogs. Overall, this study provides novel insights into the emergence of functional diversity in the BAHD acyltransferase family.

2 Materials and methods

2.1 Identification of BAHD proteins from sequenced proteomes and analysis of genomic features

For the identification of BAHD acyltransferases from the genomes analyzed in this study (Supplementary File 3) we followed the same approach used previously (Kruse et al., 2022), specifically, using the PFAM domain PF02458 with the HMMER software (Potter et al., 2018). After identification, we gathered additional genomic information from the respective general feature format (GFF) files for each species using custom Python scripts.

2.2 Motif analysis

We identified the top five enriched motifs in each clade using discriminant analysis *via* STREME v5.4.1 (Bailey, 2021) using default parameters but with following modifications: *–protein -nmotifs 5*. We used sequences from each clade as well as orthologous sequences (OGs) corresponding to those clades, which had been previously defined (Kruse et al., 2022) using OrthoFinder (Emms and Kelly, 2019). As background distribution, we used all other BAHD sequences not assigned to that specific clade. Protein structures were downloaded from the Protein Data Bank (Berman et al., 2000) or the AlphaFold protein structure database (for AtCER2 only) (https://alphafold.ebi.ac.uk/). The top 5 motifs were mapped onto the structures using the UCSF Chimera software (Pettersen et al., 2004). Whether the motifs were exposed to the acceptor/donor binding domains was determined manually based on knowledge of these regions as per the AtHCT and Dm3MAT3 structures.

2.3 Phylogenetic analysis

To generate species-specific phylogenetic trees, a protein sequence alignment of all identified BAHDs was generated using MAFFT v.7.453-with-extensions as described earlier (Kruse et al., 2022). IQ-Tree v1.6.10 (Nguyen et al., 2015) was then used to infer a phylogenetic tree using following parameters: *-st AA -nt AUTO -ntmax 12 -b 1000 -m TEST* with automatic model selection. The resulting trees were visualized using iTol v.5.6.2 (Letunic and Bork, 2021).

2.4 Blast and phmmer sequence mapping

To map the known BAHD clades (Kruse et al., 2022) to the BAHD sequences identified from the different analyzed species, we used two different approaches. In approach 1 we used blastp to map biochemically characterized BAHDs to the newly identified BAHDs from each species. Here, we used blastall with the following parameters: *-p blastp -e 1 -m 8*. Subsequently, we filtered out the best top hits and applied a filter of 40% sequence identity and 200 amino acid match length between query and target. In approach 2, we used phmmer (hmmer v3.3 package) (Potter et al., 2018) with the following parameters: *–noali -E 1e-20*. Afterwards, we filtered out hits with e-value larger than 1e-50. For comparison, we also ran phmmer without specified e-value. Finally, we used ITOL v.5.6.2. (Letunic and Bork, 2021) to map the clade assignments to the individual, species-specific BAHD trees to illustrate the spread of each clade across the analyzed species.

2.5 Gene expression analysis in Arabidopsis and rice and calculation of synonymous rate

Normalized gene expression information for Arabidopsis was downloaded from Arabidopsis RNA-seq database (ARS; http://ipf.sustech.edu.cn/pub/athrna/) (Zhang et al., 2020). Expression data for rice was gathered from: https://tenor.dna.affrc.go.jp/downloads (downloaded on June 4, 2021). Subsequently we translated RAP-DB locus IDs to MSU locus IDs using a mapping file downloaded from https://rapdb.dna.affrc.go.jp/download/irgsp1.html (downloaded on June 4, 2021). Expression values of BAHDs were isolated from each of the datasets, and were used to calculate Pearson correlation coefficient using base R v4.0.5 (R Core Team, 2021). Plots were generated in R using ggplot2. The blue line represents the best fit of a linear model (lm) and the shaded area represents the 95% confidence interval. R^2 was calculated using the lm formula in base R (R Core Team, 2021). Kruskal-Wallis rank sum tests were performed to detect significant differences using entire datasets using base R. Statistical tests between bins were performed using Kolmogorov-Smirnov (KS) test with multiple testing correction using p.adjust (method = "fdr") in R. All tests were performed in R v4.0.5 (R Core Team, 2021). For calculating substitution rates, all pairwise comparisons of paralogous BAHDs were calculated using the yn00 function in the PAML software and compiled using custom Python scripts.

2.6 Co-expression analysis for pathway prediction

The ATTED co-expression data table (Ath-u.v21-01.G18957-S27427.combat_pca_subagging.ls.d) was downloaded (Obayashi et al., 2022). This table was constructed using integrative assessment of both RNA-seq and microarray datasets as described here (https://atted.jp/static/help/download.shtml#method). The co-expression significance value of each gene is expressed as a z-score. From the Entrez Gene IDs noted in the ATTED data file, genes with z-score ≥ 3 were considered co-expressed while those with z-score ≤ 1 were considered not-co-expressed. Pathway assignments for each gene were obtained from Plant Metabolic Network (ara_pathways.20210325.txt) (Hawkins et al., 2021). For each BAHD, we first asked which other genes with pathway information were co-expressed. Using the co-expressed and not-co-expressed gene sets, we then performed an enrichment analysis to determine if a given pathway was enriched among the co-expressed genes. Statistical significance was determined using Fisher Exact Test with multiple testing correction based on Q-value (Storey, 2002).

3 Results

3.1 BAHD gene and protein features experienced substantial changes in land and non-land plant lineages

Previous results demonstrated that BAHDs expanded in land plants faster than the increase in genomic gene content *via* repeated duplications (Kruse et al., 2022). Investigation of 52 sequenced genomes revealed that this expansion was also associated with changes in gene and protein structure. The number of BAHDs increased from 1-5 copies in algal genomes to dozens to hundreds of copies in diploid plant genomes (Kruse et al., 2022). An interesting gradation in intron counts was observed (Figures 1A, B), with chlorophyte BAHDs having multiple introns, non-seed plant BAHDs generally showing a single intron, and most seed plant BAHDs having 0-1 introns. Furthermore, for chlorophytic BAHDs with introns, the average intron size is also considerably larger than expected (and for unknown reasons, the fern *Azolla filiculoides*), leading to larger overall gene locus length (Figures 1C, D). It is not clear if the *A. filiculoides* intron size increase is associated with the overall high number of transposable elements in the genome (Li et al., 2018). On average, the coding sequence length of the algal BAHDs is also larger (Figure 1E), indicating that these BAHDs may have different roles and regulatory behaviors than angiosperm BAHDs.

To investigate whether algal BAHDs have different domain structure than land plant BAHDs, we scanned all identified BAHDs for presence of any other domains as described in the Protein Family (PFAM) annotation. The PFAM database identifies 285 different architectures for this domain model (PF02458) across 791 species, with 88.2% of BAHDs not showing co-occurrence with any other domain. The algal genomes also follow this trend. We specifically analyzed 38 BAHD sequences from 28 species, of which 28 (73.6%) sequences showed a singular BAHD domain. This result suggests that most functional novelty in algal BAHDs arises due to innovations within the BAHD domain as against due to its co-occurrence with other domains. Furthermore, alignment of 13 algal BAHDs selected previously (Kruse et al., 2022) with twenty random, biochemically characterized land plant BAHDs revealed that eight BAHD algal sequences (e.g. from *Chlamydomonas reinhardtii*, *Chara braunii*, and *Micromonas pusilla*) were longer and contain sequence regions that cannot be found in land plant BAHDs, the significance of which is not clear (Supplementary Figure 1). To the best of our knowledge, only one BAHD from algae (*Chara braunii* HQT-like) has been characterized (Kruse et al., 2022). When tested *in vitro* against a panel of 12 substrates, this enzyme catalyzed only the acylation of quinate using coumaroyl-CoA, however, it is unknown if other *in vivo* substrates of this enzyme exist. CbHQT-like, despite its marked

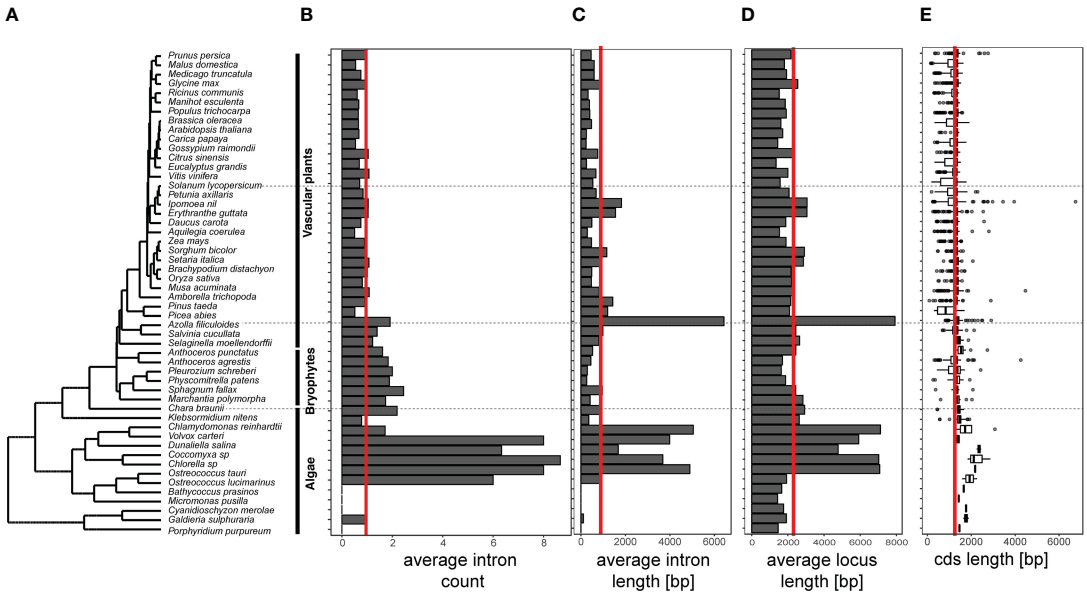


FIGURE 1
Genomic structure of BAHDs accross 52 representative genomes of plant species. **(A)** Representative species were selected from algal and land plant species and a species tree was inferred and illustrated as described in Kruse et al., 2022 and Moghe et al., 2023. Subsequently, we identified BAHDs using hmmsearch, estimated the number of introns per gene **(B)**, intron length **(C)**, locus length **(D)**, and coding sequence (cds) length **(E)** using custom Python scripts. The vertical red line indicates the average across all analyzed species.

longer sequence length (572 residues) compared to the average length of land plants (407 residues), catalyzes a typical BAHD reaction that is conserved across land plants. Currently the explanation for structural differences of algae and land plant BAHDs remains unknown.

The BAHD family was previously predicted to have expanded in land plants, resulting in seven different clades (Kruse et al., 2022; Moghe et al., 2023). These clades are functionally divergent, with their functions defined using experimentally characterized BAHDs belonging to those clades. Thus, we explored the clade-wise expansions of BAHDs over multiple species over land plant evolution.

3.2 Clade-specific expansions and duplication-divergence characterize BAHD evolution in land plants

To better understand the expansions of the seven clades in land plants, we obtained maximum likelihood trees of BAHD protein sequences from 13 species selected in a phylogeny-guided manner. Using BLAST and phmmer, we identified the best hits of each BAHD in a given species to the biochemically characterized enzymes previously defined to be in each clade (Kruse et al., 2022; Moghe et al., 2023), adopting a clade nomenclature that was updated from a system used earlier (D'Auria, 2006; Tuominen et al., 2011). Two versions of the similarity search results are shown (Figure 2) – the top hits of BLAST and phmmer (which is more sensitive than BLAST) without any filtering (Relaxed Set), and hits after filtering them with a 40% identity and 200 amino acid length threshold that, in our experience, typically filters random hits of BAHDs (Conservative Set). While the Relaxed Set assigns every enzyme in each species to a clade, the Conservative Set reveals novel, species-specific clades comprising BAHDs that have sufficiently diverged from their ancestors at the sequence level. Although there were slight differences between BLAST and phmmer, the overall trend remained the same.

The *C. braunii* BAHDs are sufficiently divergent from land plant BAHDs, forming a separate clade on their own, previously referred to as clade 0 (Kruse et al., 2022). Clade 1, involved primarily in anthocyanin and flavonoid acylation, is only detected in the Conservative Set within seed plant genomes. This observation is congruent with emergence of the anthocyanin biosynthetic pathway – in which BAHDs catalyze the last decoration steps – in seed plants, primarily angiosperms (Piatkowski et al., 2020) and agrees with our previous OG-based inference (Kruse et al., 2022; Moghe et al., 2023). Clade 2, involved in wax biosynthesis, is seen in the Conservative Set-in only angiosperm plant clades. Clade 3, involved in acylation of diverse chemical scaffolds such as sugars, flavonoids and alkaloids was restricted to dicots, with no high-confidence hits found in other species groups. The absence of clade 3 in other plant genomes than eudicots suggests that Clade 3 is primarily a dicot-specific innovation. However, additional sampling of plant genomes from monocots, outside of Poaceae, and further early diverging eudicots would be needed to confirm Clade 3 exclusivity to dicot plants. Interestingly, Clade 4, associated with amine acylation, was found to be expanded in monocot grasses compared to the sampled dicot species, possibly reflecting the high prevalence of phenolamides (hydroxycinnamic acid amides) in grasses (Peng et al., 2016; Roumani et al., 2021). Surprisingly, no members of this clade were detected in the gymnosperm *Pinus taeda* (loblolly pine), beets, poplar and Arabidopsis, the significance of which is not clear. Clades 5 and 6, involved in aromatic alcohol acylation (e.g. in phenylpropanoid biosynthesis) and aliphatic alcohol/terpene acylation are present in all land plants, consistent with the role of these building blocks in the conquest of land. As suggested by

10.3389/fpls.2023.1067613

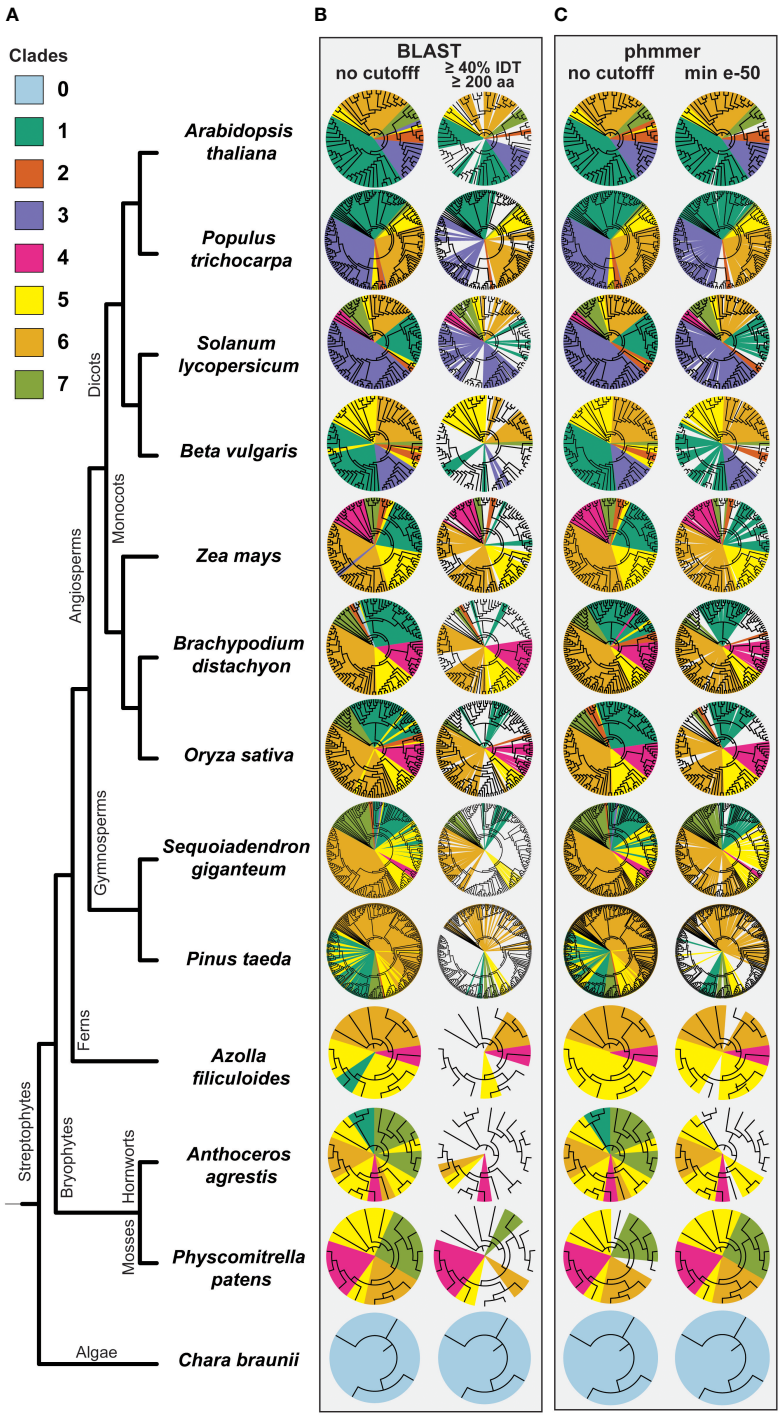


FIGURE 2
Phylogenetic distribution of BAHD clades in selected plant species. **(A)** Representative species were selected from early diverging plant lineages, gymnosperms, monocots, and dicots. Clade colors correspond to previously published clade assigne-ments of characterized BAHD members (Kruse et al, 2022). BAHD sequences were assigned to clades using **(B)** blastp with either no cutoff or a 40% IDT, 200 aa match length cutoff and **(C)** using the software phmmer with no cutoff or a minimum e-value of e-50. Each phylogenetic tree was inferred using IQTree with 1000 non-parametric bootstrap replicates, illustrated in iTOL (Letunic and Bork, 2021) and mapped clades were indicated using different colors.

previous studies, most of the biochemical and genetic pre-requisites for the evolution of true lignin and other derivatives of the general phenylpropanoid pathway were already present in the common ancestor of land plants (Weng and Chapple, 2010; Espiñeira et al., 2011; Renault et al., 2019; Kriegshauser et al., 2021; Rencoret et al., 2021). Our previous results (Kruse et al., 2022) also suggested that the ability to produce caffeoylquinate – the hydroxycinnamoyl CoA quinate transferase (HQT) activity – exists in charophytic algae and likely existed in the ancestor of land plants.

We also surveyed BAHDs in sugar beet, which produces betalain alkaloids. Betalains are produced due to a diversion of flux from arogenate, away from flavonoid and anthocyanin production

(Lopez-Nieves et al., 2018; Timoneda et al., 2019). Thus, we expected Clade 1, involved in anthocyanin and flavonoid acylation, to have contracted in that genome. In contrast, in poplar and giant sequoia – both woody tree species – we expected BAHDs involved in lignin production (clade 5) and generally, aromatic alcohol acylation, associated to terpenoid production to have increased in number. No such trends were observed (Figure 2).

The Conservative Set also revealed potential novel clades in bryophyte genomes (Figure 1B). The sampled bryophytes have 21 and 15 BAHDs, but the *in vitro* activities and physiological roles of most of these novel BAHDs have not been characterized. Such unassigned clades are also seen in other non-seed plant genomes and likely reflect the relative disparity in BAHD research in these clades. While it is possible that enzymes belonging to these clades have the same or similar activities to already known BAHD clades, at the sequence level, they have diverged substantially from any characterized BAHD enzyme to be assigned to a known clade using our clade-assignment approaches and thresholds. Such diverged clades are also seen in every other angiosperm and gymnosperm species including *A. thaliana*, highlighting the continuous functional innovation that occurs *via* duplication in this family.

3.3 Multiple motifs exposed to substrate binding cavities are enriched in BAHD clades

Identification of unique motifs may help in functional prediction of BAHDs. For example, DFGWG and HXXXD are two distinguishing and functionally important motifs of the family, and different variations of the adjoining residues appear in different clades (Figure 3). These motifs are structurally and catalytically important, respectively, with the His residue playing a key role in catalysis. To extend such insights, we asked if specific motifs were enriched among individual BAHD clades, using the sequences of biochemically characterized BAHDs in that clade vs. all other enzymes not in that clade. To ensure enough sample size for this discriminant analysis, we boosted the numbers of the BAHDs wherever required using orthologous groups of those enzymes as previously defined (Kruse et al., 2022). Several clade-specific motifs were discovered.

For each enriched motif, we asked if it was structural or likely-important for substrate binding/catalysis, based on whether the side chain of at least one residue in the motif was exposed to the acceptor/donor binding pocket. Clade 1 enzymes are prominently involved in malonyltransferase reactions whereas all other clades typically catalyze acyltransfer using aromatic (coumaryl, feruloyl, benzoyl CoA) or aliphatic (C2-C12 carbon CoA) donors. We found that 3/5 enriched motifs in this clade were located exposed to the donor CoA binding pocket. The motif TFFDXXW was also found to be enriched. Through site-directed mutagenesis and molecular dynamic simulations, we previously identified the role of the Trp residue in positioning the anthocyanidin core for acylation (Kruse et al., 2022). Another motif YFGNC, which is enriched in subclade 1a/b involved primarily in anthocyanin/flavonoid acylation (Supplementary Figure 2), was not found to be differentially enriched when assessing clade 1 as a whole, likely due to substantial functional divergence of clade 1c/d in comparison to the anthocyanin/flavonoid acylating enzymes of clade 1a/b. Aligning sequences belonging to clade 1a/b with clade 1c and 1d shows that the Cys residue, important for anthocyanin/flavonoid activity (Kruse et al., 2022), does not occur in 1c and 1d, suggesting this residue's close association with the anthocyanin/flavonoid acylating activity (Supplementary Figure 2).

In contrast to Clade 1, Clade 2 enzymes involved in wax biosynthesis had 4/5 enriched motifs located exposed to the *acceptor* binding cavity, of which one also extended in the donor binding site. This is a unique clade involved in long-chain fatty acid/alcohol acylation, for which no biochemical *in vitro* activities are available. Due to the unique hydrophobic nature of their acceptor substrates, it is likely that these motif changes reflect the acceptor site remodeling that may have occurred in these enzymes.

Clade 3 is a multifunctional, rapidly diverging clade (D'Auria, 2006; Tuominen et al., 2011; Kruse et al., 2022; Moghe et al., 2023). We see only one motif exposed to the *acceptor* binding site. The rapid divergence of these enzymes – as previously seen by the long branch lengths in a tree of characterized BAHDs (Kruse et al., 2022) – may have led to lack of any commonly enriched motifs in this clade. In contrast, Clade 4 (amine acylation) and Clade 5 (aromatic alcohol acylation), contain three and four motifs respectively that are exposed to the acceptor or donor binding pockets. Activities of Clade 6 are, overall, very similar to Clade 5 and both clades appear across all land plants (Figure 2). However, no enriched motif was found exposed to the acceptor/donor binding sites. Most of the enriched motifs in Clade 6 are likely structural in nature.

In Clades 4,5 and 6, different variations of the SXXD motif were differentially enriched. This motif is not in the active site or exposed to the substrate binding pockets but is part of a helix. The role of this motif is not clear, however, given its position and proximity to loops, we postulate that this motif acts as a hinge and may be involved in allosteric movement of the protein upon donor CoA binding, influencing the specificity of the acyltransferase reaction. Indeed, a previous study (Levsh et al., 2016) identified AtHCT Arg356 – which is just 4 aa away from the enriched SXXD motif in AtHCT – as an important driver of substrate selectivity in Clade 5 HCT enzymes. The entire motif, however, may play a key role in this allosteric movement.

This analysis was restricted at the clade level to obtain enough sample size for performing the discriminant analysis. It is possible that sub-clade-specific analyses – in some sub-clades such as clades 1a/b and 5a (Supplementary Figure 3), or clades 6c, 6a – may reveal additional function-specific motifs. Nonetheless, the motifs identified in this analysis that are exposed to the acceptor or donor sites are attractive targets for site-directed mutagenesis to enable activity engineering of these enzymes.

While sequence level information can provide insights on the ability of an enzyme to use a given substrate class, the actual substrates being used depends on the cellular localization of the enzymes. While sub-cellular and cell-type specific data is limited, analysis of condition-wise and organ-wise expression data can provide independent insights about BAHD evolution. Therefore, we first explored previously compiled expression data in rice and Arabidopsis to answer questions regarding evolution of paralog BAHD expression profiles.

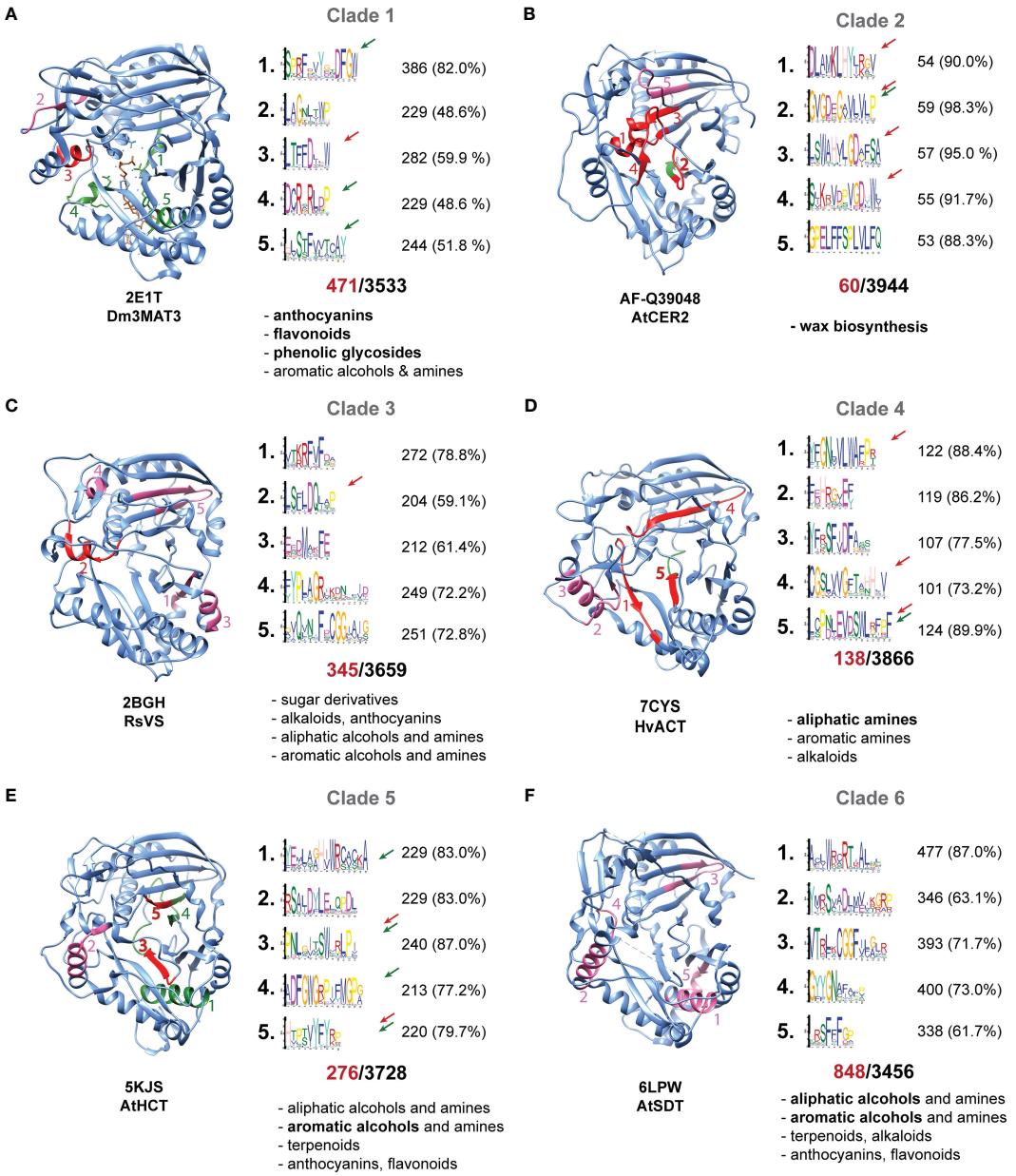


FIGURE 3
Clade-specific enriched motifs. **(A–F)** Motifs that are enriched in the sequences in the OGs of the respective clade, in comparison to all other OGs identified across 52 representative plant genomes. Motifs were identified using STREME in discriminatory mode. Only the top 5 enriched motifs are shown, and their locations on the protein structures are highlighted. Clade 6 motif locations are best guesses since some motifs are not found in AtSDT. Each protein structure is positioned with the acceptor binding pocket in the foreground and donor CoA binding pocket in the back. Motifs that are exposed to the acceptor side are highlighted with a red arrow and colored red within the structure. Motifs exposed to the donor side indicated are indicated with a green arrow and colored green. Structural motifs are colored in pink. For each clade, the number in red below the motifs indicates the number of sequences found in the corresponding OGS of that clade, and the black number indicates the number of sequences used as background. The number and percentage of sequences with the motif are given for each motif. All known substrate classes of each clade are noted, according to the nomenclature introduced in Kruse et al., 2022, with the most typical structural class bolded.

3.4 Most BAHD paralogs have diverged substantially in their expression patterns in Arabidopsis and rice

Diversification of an enzyme family can occur both at sequence and expression level. To determine how annotated BAHDs are different in their functions, we studied normalized RNA-seq data from previously published studies in *Oryza sativa* (rice) and *Arabidopsis thaliana* (Arabidopsis). In the latter, 18,916 expression data points for each of the 64 BAHDs were used to calculate pairwise Pearson's Correlation Coefficients (PCC), while in rice, 136 data points were used for 115 BAHDs. In both species, most BAHDs are uncorrelated with each other, however, multiple pairs were found to be significantly positively and negatively correlated with each other, at >95th and <5th percentile respectively of the overall correlation distribution (Figures 4A, B), highlighted in blue and red, respectively). We asked if these highly correlated genes tend to be recent gene duplicates. Mapping the largest correlated cluster onto the gene tree did not suggest any specific

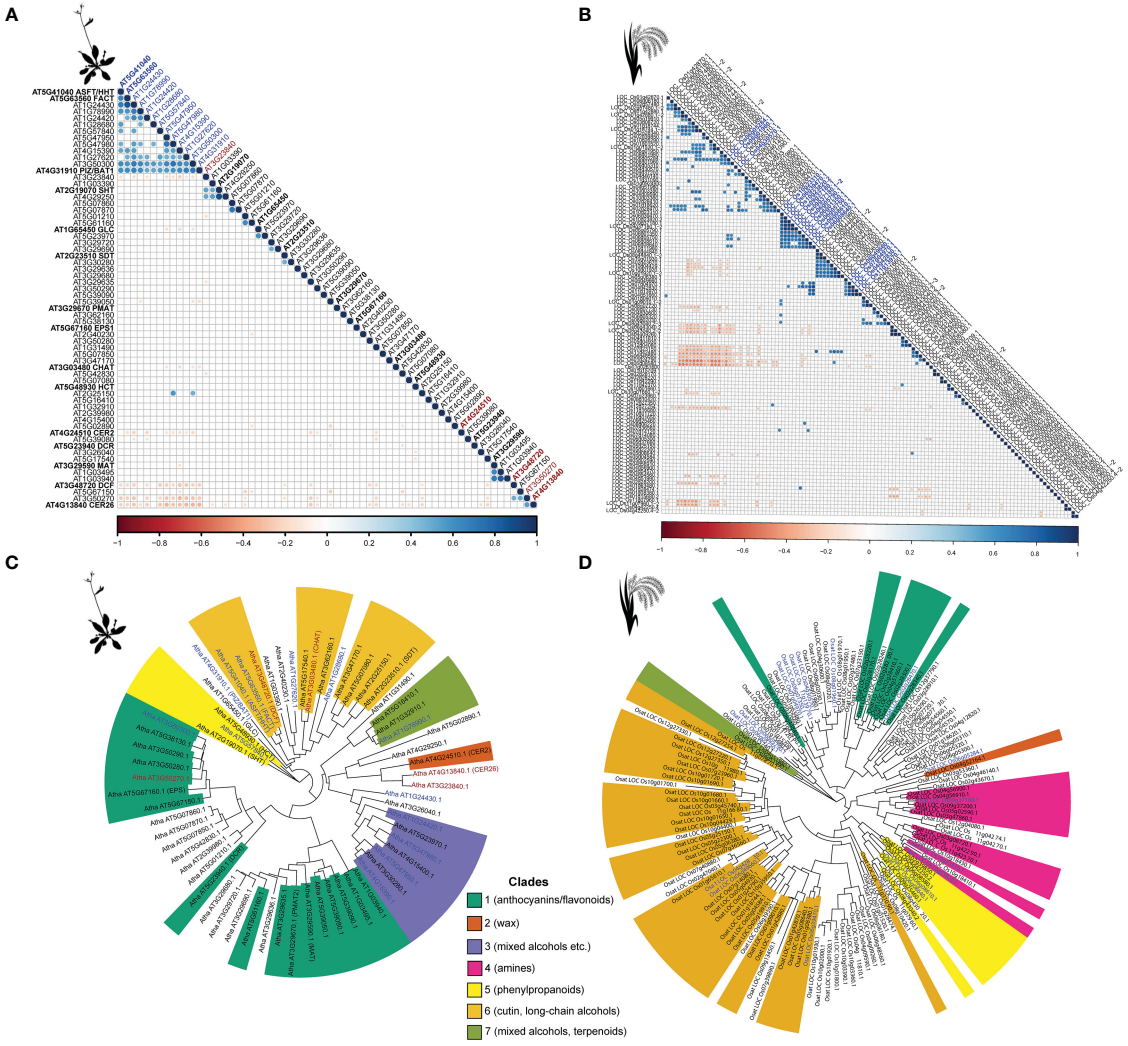


FIGURE 4
Co-expression of BAHDs in rice and Arabidopsis. **(A)** Co-expression matrix of Arabidopsis BAHDs. **(B)** Co-expression matrix of rice BAHDS. **(C)** Maximum-likelihood tree of Arabidopsis BAHDs. **(D)** Maximum-likelihood tree of rice BAHDS. Trees were inferred using IQTREE with 1000 bootstrap replicates and different colors represent different clades that were mapped using BLASTP (see Methods). Blue colored sequences were found to be positively coexpressed (PCC >= 95th percentile) and sequences colored in red are strongly negatively correlated (PCC <= 5th percentile). Sequences highlighted in bold in **(A)** have functional annotations. Clade nomenclature corresponds to the nomenclature introduced in Kruse et al., 2022 and typical substrate classes (but not all) are noted for each clade.

clustering in Arabidopsis (Figures 4C, D), however, five pairs of recently duplicated genes were found to be correlated in rice. These genes belonged to aromatic alcohol acylating (1 pair), amine acylating (1 pair) and rice-specific clades with unknown function (3 pairs). Synonymous substitution rate (Ks) between BAHDs, which is a proxy for their time since duplication, did not explain the variation in PCC (R^2: 0.04 and -0.0002 in rice and Arabidopsis, respectively) (Supplementary Figures 4A, D). Similarly, non-synonymous substitution rate (Ka) and the Ka/Ks also did not explain the variation in PCC (Supplementary Figures 4B–F). This was likely due to an overabundance of highly diverged paralogs biasing the regression (Supplementary Figure 4). Splitting paralogs into Ks bins revealed that only the most recent BAHD paralogs (0<Ks<0.2) in rice were significantly more co-expressed than BAHDs in other bins (Kolmogorov-Smirnov test, corrected p-value < 0.06, Supplementary File 1). This trend was not observed in Arabidopsis due to lack of paralogs in this Ks bin. In both species, a trend quickly became undetectable beyond Ks>0.4, suggesting that the regulation of BAHD expression changes rapidly after duplication (Figures 5A, B), corroborating earlier studies on expression divergence of gene duplicates (Ganko et al., 2007; Renny-Byfield et al., 2014). These results suggest that co-expression between BAHDs may rarely be informative of function.

3.5 Predicting BAHD functions using co-expression information

Above correlation analysis showed that most BAHDs have diverged significantly in their expression patterns from their paralogs. In our previous study (Kruse et al., 2022), we used orthology as a means to predict *in vitro* substrate utilization patterns of BAHDs and demonstrated that this approach successfully predicted correct substrate classes in 80-92% instances where a class could be assigned. To determine if co-expression with other genes can be used to obtain

10.3389/fpls.2023.1067613

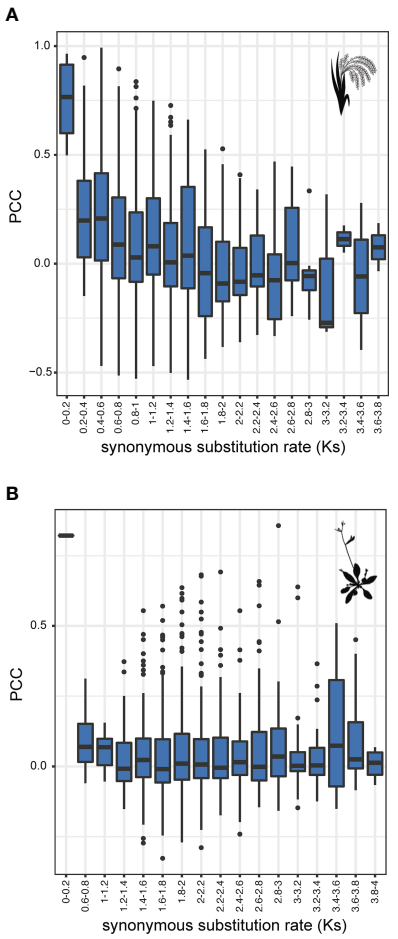


FIGURE 5
Relationship between synonymous substitutions (Ks) and expression correlation of BAHD paralogs. Expression correlation (PCC) versus synonymous substitution rate (Ks) of **(A)** rice and **(B)** Arabidopsis BAHDs. No significant differences were found when comparing across the entire dataset (Kruskal-Wallis rank sum test p=0.121 and p=0.494, respectively). However, comparing individual Ks bins for rice using pairwise Kolmogorov-Smirnov tests revealed significant differences between bin 0-0.2 in several bins up to bin 2.6-2.8 (significance level 0.05).

orthogonal evidence for *in vivo* function and help prioritize candidates for downstream experimental analyses, we investigated co-expression patterns of Arabidopsis BAHDs.

We first identified all co-expressed genes for each BAHD using ATTED – a pre-compiled database containing co-expression data for 18,957 Arabidopsis genes (Obayashi et al., 2022). In parallel, Plant Metabolic Network (PMN) pathway database was used to map all Arabidopsis genes to specific metabolic pathways. Overall, 2714 Arabidopsis genes in the ATTED database were assigned to PMN pathways. For each BAHD, we first identified all co-expressed and not-co-expressed genes mapped to PMN pathways (see Methods), and then asked if any of the matching PMN pathways were enriched among the co-expressed genes (Fisher exact test, corrected p-value < 0.05). Of the 65 BAHDs in Arabidopsis, 33 (50.7%) were previously assigned to at least one pathway in PMN (Supplementary File 2) and enrichment test could be performed for 33/65 BAHDs that were present in both ATTED and PMN databases and had highly co-expressed genes. Of these, 27 BAHDs had at least one enriched PMN metabolic pathway. We note that co-expression results may not necessarily identify the pathways a given BAHD is involved in, for example, because the actual pathway is not known, not enough/too many enzymes are mapped to the given pathway to reveal significant co-expression signal or because the biological phenomenon where the BAHD is expressed in may result in many co-expressed pathways. Furthermore, spurious correlations due to above and/or other technical reasons may lead to misassigning of genes to metabolic pathways in which they are not involved. Therefore, the co-expression results were primarily used to identify the overall metabolic class or the biological process a given BAHD may be associated with, instead of the actual *in vivo* pathway or specific substrates/molecular function. These co-expression based predictions were then combined with orthology-based predictions obtained previously (Kruse et al., 2022) to obtain greater confidence in the functional predictions.

We first asked if the co-expressed processes of the known BAHDs matched expectation, followed by assessment of other BAHDs showing similar prediction patterns. Known enzymes involved in lipid and cuticular wax/suberin biosynthesis such as ECERIFERUM 2 (CER2), DEFICIENT IN CUTIN FERULATE (DCF), PERMEABLE LEAVES 3 (PEL3), REDUCED LEVELS OF WALL-BOUND PHENOLICS 1 (RWP1), FATTY ALCOHOL : CAFFEOYL COENZYME A ACYLTRANSFERASE (FACT) were correctly predicted as being involved in lipid metabolism pathways such as those of suberin, cutin, very long chain fatty acid and acylglycerol biosynthesis/degradation (Supplementary File 2). DCF and PEL3 were also assigned "aliphatic alcohols" as the *in vitro* substrate class. AT3G23840 and AT5G02890 (both unannotated) were also significantly co-expressed with genes mapped to lipid metabolism, and therefore may be involved in similar processes. The former was mapped to the orthologous group of AtCER2, providing further support to its prediction, while only co-expression-based result is available for the latter candidate. Another characterized enzyme ACETYL COA: (*Z*)-3-HEXEN-1-OL ACETYLTRANSFERASE (AtCHAT) (D'Auria et al., 2007a) – known to be involved in 3-hexen-1-yl acetate biosynthesis – was co-expressed with other enzymes in the same pathway as well as those in cytokinin and chlorophyll degradation, the significance of which is not known. AT5G17540 received similar *in vitro* substrate class annotations but did not have any co-expressed pathways. This enzyme was previously shown to influence brassinosteroid metabolism (Zhu et al., 2013).

AT3G29590 [MALONYL COA: ANTHOCYANIDIN 5-O-GLUCOSIDE-6''-O-MALONYLTRANSFERASE; At5MAT, (D'Auria et al., 2007b)] is involved in anthocyanin acylation and was correctly mapped to the same and related pathways using co-expression. Additionally, AT1G03940/3495 (both unannotated) were also significantly co-expressed with anthocyanin-biosynthetic enzymes. Based on our substrate class prediction algorithm (Kruse et al., 2022), these enzymes are predicted to use "anthocyanins/flavonoids/phenolic glycosides", which agrees with their co-expression patterns. *A. thaliana* PHENOLIC GLUCOSIDE MALONYLTRANSFERASE (AtPMAT1) is involved in acylation of phenolic glycosides and is considered a detoxification enzyme (Taguchi et al., 2010; Gan et al., 2021). Consistent with its role, it was predicted to use "flavonoid" class (which as per previous definition (Kruse et al., 2022) also includes phenolic glycosides) and was found to be co-expressed with not only other phenolic glucoside

pathway enzymes but also with several enzymes involved in flavonoid biosynthesis, detoxification of reactive carbonyls, glutathione-mediated detoxification, flavonoid biosynthesis and abscisic acid pathways (Supplementary File 2). Based on its phylogenetic position, its substrate class is predicted to be anthocyanins/flavonoids/phenolic glycosides, which also agrees with its role. No other BAHD, however, showed similar patterns.

Characterized genes such as AtPMAT2 (also involved in phenolic glycoside biosynthesis), AtBIA1/DRL1 (involved in brassinosteroid biosynthesis (Roh et al., 2012; Zhu et al., 2013), AT2G240230 (DRL1 homolog) could not be confirmed because they either did not have a significantly co-expressed PMN pathway, *in vitro* class prediction or both. The predictions for *A. thaliana* SPERMIDINE HYDROXYCINNAMOYLTRANSFERASE (AtSHT) (Grienenberger et al., 2009; Wang et al., 2021), were incorrect using both methods.

A. thaliana ENHANCED PSEUDOMONAS SUSCEPTIBILITY (AtEPS1) (Torrens-Spence et al., 2019) involved in salicylic acid metabolism, is highly co-expressed with glucosinolate-biosynthetic enzymes. This may be explained by the involvement of both compound classes in defense responses e.g. by salicylic acid inducing glucosinolate accumulation (reviewed in Halkier and Gershenzon, 2006; Textor and Gershenzon, 2009). Multiple BAHDs (11/33, 33%) were annotated in PMN to be involved in simple coumarins and chlorogenic acid biosynthesis. The co-expression analysis predicted defense response roles to many of these BAHDs (AT3G50280, AT5G67150, AT3G50270) due to their co-expression with flavonoid, glutathione-mediated detoxification, glucosinolate and jasmonate biosynthetic pathways (Supplementary File 2).

These results suggest – based on functional analysis of previously characterized enzymes – that combining co-expression with pathway models and *in vitro* activity based predictions can generate useful preliminary hypotheses about BAHD roles in specific metabolic pathways and/or biological processes. While not all co-expression-based predictions are accurate, combining them with orthology-based predictions can help increase confidence in the BAHD's biochemical function. Nonetheless, further wet-lab characterization is required to validate functional predictions of the yet-uncharacterized enzymes.

4 Discussion

Processes inherent in the evolution of enzyme families – gene duplication-divergence, promiscuity, allelic divergence – are some of the biggest drivers of metabolic diversification in plants (Weng et al., 2012; Weng, 2014; Copley, 2015; Moghe and Last, 2015; Copley, 2020; Copley, 2021). A better understanding of these processes can help improve models for functional prediction of enzymes involved in metabolism (de Crécy-lagard et al., 2022). In this study, we sought to determine how the large BAHD acyltransferase family has evolved in plants (Supplementary File 3), and whether there are any sequence and/or expression features that can aid functional prediction.

We found that only ~1-5 BAHDs may have existed in the common ancestor of land plants and algae but their numbers quickly increased in land plants (Kruse et al., 2022) with a concomitant change in gene structure, producing shorter coding sequences and typically fewer introns than algal sequences. It is not clear, however, what the ancestral state for the gene structure was. Both charophytic algae (*C. braunii, K. nitens*) show shorter introns and genetic loci than chlorophytic algae. Therefore, it is also possible that the intron size, locus length, CDS length and number of introns increased in chlorophytic algae. The significance of these differences is unknown, especially since no BAHD functions and structure-function studies have been reported from chlorophytic algae.

In land plants, our results show that BAHD expansions occurred differently in different species. For example, BAHDs involved in phenylpropanoid biosynthesis and aliphatic alcohol acylation are present across all land plants (and therefore likely ancestral). However, BAHDs orthologous to known amine acylating enzymes such as agmatine coumaroyltransferase and spermidine coumaroyltransferase – despite their orthologous group being present in all land plants –have expanded specifically in monocots (Poaceae) (Figure 2). Hydroxycinnamic acid amides (HCAAs) are known to be important in grasses for pest and pathogen defense as well as for maintaining cell wall integrity (Roumani et al., 2021). Homologs of the characterized N-acyltransferases are also over-represented among BAHDs in *Physcomitrella patens* (moss) – it would be interesting to assess what roles these enzymes play in mosses and whether they too have the N-acyltransferase activities. Similarly, Clade 3 – which is involved in multiple specialized metabolic pathways – is likely dicot-specific and has been crucial in evolution of new metabolic classes such as acylsugars, alkaloids such as capsaicin, vinorine and cocaine, triterpenoids e.g. thalianol and arabidiol, and several anthocyanin acylating activities. In addition, we identified several clades whose members have sufficiently diverged from experimentally characterized enzymes to be placed into known clades. This is especially true in ferns and bryophytes but also true in other species. Such unassigned clades, which could either be an artefact of our technical thresholds or could indeed represent novel BAHD activities, need to be prioritized for functional assays for further understanding of the roles BAHDs play in plant metabolism.

Differential motif enrichment analysis identified unique motifs in the active site, acceptor and donor binding regions, internal structural regions as well as external handles that likely alter protein structure upon donor binding. It needs to be noted that these are simply the enriched regions; the BAHD sequences experience a lot more sequence changes. Despite such perturbations, the overall activity – acylation – has remained essentially the same, pointing to the mutational robustness of the BAHD fold. Residues identified in this study serve as a starting point for more detailed structure-function studies and enzyme engineering in the BAHD family (Ben-Hur and Brutlag, 2006; Kruse et al., 2022).

In addition to sequence features, we asked if expression-based features could be used as predictive signals of function. Due to extensive duplications in the BAHD family, we first sought to understand how expression patterns change with time. Our results suggest that very young duplicates, as expected, have similar expression e.g. while similar expression in more recent rice paralogs (Ks<0.2) is common, there is no differentiation power left at Ks>0.2, indicating background levels of expression divergence are quickly reached in BAHD paralogs. Most characterized BAHDs show substantial substrate promiscuity when assayed *in vitro* (D'Auria, 2006; Kruse et al., 2022; Moghe et al., 2023), suggesting that other

mechanisms e.g., altering spatio-temporal expression may play a role in modulating the *in vivo* function of a duplicated BAHD by exposing the enzyme to a different metabolic microenvironment.

While expression similarity between BAHDs has little explanatory power in predicting functional similarity, co-expression with genes in other pathways is indicative of function but needs orthogonal evidences for greater confidence in the prediction. We used Arabidopsis as a test given the wealth of previously characterized enzymes that could be used to test the functional predictions. We found several examples in lipid, anthocyanin, phenylpropanoid and amide biosynthesis where co-expression yielded accurate pathways based on prior knowledge. Most of these predictions were correct, albeit at different levels of functional resolution. The co-expression analysis also yielded novel functional predictions that can be tested using experimental approaches. Such co-expression analysis coupled with orthology-based information to predict *in vitro* and *in vivo* functions, can be significantly more impactful in species where, unlike *Arabidopsis thaliana*, substantial molecular analysis is not possible or has not been carried out before.

Overall, our study provides new insights into the evolution of BAHD acyltransferases, and provides a template to improve BAHD functional annotations. With large enzyme families, selecting impactful targets to characterize and specific hypotheses to test is important. These specifications can help extend our knowledge to clades/parts of the family that are not significantly researched into, and therefore would enable the discovery of novel activities. The functional prediction pipeline outlined in this study – combining expression patterns, pathway knowledge and *in vitro* substrate class prediction – can help to significantly reduce the time to characterize unknown enzymes. Similar approaches can also be applied to other large enzyme families such as CYP450, methyltransferases or UDP glycosyltransferases, expanding our understanding of large enzyme family evolution and providing an impetus to their application in synthetic biology.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

LK and GM conceived the initial project idea. LK and GM conducted the final analysis and prepared the figures. BF and JC performed initial analysis and developed analysis pipelines. LK and GM wrote the manuscript. BF and JC provided comments, edited and approved the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Funding for this project was provided by the Deutsche Forschungsgemeinschaft (DFG) Fellowship #411255989 to LK and Cornell University to GM. Contribution of JC to this project was enabled by NSF-REU DBI-1850796 to Georg Jander (Boyce Thompson Institute).

Acknowledgments

We acknowledge the role of Cornell BioHPC in assisting with the computational infrastructure needed for this project's successful completion. We also thank Dr. Jessica Garzke for helpful discussions on statistics and the R scripts.

Conflict of interest

The 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1067613/full#supplementary-material

SUPPLEMENTARY FILE 1
Excel file containing the results of the PCC vs Ks analyses.

SUPPLEMENTARY FILE 2
Excel file containing the results of the co-expression prediction pipeline.

SUPPLEMENTARY FILE 3
Excel file providing links to the used genome sequence repositories.

References

Bailey, T. L. (2021). STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* 37, 2834–2840. doi: 10.1093/bioinformatics/btab203

Ben-Hur, A., and Brutlag, D. (2006). "Sequence motifs: Highly predictive features of protein function," in *Feature extraction: Foundations and applications studies in fuzziness and soft computing*. Eds. I. Guyon, M. Nikravesh, S. Gunn and L. A. Zadeh (Berlin, Heidelberg: Springer). doi: 10.1007/978-3-540-35488-8_32

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bontpart, T., Cheynier, V., Ageorges, A., and Terrier, N. (2015). BAHD or SCPL acyltransferase? What a dilemma for acylation in the world of plant phenolic compounds. *New Phytol.* 208, 695–707. doi: 10.1111/nph.13498

Copley, S. D. (2015). An evolutionary biochemist's perspective on promiscuity. *Trends Biochem. Sci.* 40, 72–78. doi: 10.1016/j.tibs.2014.12.004

Copley, S. D. (2020). Evolution of new enzymes by gene duplication and divergence. *FEBS J.* 287, 1262–1283. doi: 10.1111/febs.15299

Copley, S. D. (2021). Setting the stage for evolution of a new enzyme. *Curr. Opin. Struct. Biol.* 69, 41–49. doi: 10.1016/j.sbi.2021.03.001

D'Auria, J. C. (2006). Acyltransferases in plants: A good time to be BAHD. *Curr. Opin. Plant Biol.* 9, 331–340. doi: 10.1016/j.pbi.2006.03.016

D'Auria, J. C., Pichersky, E., Schaub, A., Hansel, A., and Gershenzon, J. (2007a). Characterization of a BAHD acyltransferase responsible for producing the green leaf volatile (Z)-3-hexen-1-yl acetate in arabidopsis thaliana. *Plant J.* 49, 194–207. doi: 10.1111/j.1365-313X.2006.02946.x

D'Auria, J. C., Reichelt, M., Luck, K., Svatos, A., and Gershenzon, J. (2007b). Identification and characterization of the BAHD acyltransferase malonyl CoA: anthocyanidin 5-O-glucoside-6''-O-malonyltransferase (At5MAT) in arabidopsis thaliana. *FEBS Lett.* 581, 872–878. doi: 10.1016/j.febslet.2007.01.060

de Crécy-lagard, V., Amorin de Hegedus, R., Arighi, C., Babor, J., Bateman, A., Blaby, I., et al. (2022). A roadmap for the functional annotation of protein families: A community perspective. *Database* 2022, baac062. doi: 10.1093/database/baac062

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Espiñeira, J. M., Novo Uzal, E., Gómez Ros, L. V., Carrión, J. S., Merino, F., Ros Barceló, A., et al. (2011). Distribution of lignin monomers and the evolution of lignification among lower plants. *Plant Biol.* 13, 59–68. doi: 10.1111/j.1438-8677.2010.00345.x

Fan, P., Miller, A. M., Liu, X., Jones, A. D., and Last, R. L. (2017). Evolution of a flipped pathway creates metabolic innovation in tomato trichomes through BAHD enzyme promiscuity. *Nat. Commun.* 8, 2080. doi: 10.1038/s41467-017-02045-7

Fan, P., Miller, A. M., Schilmiller, A. L., Liu, X., Ofner, I., Jones, A. D., et al. (2016). *In vitro* reconstruction and analysis of evolutionary variation of the tomato acylsucrose metabolic network. *Proc. Natl. Acad. Sci. U.S.A.* 113, E239–E248. doi: 10.1073/pnas.1517930113

Ganko, E. W., Meyers, B. C., and Vision, T. J. (2007). Divergence in expression between duplicated genes in arabidopsis. *Mol. Biol. Evol.* 24, 2298–2309. doi: 10.1093/molbev/msm158

Gan, S., Rozhon, W., Varga, E., Halder, J., Berthiller, F., and Poppenberger, B. (2021). The acyltransferase PMAT1 malonylates brassinolide glucoside. *J. Biol. Chem.* 296, 100424. doi: 10.1016/j.jbc.2021.100424

Grienenberger, E., Besseau, S., Geoffroy, P., Debayle, D., Heintz, D., Lapierre, C., et al. (2009). A BAHD acyltransferase is expressed in the tapetum of arabidopsis anthers and is involved in the synthesis of hydroxycinnamoyl spermidines. *Plant J.* 58, 246–259. doi: 10.1111/j.1365-313X.2008.03773.x

Halkier, B. A., and Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. *Annu. Rev. Plant Biol.* 57, 303–333. doi: 10.1146/annurev.arplant.57.032905.105228

Hawkins, C., Ginzburg, D., Zhao, K., Dwyer, W., Xue, B., Xu, A., et al. (2021). Plant metabolic network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *J. Integr. Plant Biol.* 63, 1888–1905. doi: 10.1111/jipb.13163

Kim, J., Kang, K., Gonzales-Vigil, E., Shi, F., Jones, A. D., Barry, C. S., et al. (2012). Striking natural diversity in glandular trichome acylsugar composition is shaped by variation at the Acyltransferase2 locus in the wild tomato *Solanum habrochaites*. *Plant Physiol.* 160, 1854–1870. doi: 10.1104/pp.112.204735

Kriegshauser, L., Knosp, S., Grienenberger, E., Tatsumi, K., Gütle, D. D., Sørensen, I., et al. (2021). Function of the HYDROXYCINNAMOYL-CoA:SHIKIMATE HYDROXYCINNAMOYL TRANSFERASE is evolutionarily conserved in embryophytes. *Plant Cell* 33, 1472–1491. doi: 10.1093/plcell/koab044

Kruse, L. H., Weigle, A. T., Irfan, M., Martínez-Gómez, J., Chobirko, J. D., Schaffer, J. E., et al. (2022). Orthology-based analysis helps map evolutionary diversification and predict substrate class use of BAHD acyltransferases. *Plant J.* 111, 1453–1468. doi: 10.1111/tpj.15902

Landis, J. B., Miller, C. M., Broz, A. K., Bennett, A. A., Carrasquilla-Garcia, N., Cook, D. R., et al. (2021). Migration through a major Andean ecogeographic disruption as a driver of genetic and phenotypic diversity in a wild tomato species. *Mol. Biol. Evol* 38, 3202–3219. doi: 10.1093/molbev/msab092

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* 49, W293–W296. doi: 10.1093/nar/gkab301

Levsh, O., Chiang, Y.-C., Tung, C. F., Noel, J. P., Wang, Y., and Weng, J.-K. (2016). Dynamic conformational states dictate selectivity toward the native substrate in a substrate-permissive acyltransferase. *Biochemistry* 55, 6314–6326. doi: 10.1021/acs.biochem.6b00887

Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.-M., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* 4, 460–472. doi: 10.1038/s41477-018-0188-8

Lopez-Nieves, S., Yang, Y., Timoneda, A., Wang, M., Feng, T., Smith, S. A., et al. (2018). Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in caryophyllales. *New Phytol.* 217, 896–908. doi: 10.1111/nph.14822

Milo, R., and Last, R. L. (2012). Achieving diversity in the face of constraints: Lessons from metabolism. *Science* 336, 1663–1667. doi: 10.1126/science.1217665

Moghe, G. D., and Kruse, L. H. (2018). The study of plant specialized metabolism: Challenges and prospects in the genomics era. *Am. J. Bot.* 105, 959–962. doi: 10.1002/ajb2.1101

Moghe, G., Kruse, L. H., Petersen, M., Scossa, F., Fernie, A. R., Gaquerel, E., et al. (2023). BAHD company: The ever-expanding roles of the BAHD acyltransferase gene family in plants. *Annu. Rev. Plant Biol.* 74:1. doi: 10.1146/annurev-arplant-062922-050122

Moghe, G. D., and Last, R. L. (2015). Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiol.* 169, 1512–1523. doi: 10.1104/pp.15.00994

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y., and Kinoshita, K. (2022). ATTED-II v11: A plant gene coexpression database using a sample balancing technique by subagging of principal components. *Plant Cell Physiol.* 63, 869–881. doi: 10.1093/pcp/pcac041

Peng, M., Gao, Y., Chen, W., Wang, W., Shen, S., Shi, J., et al. (2016). Evolutionarily distinct BAHD n-acyltransferases are responsible for natural variation of aromatic amine conjugates in rice. *Plant Cell* 28, 1533–1550. doi: 10.1105/tpc.16.00265

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Piatkowski, B. T., Imwattana, K., Tripp, E. A., Weston, D. J., Healey, A., Schmutz, J., et al. (2020). Phylogenomics reveals convergent evolution of red-violet coloration in land plants and the origins of the anthocyanin biosynthetic pathway. *Mol. Phylogenet. Evol.* 151, 106904. doi: 10.1016/j.ympev.2020.106904

Pichersky, E., and Raguso, R. A. (2018). Why do plants produce so many terpenoid compounds? *New Phytol.* 220, 692–702. doi: 10.1111/nph.14178

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

R Core Team (2021) *R: a language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing*. Available at: http://www.R-project.org/.

Renault, H., Werck-Reichhart, D., and Weng, J.-K. (2019). Harnessing lignin evolution for biotechnological applications. *Curr. Opin. Biotechnol.* 56, 105–111. doi: 10.1016/j.copbio.2018.10.011

Rencoret, J., Gutiérrez, A., Marques, G., del Río, J. C., Tobimatsu, Y., Lam, P. Y., et al. (2021)New insights on structures forming the lignin-like fractions of ancestral plants (Accessed December 7, 2022).

Renny-Byfield, S., Gallagher, J. P., Grover, C. E., Szadkowski, E., Page, J. T., Udall, J. A., et al. (2014). Ancient gene duplicates in *Gossypium* (Cotton) exhibit near-complete expression divergence. *Genome Biol. Evol.* 6, 559–571. doi: 10.1093/gbe/evu037

Roh, H., Jeong, C. W., Fujioka, S., Kim, Y. K., Lee, S., Ahn, J. H., et al. (2012). Genetic evidence for the reduction of brassinosteroid levels by a BAHD acyltransferase-like protein in Arabidopsis1[W][OA]. *Plant Physiol.* 159, 696–709. doi: 10.1104/pp.112.197202

Roumani, M., Besseau, S., Gagneul, D., Robin, C., and Larbat, R. (2021). Phenolamides in plants: An update on their function, regulation, and origin of their biosynthetic enzymes. *J. Exp. Bot.* 72, 2334–2355. doi: 10.1093/jxb/eraa582

Schilmiller, A. L., Moghe, G. D., Fan, P., Ghosh, B., Ning, J., Jones, A. D., et al. (2015). Functionally divergent alleles and duplicated loci encoding an acyltransferase contribute to acylsugar metabolite diversity in *Solanum* trichomes. *Plant Cell* 27, 1002–1017. doi: 10.1105/tpc.15.00087

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc B.* 64, 479–498. doi: 10.1111/1467-9868.00346

Taguchi, G., Ubukata, T., Nozue, H., Kobayashi, Y., Takahi, M., Yamamoto, H., et al. (2010). Malonylation is a key reaction in the metabolism of xenobiotic phenolic glucosides in arabidopsis and tobacco. *Plant J.* 63, 1031–1041. doi: 10.1111/j.1365-313X.2010.04298.x

Textor, S., and Gershenzon, J. (2009). Herbivore induction of the glucosinolate–myrosinase defense system: Major trends, biochemical bases and ecological significance. *Phytochem. Rev.* 8, 149–170. doi: 10.1007/s11101-008-9117-1

Timoneda, A., Feng, T., Sheehan, H., Walker-Hale, N., Pucker, B., Lopez-Nieves, S., et al. (2019). The evolution of betalain biosynthesis in caryophyllales. *New Phytol.* 224, 71–85. doi: 10.1111/nph.15980

Torrens-Spence, M. P., Bobokalonova, A., Carballo, V., Glinkerman, C. M., Pluskal, T., Shen, A., et al. (2019). PBS3 and EPS1 complete salicylic acid biosynthesis from isochorismate in arabidopsis. *Mol. Plant* 12, 1577–1586. doi: 10.1016/j.molp.2019.11.005

Tuominen, L. K., Johnson, V. E., and Tsai, C.-J. (2011). Differential phylogenetic expansions in BAHD acyltransferases across five angiosperm taxa and evidence of divergent expression among *Populus* paralogues. *BMC Genomics* 12, 236. doi: 10.1186/1471-2164-12-236

Wang, C., Li, J., Ma, M., Lin, Z., Hu, W., Lin, W., et al. (2021). Structural and biochemical insights into two BAHD acyltransferases (AtSHT and AtSDT) involved in phenolamide biosynthesis. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.610118

Weng, J.-K. (2014). The evolutionary paths towards complexity: A metabolic perspective. *New Phytol.* 201, 1141–1149. doi: 10.1111/nph.12416

Weng, J.-K., and Chapple, C. (2010). The origin and evolution of lignin biosynthesis. *New Phytol.* 187, 273–285. doi: 10.1111/j.1469-8137.2010.03327.x

Weng, J.-K., Philippe, R. N., and Noel, J. P. (2012). The rise of chemodiversity in plants. *Science* 336, 1667–1670. doi: 10.1126/science.1217411

Zhang, H., Zhang, F., Yu, Y., Feng, L., Jia, J., Liu, B., et al. (2020). A comprehensive online database for exploring ~20,000 public arabidopsis RNA-seq libraries. *Mol. Plant* 13, 1231–1233. doi: 10.1016/j.molp.2020.08.001

Zhu, W., Wang, H., Fujioka, S., Zhou, T., Tian, H., Tian, W., et al. (2013). Homeostasis of brassinosteroids regulated by DRL1, a putative acyltransferase in arabidopsis. *Mol. Plant* 6, 546–558. doi: 10.1093/mp/sss144

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 March 2023
DOI 10.3389/fpls.2023.1142840

OPEN ACCESS

EDITED BY
Wei Li,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Ke Wang,
Anhui Agricultural University, China
Juanjuan Yu,
Henan Normal University, China

*CORRESPONDENCE
Hu Zhao
zhaohu8196@sina.com

SPECIALTY SECTION
This article was submitted to Plant Metabolism and Chemodiversity, a section of the journal Frontiers in Plant Science

RECEIVED 12 January 2023
ACCEPTED 06 March 2023
PUBLISHED 20 March 2023

CITATION
Zhao H, Shen C, Hao Q, Fan M, Liu X and Wang J (2023) Metabolic profiling and gene expression analysis reveal the quality deterioration of postharvest toon buds between two different storage temperatures.
Front. Plant Sci. 14:1142840.
doi: 10.3389/fpls.2023.1142840

Metabolic profiling and gene expression analysis reveal the quality deterioration of postharvest toon buds between two different storage temperatures

Hu Zhao*, Cheng Shen, Qingping Hao, Mingqin Fan, Xiaoli Liu and Juan Wang

Biology and Food Engineering College, Fuyang Normal University, Engineering Technology Research Center of Anti-Aging Chinese Herb, Fuyang, Anhui, China

Toon buds, a popular woody vegetable, contain large amounts of nutrients. However, toon buds have strong respiratory metabolism after harvest and are highly prone to decay, resulting in quality deterioration. Low temperature can effectively inhibit postharvest senescence of toon buds. GC-TOF-MS combined with quantitative real-time PCR was used to elucidate the toon bud deterioration mechanism after harvest by analyzing the difference in the relative contents of primary metabolites and their derivatives, and the expression of key genes associated with metabolic pathways in toon buds between low temperature and room temperature storages for 72 h. Results showed that the ethylene synthesis in toon buds accelerated under room temperature storage, along with significant changes in the primary metabolic pathway. The catabolism of amino acids, fatty acids, and cell membrane phospholipids was accelerated, and the gluconeogenesis synthesis was strengthened. Moreover, the sucrose synthesis was increased, the glycolysis and TCA cycle were broken down, and the pentose phosphate pathway was vigorous. As metabolic intermediates, organic acids were considerably accumulated. Moreover, varieties of toxic compounds were produced in parallel with the activation of aromatic compounds. This work provided a comprehensive understanding of the metabolic regulation, thereby revealing how low and room temperatures differentially influenced the quality deterioration of postharvest toon buds.

KEYWORDS

metabolomics, gene expression, quality deterioration, storage temperature, *toona sinensis*

Highlights

- Ethylene production and respiration rate in postharvest toon buds were accelerated.
- A total of 305 metabolites were detected in the metabolic data.
- Sucrose and organic acids were accumulated in postharvest toon buds.
- Catabolism of amino acids was strengthened in postharvest toon buds.
- Some key genes were differentially expressed between two storage temperatures.

1 Introduction

Toona sinensis, also called Xiangchun in Chinese, is an important woody vegetable widely planted in Asian countries. In food production, toon buds are used as raw materials for sauce and functional food by consumers because of their bright color, distinctive flavor, and multiple nutrients (Jiang et al., 2019; Su et al., 2020).

As a woody vegetable, toon buds are susceptible to decay and quality deterioration during postharvest storage because of their high respiration rate and water content (Zhao et al., 2018; Lin S. H. et al., 2019). Many preservative methods, such as modified atmosphere, film coating, and various chemical reagent treatments, have been employed to prolong the shelf life of toon buds. However, cold storage has been considered as an effective and economical method (Zhang et al., 2009; Yang et al., 2011; Zhu and Gao, 2017). LT contributes to decreasing respiration, slowing metabolism and maintaining the quality of toon buds (Wang et al., 2019). However, the quality of toon buds inevitably deteriorates with the increasing storage time (Hu et al., 2019). Nevertheless, two storage temperature models, i.e., room temperature (RT) and LT, provide an ideal model for deeply analyzing the molecular mechanism of quality deterioration of toon buds. Previous studies have mainly investigated the secondary metabolites in postharvest toon buds. These metabolites determine the color and flavor quality of toon buds (Yang et al., 2011; Yang et al., 2019). After cold storage, the total content of flavonoids and volatile terpenoids and their oxidates increase significantly with prolonged storage time (Zhao et al., 2019; Zhao et al., 2021). However, how the main carbohydrates, organic acids, amino acids, fatty acids, and their metabolic mechanism change in postharvest toon buds during storage remains largely unknown.

Recently, a large number of literatures have reported the physiological, biochemical, and molecular regulatory mechanisms of quality deterioration and postharvest senescence of fruits and vegetables (Tang et al., 2016; Guo et al., 2019; Pott et al., 2020). Tang et al. (2016) revealed that RT storage enhanced ABA and ethylene signaling pathways *via* the upregulation of *PYLs*, *ABI5*, and *ERFs* on Powell orange pulp senescence. The RT-stored upregulated genes were involved in primary metabolism including sucrose metabolism, glycolysis, gluconeogenesis, and fermentation pathways, resulting in declining levels of sucrose and organic acids, such as malate, citrate, and α-ketoglutaric acid, and accumulated hexoses on Powell orange pulp. Postharvest fruit respiration directly affects primary metabolic pathways, including glycolysis and the tricarboxylic acid cycle (TCA), which account for changes in sugar, amino acid, and organic acid levels (Zhang et al., 2011; Goulas et al., 2015). Guo et al. (2019) reported that the metabolic pathways related to carbohydrates, organic acids, and amino acids might be highly active in RT-stored litchi pulps, whereas the metabolic pathways related to aliphatic metabolites and nucleotides might be highly active in LT-stored litchi pulps. The genotype tomato with malate dehydrogenase (MDH) deficiency showed higher malate content and poorer postharvest behavior than non-transformed fruits (Osorio et al., 2019). This finding indicates malate's role in postharvest responses to RT storage. In the present study, the metabolites in the toon buds at harvest (0 h) were determined through gas chromatograph coupled with a time-of-flight mass spectrometer (GC-TOF-MS) after storage at RT and LT for 12, 24, 48, and 72 h. The DEMs of toon buds after 12, 24, 48, and 72 h of storage at RT and LT were identified compared with those at harvest 0 h (control). Furthermore, the differential expression of the critical genes involved in the primary metabolism and ethylene signaling pathway was revealed between two storage temperature models. In summary, our study shed light on the differential metabolites and candidate genes associated with the quality deterioration of toon buds in both storage models.

2 Materials and methods

2.1 Plant materials

Freshly harvested toon buds of the 'Heiyouchun' cultivar were obtained from the *T. sinensis* nursery base in Xin Town of Taihe County in Anhui Province, in April 2021. More than 300 toon buds with uniform color and size but without visual blemishes and mechanical damage were selected. The harvested toon buds collected in ice boxes were immediately transported to the laboratory (control). Then, they were randomly divided into two groups. One group was placed in a hermetic plastic container at RT (20°C ± 0.5°C) with a relative humidity 80%–90%. In comparison, the other group was stored in a refrigerator (4°C ± 0.5°C) for the LT storage with the same humidity. The sampling time points at both treatments were set at 12, 24, 48, and 72 h after storage. Six biological replicates were designed with five toon buds each. The samples were collected at each time point for respiration rate and ethylene content measurements. The quick-freeze samples treated with liquid nitrogen were stored at −80°C and employed for subsequent metabolites and RNA extraction.

2.2 Toon bud appearance evaluation

Toon bud appearance was evaluated as previously described by Zhao et al. (2018). The decay symptoms after storage were visually determined as injury ranks using a five-grade marking system based on the percentage of brown leaves or shoots: 0 = intact buds without brown or rotten tissue; 1 = 1% to 25% of the bud-damaged tissue; 2 = 26% to

50% of the bud-damaged tissue; 3 = 51% to 75% the bud-damaged tissue; 4 ≥ 76% of the bud-damaged tissue. The decay index (DI) was calculated following the formula described by Zhao et al. (2018).

2.3 Determining respiration rate and ethylene production

Respiration rates were measured according to the modified small-skep-method (MSSM) (Li et al., 2015). Three toon buds with uniform size from each sample were kept in a 0.5-l airtight wide-mouth bottle. The released CO_2 from postharvest toon buds was absorbed using Ba $(OH)_2$ solution, and the residual $Ba(OH)_2$ was titrated with oxalic acid solution. The amount of CO_2 could be calculated from the difference between the oxalic acid solution consumed by the blank and the samples. The result was expressed as mg kg^{-1} h^{-1} of CO_2. Ethylene production was measured *via* gas chromatography (GC). The samples were placed in 0.5-l airtight plastic bottles (each containing three toon buds) for five time points at both storage temperatures. Then, 0.1 ml of the headspace gas was injected into an Agilent 7890A GC (Agilent, Santa Clara, CA, USA) equipped with a HP-5MS packed column and a flame ionization detector. The amount of ethylene production was calculated from a calibration curve of standard ethylene gas and expressed as ng kg^{-1} s^{-1} of fresh weight.

2.4 Non-target GC–TOF-MS analysis

2.4.1 Sample preparation and extraction

Hydrophilic metabolites were extracted from 50 mg of the powered toon buds by adding 500 µl of 3:1 (v/v) methanol: ddH_2O as a precold extraction mixture. Then, 10 µl ribitol (0.5 g l^{-1} stock solution) was added to the extraction mixture as the internal standard (IS). The extraction was mixed using a thermomixer compact (Eppendorf AG, Germany) for 30 s. A steel ball was added to the extraction to extract the metabolites fully. Moreover, the sample was treated with a 45-Hz grinding instrument for 4 min and ultrasonicated with an ice water bath for 5 min (repeated three times). After centrifugation at 4°C for 15 min at 12,000 rpm, 300 µl supernatant was transferred to a fresh tube. Then, 100 µl of each sample was collected, pooled, and evaporated in a vacuum concentrator to prepare the quality control (QC) sample. The dried samples were redissolved in 80 µl of 20 g l^{-1} methoxyamine hydrochloride in pyridine and incubated at 80°C for 30 min while shaking. Then, derivatization of the mixture was achieved through incubation with 100 µl of BSTFA regent (1% TMCS, v/v) at 70°C for 1.5 h. After the samples were gradually cooled to RT, 5 µl of FAMEs (in chloroform) was added to the QC sample. Then, all samples were subjected to Agilent 7890 GC coupled with a time-of-flight mass spectrometer (GC-TOF-MS) for analysis.

2.4.2 Injection parameters on GC-TOF-MS

A volume of 1 µl aliquot of derivatized sample was injected in splitless mode and separated on a 30 m × 0.25 mm DB-5MS capillary column coated with a 0.25-µm CP-Sil 8 CB low bleed (Varian Inc., Palo Alto, CA, USA). Helium was used as the carrier gas. The front inlet purge flow was 3 ml min^{-1}. The gas flow rate through the column was 1 ml min^{-1}. The initial temperature was kept at 50°C for 1 min. Then, it was raised to 310°C at a rate of 10°C min^{-1} and kept for 8 min at 310°C. The injection, transfer line, and ion source temperatures were 280°C, 280°C, and 250°C, respectively. The energy was −70 eV in the electron impact mode. The mass spectrometry data were acquired in the full-scan mode with the 50–500-m/z range at 12.5 spectra per second after a solvent delay of 6.27 min.

2.4.3 Metabolite analysis by GC-TOF-MS

Raw data analysis, including peak extraction, baseline adjustment, deconvolution, alignment, and integration, was completed with ChromaTOF (V 4.3x, LECO) software. The LECO-Fiehn Rtx5 database was used for metabolite identification by matching the mass spectrum and retention index (Kind et al., 2009). Finally, the peaks detected in less than half of the QC samples or relative standard deviation (RSD) in>30% of the QC samples was removed (Dunn et al., 2011).

2.5 RNA isolation and quantitative real-time PCR analysis

The relative expression levels of the genes selected were analyzed following the method by Zhao et al. (2017) with slight modifications. The total RNA was isolated from 100 mg of toon bud samples with RNAprep Pure Plant Plus Kit (polysaccharide- and polyphenolics-rich) (Tiangen, Beijing). Then, it was reversely transcribed into single cDNA at 42°C for 15 min by using FastKing gDNA Dispelling RT SuperMix (Tiangen, Beijing) according to the manufacturer's protocols. Then, the cDNA was diluted 10-fold as templates for the quantitative real-time PCR (qRT-PCR) analysis *via* SuperReal PreMix Plus (SYBR Green) (Tiangen, Beijing) performed by the Bio-Rad, CFX96 Connect Optics Module Real-Time System (Thermo, CA). The *actin* gene of *T. sinensis* was used as the internal control for evaluating the relative expression of the specific genes. The gene-specific primers used are listed in Supplementary Table S1 (Zhao et al., 2017). The expression levels of critical genes involved in the primary metabolism and ethylene signaling were analyzed by qRT-PCR. The $2^{-\triangle\triangle}$Ct method was applied to calculate the relative expression levels of these genes (Zhao et al., 2017). Each sample also contained three biological replicates and three technological replicates. The relative expression levels were represented by mean ± standard error of the mean (SEM) values (Zhao et al., 2017).

2.6 Statistical analysis

The data of metabolites presented in this study were the mean values of six biological replicates. The GC-TOF-MS data were processed with SIMCA software (Version 15.0, Umetrics, Umea, Sweden) to obtain the result of multivariate statistical analysis. Principal component analysis (PCA) was performed to compare the DEM level between both samples. Then, DEMs were identified and confirmed between the two samples using the orthogonal partial least square discriminant analysis (OPLS–

DA) with the threshold value of variable importance in projection (VIP) >1 and $P < 0.05$. Subsequently, the expression patterns of DEMs were demonstrated by hierarchical clustering analysis (HCA) and heat maps. The physiological data and qRT-PCR analysis were the mean values of three replicates. A two-way ANOVA was performed. The Fisher test ($P < 0.05$) was used for mean comparisons between the two samples using SPSS 22.0. The different letters between means were indicative of significant difference at a significant level of $P < 0.05$.

3 Results

3.1 Alleviation of postharvest senescence and rotting of toon buds by LT

Toon bud browning is a common problem affecting quality and consumer preference. The visual effects of LT storage on delaying toon bud senescence, browning, and rotting are shown in Figure 1A. The toon buds under LT storage still retained a green appearance after 72 h. In comparison, the toon buds stored in RT developed severe browning and rotting. As illustrated in Figure 1B, the DI of the LT-stored toon buds was 17.8% in 72 h, significantly lower than that of RT-stored toon buds (68.7%).

3.2 Inhibition of ethylene production and respiration rate in postharvest toon buds by LT

When the storage temperature regime was not considered during the whole storage period, the ethylene concentrations increased with prolonged storage time (Figure 2A). However, the ethylene release rate significantly differed between RT and LT storages. When toon buds were under LT conditions, the ethylene release rate was very low during the whole storage period. The ethylene release rate was only 2.34 ng kg^{-1} s^{-1} after 72 h of storage. However, the ethylene release rate increased exponentially after a relatively stable period of 24–48 h under RT storage. The ethylene release rate reached 39.90 ng kg^{-1} s^{-1} at the end of the storage period (Figure 2A).

Unlike the ethylene release rate, the respiration rate demonstrated various downward trends in both groups (Figure 2B). During 0–48 h of RT storage, the respiration rate of the LT storage was lower than that of the RT storage. The finding indicated that the respiratory consumption of toon buds was inhibited under LT conditions. After 72 h, the respiration rate in RT was lower than that in LT.

3.3 Identification of the metabolites in postharvest toon buds

A total of 305 metabolites were detected in the raw data, of which 150 matched the known metabolites in the database, including carbohydrates, organic acids, amines, aldehydes and alcohols, amino acids, fatty acids, alkaloids, pyridines and purines, and their derivatives. The remaining 155 metabolites were named as "unknown" or "analyte." Their structural features needed to be further identified (Supplementary Table S2). The correlation coefficients (six biological duplicate samples) within a group exceeded a threshold value (0.7) within the range of 0.77–0.98. This finding suggested that the data within the group showed good repeatability and could be used for subsequent differential metabolite screening (Supplementary Table S3).

3.4 Multivariate analysis of metabolomic data in postharvest toon buds

The metabolic data between groups (control and 12–24 h of LT-stored groups, 48 h of RT-stored and 72 h of RT-stored groups,

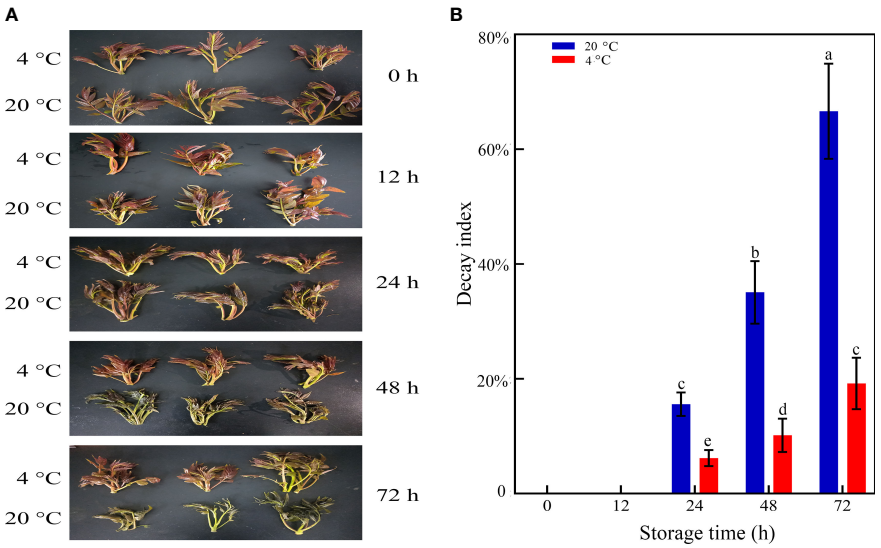


FIGURE 1
Appearance **(A)** and DI **(B)** of toon buds at 4°C and 20°C storage after 0, 12, 24, 48, and 72 h. The values are presented as the mean ± SEM. The bars with different lowercase letters are significantly different at $P < 0.05$.

10.3389/fpls.2023.1142840

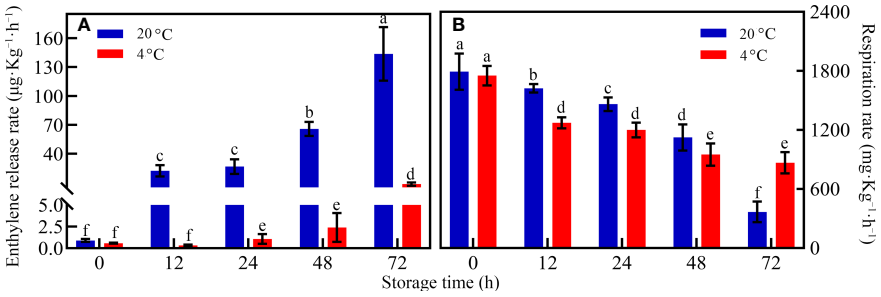


FIGURE 2
Change in the **(A)** ethylene release rate and respiration rate **(B)** of toon buds at 4°C and 20°C storage after 0, 12, 24, 48, and 72 h. The values are presented as the mean ± SEM. The bars with different lowercase letters are significantly different at $P < 0.05$.

48–72 h of LT-stored and 12–24 h of RT stored groups) had high correlation coefficients, indicating that these groups had similar metabolite accumulation patterns (Supplementary Figure S1). PCA was performed to reveal the distribution trends among various samples (Supplementary Figure S1). The differences between the samples could not be explained through the visual discrimination generated by the PCA. The orthogonal projections to latent structures-discriminant analysis (OPLS-DA) and permutation plots indicated the clear separation of the eight coupled treatments between control and other storage times (Figure 3). In general, an R2Y value of 0.65 or more and Q2Y of 0.5 or more indicated a satisfactory ability for quantitative prediction. High Q2, R2X, and R2Y values were observed in the comparison between control (0 h) and other samples. This observation indicated major time-dependent relationships in the metabolic profiles of postharvest toon buds.

3.5 DEMs in RT- and LT-stored toon buds compared with control

A differential metabolomics method based on VIP > 1 and $P < 0.05$ was used to screen the DEMs of postharvest toon buds. The total differential metabolites of LT vs. control, RT vs. control, and RT vs. LT at the same storage time point were 192 (Figures 4A, D), 234 (Figures 4B, E), and 277 (Figures 4C, F), respectively.

We built an HCA to unveil the diversity in the metabolite profiles among different experimental samples (Figure 5). Among the known differential metabolites, 16 carbohydrates (Figure 5A), 45 known acid metabolites (Figure 5B), 13 amino acids and their 18 derivatives (Figure 5C), 36 secondary metabolites (Figure 5D), and nine pyridines, pyrimidines, purines, and their derivatives (Figure 5E) were identified.

As shown in Figure 5A, the change patterns of various carbohydrates in toon buds under RT and LT storage were different. Sorbose, fructose, sucrose, lactose, D-talose, and maltose displayed greater upregulation in RT storage than in LT storage. The decrease in the relative contents of lysose, ribose, and fucose in toon buds during RT storage was more than that at LT storage. Digitoxose and 3,6-anhydro-D-galactose were upregulated under RT storage and downregulated under LT storage. D-Glucoheptose and isomaltose were downregulated under RT storage but not significantly.

As shown in Figure 5B, 16 organic acids related to glucose metabolism, including glucose 1-phosphate, glucose 6-phosphate, and 6-phosphogluconic acid, accumulated in different levels between LT and RT groups. The relative contents of inorganic sulfuric acid increased at RT, particularly after 48 h of storage. However, they decreased under LT storage. Phosphoric acid decreased rapidly during RT storage but changed slightly under LT storage.

Except for aspartic acid, glutamic acid, and glutamine, most amino acids and their derivatives exhibited a declining trend during storage. Moreover, the decrease in amino acid contents in the RT storage was greater than that in the LT storage. The relative contents of essential fatty acids such as palmitic acid and linolenic acid also showed a similar downward trend. However, the change in stearic acid was not remarkable during storage (Figure 5C).

Except for the primary metabolites, 36 secondary metabolites were classified into alcohol, phenol, aldehyde, amine, and their oxides (Figure 5D).

Given the catabolism of vitamins and uracil, 2-hydroxypyridine, 4-hydroxypyridine, and orotic acid accumulated significantly in toon buds during RT storage (Figure 5E). In contrast, purine and pyrimidine such as uracil, allantoic acid, hypoxanthine nucleoside, guanine nucleoside, and purine nucleoside decreased significantly during RT storage.

3.6 Differentially expressed genes related to the primary metabolism, shikimic acid, and ethylene biosynthesis

As sucrose, shikimic acid, and ethylene accumulated significantly during storage, the expression levels of genes related to these metabolism pathways were further investigated (Figure 6). The expression levels of genes related to sucrose biosynthesis, including sucrose-phosphate synthase (*SPS*), sucrose synthase (*SuSy*), and sucrose-phosphatase (*SPP*) were significantly upregulated under RT storage. Three glucose and xylose isomerized-related genes, namely, uridine diphosphate-glucose 4-epimerase (*UGE*), glucose-6-phosphate 1-epimerase (*GPE*), and

10.3389/fpls.2023.1142840

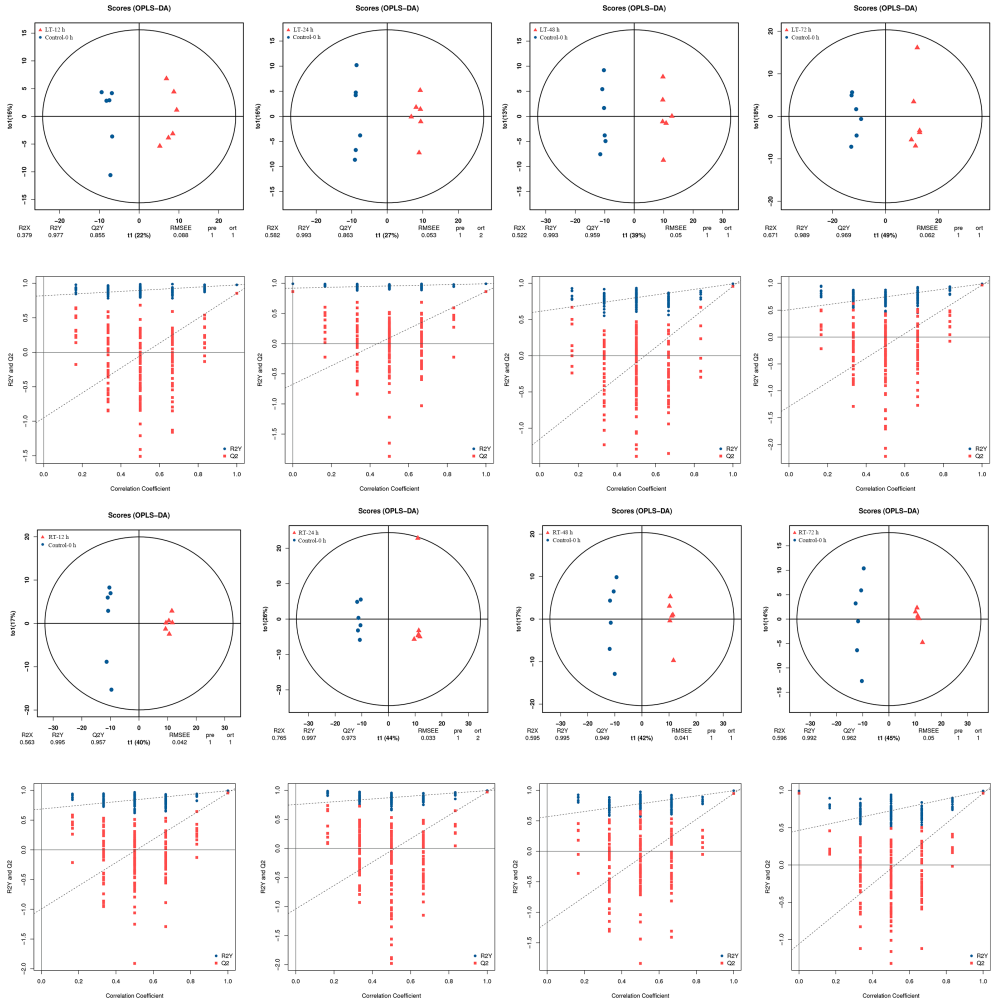


FIGURE 3
Metabolomic analysis of postharvest toon buds using GC-TOF-MS, including the OPLS-DA plots of RT, LT-stored samples, and control (top figure). Performance of the permutation tests validated from the OPLS-DA model (bottom figure).

xylose isomerase (*XI*), play critical roles in monosaccharide configuration transformation during storage. Compared with the control, *UGE*, *GPE*, and *XI* were upregulated significantly, particularly after 24 h of storage. A group of genes related to gluconeogenesis and the TCA cycle was also analyzed. Compared with the control and LT storage, the expression levels of phosphoenolpyruvate carboxykinase (*PCK*), phosphoenolpyruvate carboxylase (*PEPC*), citrate synthase (*CS*), isocitrate dehydrogenase (*ICD*), isocitrate lyase (*ICL*), and *MDH* were obviously upregulated during RT storage. In particular, their relative expression levels reached the maximum after 24–48 h during RT storage. This finding indicated that gluconeogenesis was strengthened and the balance of the TCA cycle was disturbed within 24 or 48 h of RT storage. Oppositely, LT effectively suppressed gluconeogenesis and maintained the relative balance of the TCA cycle.

In addition to those genes related to sugar metabolism, two key enzyme genes involved in amino acid metabolism, including glutamate synthase (*GS*) and arginine decarboxylase (*ADC*), were observed. Similar to *ICL*, the *GS* expression level of RT storage was significantly higher than that of LT storage. The *GS* expression displayed an increasing trend with prolonged RT storage time. Unlike *GS* expression, *ADC* expression reached the maximum under 12 h of RT storage. Then, the *ADC* expression began to decrease gradually. Its expression level during RT storage was lower than that during LT storage. The expression level of galactinol synthase (*GoLS*) decreased gradually with the extended RT storage time. This finding was consistent with the relative content decrease of galactinol. Moreover, 4-hydroxyphenylpyruvate dioxygenase (*HPPD*), an essential enzyme gene involved in aromatic compound biosynthesis, significantly increased during RT storage. However, LT storage effectively inhibited *HDDP* expression. Thus, LT storage contributed to the reduction of the secondary acids containing aromatic metabolites and the preservation of postharvest toon bud quality.

As mentioned above, the ethylene release rate rapidly increased during RT storage. The expression patterns of the three genes involved in the ethylene signaling transduction pathway were further investigated. Two ethylene biosynthetic-genes, namely, 1-aminocyclopropane-1-carboxylate synthase (*ACS*) and 1-aminocyclopropane-1-carboxylate oxidase (*ACO*), and an

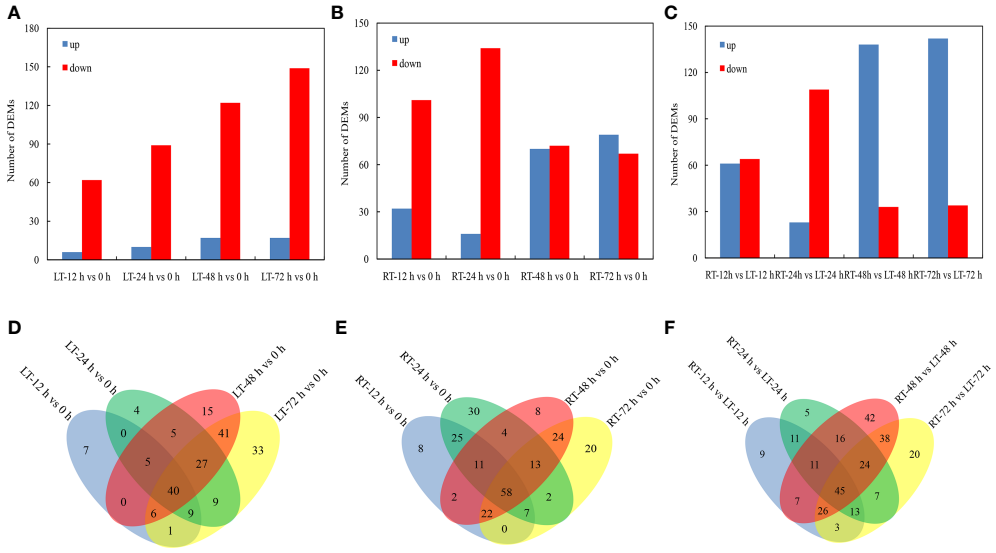


FIGURE 4
Distribution of DEMs at different storage times after harvest. Identification and number statistics of DEMs based on VIP > 1 and $P < 0.05$ **(A–C)**. According to Venn diagram analysis, the DEMs overlapped at different times under RT and LT storages after harvest **(D–F)**.

ethylene-responsive transcription factor (*ERF*) were obviously upregulated during storage. Moreover, their expression levels at RT storage were much higher than those at LT storage. The results indicated that ethylene-triggered events in toon bud may happen after postharvest storage, leading to toon bud senescence and quality deterioration.

4 Discussion

4.1 Changes in the appearance and physiological characteristics of toon bud during storage at RT and LT

Toon buds are widely favored as delicious and nutritious vegetables by consumers. However, postharvest toon buds are susceptible to decay because of their active respiration and vigorous metabolism, resulting in the loss of nutritional value following quality deterioration during RT storage (Zhao et al., 2018; Jiang et al., 2019; Lin S. H. et al., 2019). LT storage can effectively inhibit the respiration of harvested toon buds, maintain their quality, and prolong their postharvest shelf life. Our results further verified that LT could reduce the respiration rate of toon buds and strongly repress ethylene release after harvest.

4.2 Changes in the sugar metabolism of toon buds during storage at RT and LT

The sugars with high sweetness, such as sucrose and fructose, accumulated rapidly with the prolonged storage time, particularly at RT storage. In comparison, the sugars with low sweetness related to stress showed a downward trend. This finding implied that the rot of toon buds was aggravated and the protective effect of osmolytes was weakened. Sugars are the main respiratory substrates for energy supply and osmolytes (Guo et al., 2019; Uarrota et al., 2019). Sucrose and fructose accumulated rapidly with the extended storage time. The possible reason was that the hydrolysis of starch provided glucose for sucrose synthesis. Alternatively, glucose 1-phosphate or glucose 6-phosphate flowed through the gluconeogenesis pathway to promote sucrose synthesis further (Farcuh et al., 2018; Luo et al., 2021). Figure 6 shows that the expression levels of the key enzyme genes of sucrose synthesis increased rapidly during RT storage. However, their expression levels at LT storage were significantly lower than those at RT storage. The increase level of sucrose was quite limited under LT storage conditions. Many studies have shown that ethylene-dependent sucrose accumulation promoted postharvest ripening and senescence of fruits and vegetables (Foukaraki et al., 2015; Xu et al., 2016; Farcuh et al., 2018). Ethylene release rate increased during postharvest ripening of apples, resulting in sucrose and fructose accumulations (Sun et al., 2021). Exogenous ethylene application stimulated the expression of *SuSy* in postharvest blueberries (Wang et al., 2020). Our results were consistent with the above reports.

The intermediates' accumulation in the glycolysis indicated that the metabolic pathway gradually weakened (Luo et al., 2021). The gene encoding glucose-6-phosphate 1-dehydrogenase (*GPD*) is the key enzyme of PPP metabolism (Perotti et al., 2015). During storage for 24–48 h, the *GPD* expression level at RT was significantly higher than that at LT. The results suggested that the PPP metabolism was enhanced in stored toon buds. The transcript level of the *XI* gene involved in the PPP metabolism also displayed an increasing trend at RT. Its expression was extremely low under LT storage, reflecting the strengthening of PPP metabolism under RT storage.

Previous studies have shown that the respiratory metabolism of good fruit and vegetable harvested was mainly based on EMP and

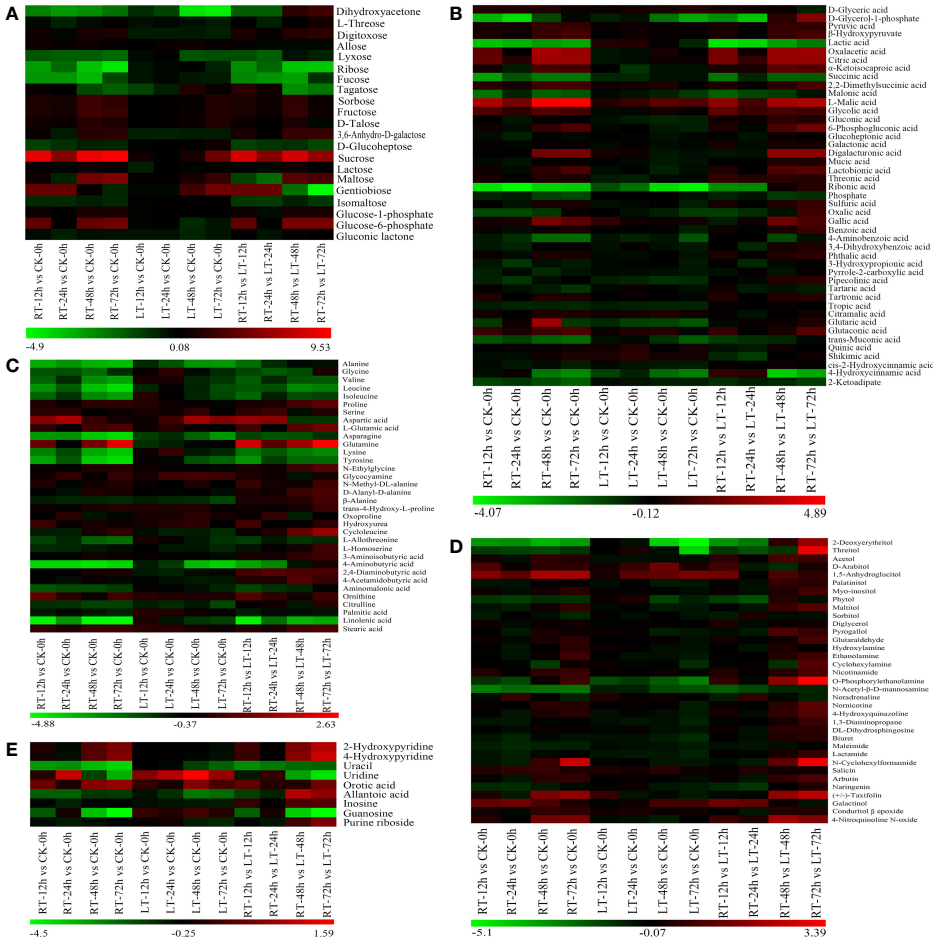


FIGURE 5
Hierarchical cluster analysis of changes in various carbohydrates **(A)**, organic acids **(B)**, amino acids **(C)**, alcohol, aldehyde, amine, and their derivatives **(D)**, and pyridine, pyrimidine and purine **(E)** in toon buds during storage. Red indicates high abundance. The metabolites with low relative abundance are shown in green. (The color key scale is at the bottom of the heatmap).

TCA cycle pathways, and PPP was used as the auxiliary respiratory metabolism (Guo et al., 2019; Lin L. J. et al., 2019; Guo et al., 2022). However, PPP became the active respiratory metabolism for postharvest deteriorated fruit and vegetable tissues. In our results, the toon buds at the early stage of postharvest storage (0–24 h) were robust and healthy. Moreover, the respiratory pathway was mainly through the EMP-TCA cycle to provide sufficient energy for the life activities of the tissues. After 24 h of storage, the browning and quality deterioration of the harvested toon buds began, and the respiratory metabolism considerably depended on PPP.

4.3 Changes in amino acid and organic acid metabolisms of toon buds during storage at RT and LT

In addition, various essential amino acids decomposed severely with increasing storage time, resulting in a decrease in their contents. However, the levels of amino acids maintained at LT storage were higher than those at RT. For example, alanine and aspartic acid were transformed into pyruvate and oxaloacetic acid, respectively. Other amino acids, such as valine, leucine, isoleucine, and lysine, were transformed into α-ketoisocaproic acid through the combined transamination (Bekele et al., 2015; Ren et al., 2020). The *GS* and *ADC* in toon buds at RT storage exhibited significantly higher expressions than those at LT storage. This scenario further aggravated the accumulation of organic acids, such as oxaloacetic acid, citric acid, α-ketoisocaproic acid, and malic acid in the TCA cycle. However, this metabolic flow slowed down under the LT storage of toon buds. Many studies have shown that during postharvest storage, the upregulated expression of various enzyme genes in the TCA cycle accelerated respiratory consumption and reduced the contents of various organic acids (Lin et al., 2015; Liu et al., 2016; Yao et al., 2018). Our results did not exactly accord with previous reports. The possible reason was the upregulated expression of *PEPC* and *PCK* genes in the gluconeogenesis pathway, which converted pyruvate into oxaloacetic acid and reduced it to form malic acid (Perotti et al., 2015; Han et al., 2018).

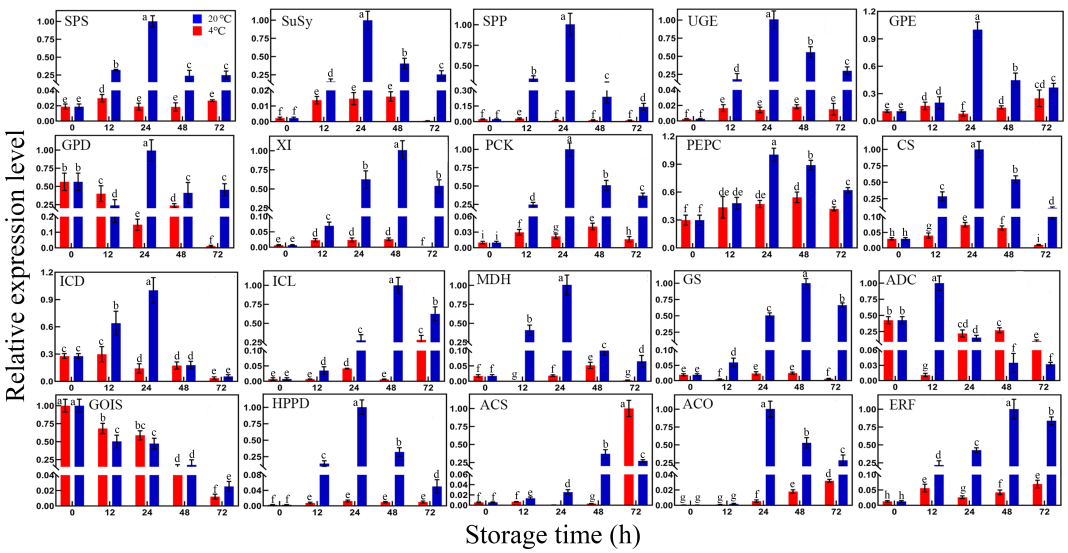


FIGURE 6
Differential expression of genes related to sucrose biosynthesis, sugar, organic acid, amino acid metabolism, ethylene biosynthesis, and signal transduction pathway in toon buds between RT and LT storages. SPS, sucrose-phosphate synthase; SuSy, sucrose synthase; SPP, sucrose-phosphatase; UGE, uridine diphosphate-glucose 4-epimerase; GPE, glucose-6-phosphate 1-epimerase; GPD, glucose-6-phosphate 1-dehydrogenase; XI, xylose isomerase; PCK, phosphoenolpyruvate carboxykinase; PEPC, phosphoenolpyruvate carboxylase; CS, citrate synthase; ICD, isocitrate dehydrogenase; ICL, isocitrate lyase; MDH, malate dehydrogenase; GS, glutamate synthase; ADC, arginine decarboxylase; GoLS, galactinol synthase; HPPD, 4-hydroxyphenylpyruvate dioxygenase; ACS, 1-aminocyclopropane-1-carboxylate synthase; ACO, 1-aminocyclopropane-1-carboxylate oxidase; ERF, ethylene-responsive transcription factor. The values are the means ± SEM from biological replicates. The bars with different lowercase letters are significantly different at $P < 0.05$.

The expression level of *ICL*, the key gene in the glyoxylic acid cycle in the bypass of the TCA cycle pathway, was also highly expressed under RT storage compared with LT storage. This finding indicated that the glyoxylic acid cycle involved in fatty acid catabolism was also strengthened. Intermediate accumulation in the EMP–TCA pathway strengthened gluconeogenesis, transforming non-sugar substances into sugars such as fructose and sucrose (Bekele et al., 2015; Liu et al., 2016).

4.4 Accumulation of aldehydes, acids, amines, and other compounds in toon buds during storage at RT and LT

The superfluous intermediates derived from sugars and amino acids, such as various organic acids, were metabolized into various toxic compounds, such as aldehydes, acids, and amines, through side reaction. This phenomenon further accelerated the quality deterioration of postharvest toon buds. However, this process slows down significantly under LT storage. Moreover, the relative levels of aldehydes, acids, amines, and their derivatives were significantly lower than those under RT storage. Phosphoenolpyruvate, the intermediate of gluconeogenesis pathway, combined with erythrose 4-phosphate, the intermediate of PPP pathway, to form other phenolic acids in the shikimic acid pathway (Becerra-Moreno et al., 2015). The upregulated expression of *HPPD* indicated that the transformation from primary metabolism to secondary metabolism was accelerated in toon buds. This result was consistent with our previous report (Zhao et al., 2021).

On the basis of the above results, we preliminarily described the changes in the molecular metabolic pathways of postharvest toon buds. As demonstrated in Figure 7, the sucrose metabolism, hexose isomerization, gluconeogenesis pathway, and PPP metabolic pathway that centered on glucose 1-phosphate or glucose 6-phosphate were strengthened. In comparison, the glycolysis pathway and the TCA cycle gradually weakened with the prolonged storage time. The increased catabolism of fatty acids and amino acids and the active side reactions, such as transamination, deamination, dehydrogenation, and oxidation, led to the accumulation of ketones, acids, aldehydes, and amines. The metabolic transformation of aromatic amino acids and the accelerated catabolism of pyrimidine or purine nucleotides led to toxic compound accumulation and cell energy consumption. All these circumstances led to the loss of nutrition and quality deterioration of toon buds after harvest.

5 Conclusions

In summary, the molecular mechanism of quality deterioration in toon buds was explored by comparatively analyzing the metabolic differences in toon buds between RT and LT storages. The results showed that compared with RT storage, LT storage could effectively inhibit the respiration rate

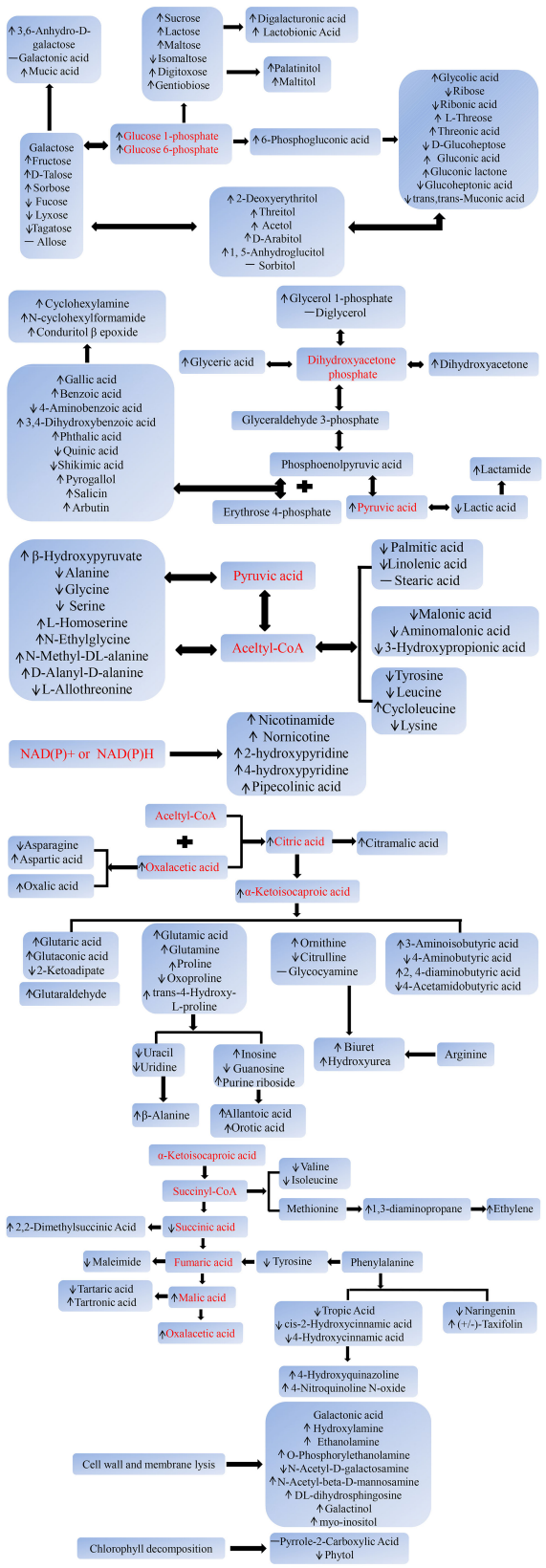


FIGURE 7
A summary of the significantly changed metabolic pathways in RT and LT storages. Narrow: upregulated or downregulated; dash: unchanged.

and the ethylene release rate of toon buds. Moreover, the catabolism of various essential amino acids and fatty acids maintained a comparatively stable level. The expression levels of the key enzyme genes associated with sugar metabolism in LT-stored toon buds were significantly lower than those in RT-stored toon buds, resulting in the limited accumulation of sucrose and various organic acids. However, the primary metabolism displayed the opposite trend in the RT-stored toon buds. Compared with LT storage, RT storage promoted the metabolism of branched-chain synthesis pathways, such as shikimic acid, resulting in the accumulation of various aromatic compounds. In addition, various toxic compounds formed by oxidation, decarboxylation, and transamination rapidly increased under RT storage, eventually leading to quality deterioration and nutrition loss of postharvest toon buds.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

HZ conceived and designed this experiment. HZ drafted the manuscript. CS and QH collected samples of toon sprouts. MF and XL extracted and assayed flavonoid components and volatile terpenoid compounds. JW carried out. qRT-PCR experiments and analyzed the data. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the Natural Science Key Foundations of the Anhui Bureau of Education, the Institution of Higher Education Outstanding Top Talent Cultivation funding project of the Anhui Bureau of Education, and the Science and Technology Special Project of Municipal – University Cooperation (Fuyang Normal University-Fuyang City) (Nos. KJ2021A0683, KJ2020A0545, gxgwfx2020049, and SXHZ202107), and the Innovation Program for College Students (No. S202110371080).

Acknowledgments

The authors would like to thank Dr. Li Wang from the School of Tea and Food Science & Technology, Anhui Agricultural University, for critically reading the manuscript.

Conflict of interest

The 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1142840/full#supplementary-material

References

Becerra-Moreno, A., Redondo-Gil, M., Benavides, J., Nair, V., Cisneros-Zevallos, L., and Jacobo-Velázquez, D. A. (2015). Combined effect of water loss and wounding stress on gene activation of metabolic pathways associated with phenolic biosynthesis in carrot. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00837

Bekele, E. A., Beshir, W. F., Hertog, M., Nicolai, B. M., and Geeraerd, A. H. (2015). Metabolic profiling reveals ethylene mediated metabolic changes and a coordinated adaptive mechanism of 'Jonagold' apple to low oxygen stress. *Physiol. Plant* 155, 232–247. doi: 10.1111/ppl.12351

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). 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. doi: 10.1038/nprot.2011.335

Farcuh, M., Rivero, R. M., Sadka, A., and Blumwald, E. (2018). Ethylene regulation of sugar metabolism in climacteric and non-climacteric plums. *Postharv. Biol. Tech.* 139, 20–30. doi: 10.1016/j.postharvbio.2018.01.012

Foukaraki, S. G., Cools, K., Chope, G. A., and Terry, L. A. (2015). Impact of ethylene and 1-MCP on sprouting and sugar accumulation in stored potatoes. *Postharvest Biol. Technol.* 114, 95–103. doi: 10.1016/j.postharvbio.2015.11.013

Goulas, V., Minas, I. S., Kourdoulas, P. M., Lazaridou, A., Molassiotis, A. N., Gerothanassis, I. P., et al. (2015). ^{1}H NMR metabolic fingerprinting to probe temporal postharvest changes on qualitative attributes and phytochemical profile of sweet cherry fruit. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00959

Guo, X. M., Luo, T., Han, D. M., and Wu, Z. X. (2019). Analysis of metabolomics associated with quality differences between room–temperature– and low–temperature–stored litchi pulps. *Food Sci. Nutr.* 7, 3560–3569. doi: 10.1002/fsn3.1208

Guo, X. M., Luo, T., Han, D. M., Zhu, D. F., Li, Z. Y., Wu, Z. Y., et al. (2022). Multi-omics analysis revealed room temperature storage affected the quality of litchi by altering carbohydrate metabolism. *Sci. Hortic.* 293, 110663. doi: 10.1016/j.scienta.2021.110663

Han, S. K., Nan, Y. Y., Qu, W., He, Y. H., Ban, Q. Y., Lv, Y. R., et al. (2018). Exogenous gamma-aminobutyric acid treatment that contributes to regulation of malate metabolism and ethylene synthesis in apple fruit during storage. *J. Agr. Food Chem.* 66, 13473–13482. doi: 10.1021/acs.jafc.8b04674

Hu, X., Liu, X. L., He, M. Y., Zhao, W., Zhao, Z. Y., Yang, J. X., et al. (2019). Postharvest physiological and biochemical changes and preservation techniques of *Toona sinensis*. *Food Ferment. Ind* 45, 286–291. doi: 10.13995/j.cnki.11-1802/ts.019326

Jiang, X. X., Zhang, B. B., Lei, M. H., Zhang, J. J., and Zhang, J. F. (2019). Analysis of nutrient composition and antioxidant characteristics in the tender shoots of Chinese toon picked under different conditions. *LWT-Food Sci. Technol.* 109, 137–144. doi: 10.1016/j.lwt.2019.03.055

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., et al. (2009). FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* 81, 10038–10048. doi: 10.1021/ac9019522

Li, H. X., Xu, J. W., Cai, M. L., Zeng, H. L., and Wang, X. K. (2015). Improved method for measuring respiratory rate of plant seeds using small-skep-method. *J. Biol.* 32, 100–106. doi: 10.3969/j.issn.2095–1736.2015.01.100

Lin, S. H., Chen, C. K., Luo, H. X., Xu, W. T., Zhang, H. J., Tian, J. J., et al. (2019). The combined effect of ozone treatment and polyethylene packaging on postharvest quality and biodiversity of *Toona sinensis* (A.Juss.) M.Roem. *Postharvest Biol. Technol.* 154, 1–10. doi: 10.1016/j.postharvbio.2019.04.010

Lin, L. J., Lin, Y. X., Lin, H. T., Lin, M. S., Ritenour, M. A., Chen, Y. H., et al. (2019). Comparison between 'Fuyan' and 'Dongbi' longans in aril breakdown and respiration metabolism. *Postharvest Biol. Technol.* 153, 176–182. doi: 10.1016/j.postharvbio.2019.04.008

Lin, Q., Wang, C. Y., Dong, W. C., Jiang, Q., Wang, D. L., Li, S. J., et al. (2015). Transcriptome and metabolome analyses of sugar and organic acid metabolism in ponkan (*Citrus reticulata*) fruit during fruit maturation. *Gene* 554, 64–74. doi: 10.1016/j.gene.2014.10.025

Liu, R. L., Wang, Y. Y., Qin, G. Z., and Tian, S. P. (2016). Molecular basis of 1-methylcyclopropene regulating organic acid metabolism in apple fruit during storage. *Postharvest Biol. Technol.* 117, 57–63. doi: 10.1016/j.postharvbio.2016.02.001

Luo, T., Shuai, L., Lai, T. T., Liao, L. Y., Li, J., Duan, Z. H., et al. (2021). Up-regulated glycolysis, TCA, fermentation and energy metabolism promoted the sugar receding in 'Shixia' longan (*Dimocarpus longan* lour.) pulp. *Sci. Hortic.* 281, 109998. doi: 10.1016/j.scienta.2021.109998

Osorio, S., Carneiro, R. T., Lytovchenko, A., Mcquinn, R., Sørensen, I., Vallarino, J. G., et al. (2019). Genetic and metabolic effects of ripening mutations and vine detachment on tomato fruit quality. *Plant Biotechnol. J.* 18, 106–118. doi: 10.1111/pbi.13176

Perotti, V. E., Moreno, A. S., Trípodi, K., Del Vecchio, H. A., Meier, G., Bello, F., et al. (2015). Biochemical characterization of the flavedo of heat-treated Valencia orange during postharvest cold storage. *Postharvest Biol. Technol.* 99, 80–87. doi: 10.1016/j.postharvbio.2014.08.007

Pott, D. M., Vallarino, J. G., and Osorio, S. (2020). Metabolite changes during postharvest storage: effects on fruit quality traits. *Metabolites* 10, 187. doi: 10.3390/metabo10050187

Ren, L., Zhang, T. T., Wu, H. X., Ge, Y. X., Zhao, X. H., Shen, X. D., et al. (2020). Exploring the metabolic changes in sweet potato during postharvest storage using a widely targeted metabolomics approach. *J. Food Process. Preserv.* 45. doi: 10.1111/jfpp.15118

Su, S., Wang, L. J., Ni, J. W., Geng, Y. H., and Xu, X. Q. (2020). Diversity of red, green and black cultivars of Chinese toon [*Toona sinensis* (A. juss.) roem]: anthocyanins, favonols and antioxidant activity. *J. Food Meas. Charact.* 14, 3206–3215. doi: 10.1007/s11694-020-00560-8

Sun, Y. J., Shi, Z. D., Jiang, Y. P., Zhang, X. H., Li, X. A., and Li, F. J. (2021). Effects of preharvest regulation of ethylene on carbohydrate metabolism of apple (*Malus domestica* borkh cv. starkrimson) fruit at harvest and during storage. *Sci. Hortic.* 276, 109748. doi: 10.1016/j.scienta.2020.109748

Tang, N., Deng, W., Hu, N., Chen, N., and Li, Z. G. (2016). Metabolite and transcriptomic analysis reveals metabolic and regulatory features associated with Powell orange pulp deterioration during room temperature and cold storage. *Postharvest Biol. Technol.* 112, 75–86. doi: 10.1016/j.postharvbio.2015.10.008

Uarrota, G. V., Fuentealba, C., Hernández, I., Defilippi-Bruzzone, B., Meneses, C., Campos-Vargas, R., et al. (2019). Integration of proteomics and metabolomics data of early and middle season hass avocados under heat treatment. *Food Chem.* 289, 512–521. doi: 10.1016/j.foodchem.2019.03.090

Wang, L. Q., Lin, S. H., Chen, C. K., Zhang, H. J., Luo, H. X., and Xu, W. T. (2019). Effects of three different preservation methods on storage quality of Toona sinensis. *Food Res. Dev. Food Res.* 40, 150–155.

Wang, S. Y., Zhou, Q., Zhou, X., Zhang, F., and Ji, S. J. (2020). Ethylene plays an important role in the softening and sucrose metabolism of blueberries postharvest. *Food Chem.* 310, 125965–125973. doi: 10.1016/j.foodchem.2019.125965

Xu, F., Wang, H. F., Tang, Y. C., Dong, S. Q., Qiao, X., Chen, X. H., et al. (2016). Effect of 1-methylcyclopropene on senescence and sugar metabolism in harvested broccoli florets. *Postharvest Biol. Technol.* 116, 45–49. doi: 10.1016/j.postharvbio.2016.01.004

Yang, Y., Wang, J., Xing, Z. E., Dai, Y. Q., and Chen, M. (2011). Identification of phenolics in Chinese toon and analysis of their content changes during storage. *Food Chem.* 128, 831–838. doi: 10.1016/j.foodchem.2011.03.071

Yang, H., Zhao, S. H., Shi, G. Y., Zhang, L., Wang, X. M., and Wang, Z. G. (2019). Effect of near freezing-point storage on quality and key flavor substances of *Toona sinensis* bud. *Storage Process* 19, 46–52. doi: 10.3969/j.issn.1009-6221.2019.05.008

Yao, S. X., Cao, S. X., Xie, J., Deng, L. L., and Zeng, K. F. (2018). Alteration of sugar and organic acid metabolism in postharvest granulation of Ponkan fruit revealed by transcriptome profiling. *Postharv. Biol. Tech.* 139, 2–11. doi: 10.1016/j.postharvbio.2018.01.003

Zhang, J., Wang, X., Yu, O., Tang, J., Gu, X., Wan, X., et al. (2011). Metabolic profiling of strawberry (*Fragaria* × *ananassa* duch.) during fruit development and maturation. *J. Exp. Bot.* 62, 1103–1118. doi: 10.1093/jxb/erq343

Zhang, X. M., Zhao, F. C., Li, H. L., and Liu, Y. Y. (2009). Effects of Allium macrostemon bunge extracts on the preservation of Toona sinensis. chin. *Agric. Sci. Bull.* 25, 55–58.

Zhao, H., Feng, S. S., Zhou, W., and Kai, G. Y. (2019). Transcriptomic analysis of postharvest toon buds and key enzymes involved in terpenoid biosynthesis during cold storage. *Sci. Hortic.* 257, 108747. doi: 10.1016/j.scienta.2019.108747

Zhao, H., Lv, W. J., Fan, Y. L., and Li, H. Q. (2018). Gibberellic acid enhances postharvest toon sprout tolerance to chilling stress by increasing the antioxidant capacity during the short-term cold storage. *Sci. Hortic.* 237, 184–191. doi: 10.1016/j.scienta.2018.04.018

Zhao, H., Ren, L. P., Fan, X. Y., Tang, K. J., and Li, B. (2017). Identification of putative flavonoid-biosynthetic genes through transcriptome analysis of taihe *Toona sinensis* bud. *Acta Physiol. Plant* 39, 122. doi: 10.1007/s11738-017-2422-9

Zhao, H., Shi, X. P., Shen, C., Chen, C. F., Liu, J. Y., and Qu, C. Q. (2021). High-throughput sequencing analysis reveals effects of short-term low temperature storage on miRNA-mediated flavonoid accumulation in postharvest toon buds. *Plant Gene* 26, 100291. doi: 10.1016/j.plgene.2021.100291

Zhu, Y. Q., and Gao, J. (2017). Effect of package film on the quality of postharvest chinese toon tender shoots storage. *J. Food Qual.* 2017, 5605202. doi: 10.1155/2017/5605202

10.3389/fpls.2023.1142840

Appendix A. Supplementary data

Glossary

ACO	1-aminocyclopropane-1-carboxylate oxidase
ACS	1-aminocyclopropane-1-carboxylate synthase
ADC	arginine decarboxylase
CS	citrate synthase
EMP	Embden–Meyerhof pathway
ERF	ethylene-responsive transcription factor
GC	gas chromatograph
GC-TOF-MS	gas chromatograph coupled with a time-of-flight mass spectrometer
GoLS	galactinol synthase
GPD	glucose-6-phosphate 1-dehydrogenase
GPE	glucose-6-phosphate 1-epimerase
GS	glutamate synthase
DEMs	differentially expressed metabolites
DI	decay index
HCA	hierarchical clustering analysis
HPPD	4-hydroxyphenylpyruvate dioxygenase
ICD	isocitrate dehydrogenase
ICL	isocitrate lyase
IS	internal standard
LT	low temperature
MDH	malate dehydrogenase
MSSM	modified small-skep-method
OPLS–DA	orthogonal partial least squares discriminant analysis
PCA	principal component analysis
PCK	phosphoenolpyruvate carboxykinase
PEPC	phosphoenolpyruvate carboxylase
PPP	pentose phosphate pathway
qRT-PCR	quantitative real-time PCR
RSD	relative standard deviation
RT	room temperature
SPP	sucrose-phosphatase
SPS	sucrose-phosphate synthase
SuSy	sucrose synthase
UGE	uridine diphosphate-glucose 4-epimerase
VIP	variable importance in the projection
XI	xylose isomerase

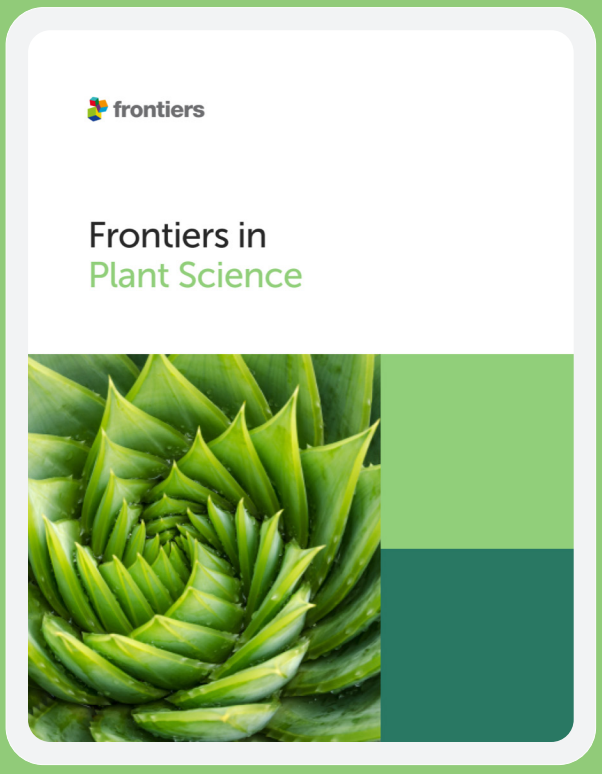
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact