DIVERSE EXPRESSION OF ISOFLAVONOID-RELATED GENES BASED ON TRANSCRIPTOMIC DATASETS OF *Pueraria mirifica* CULTIVARS

H. T. T. Huynh[1, 2 *], N. T. B. Nguyen[1], H. H. Ha[1], O. T. K. Pham[3], C. X. Nguyen[4] and H. H. Nguyen[1,2]

[1]Institute of Biology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[2]Department of Biotechnology, Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[3]Pollution Control Department, Ministry of Agriculture and Environment, Hanoi, Vietnam
[4]Faculty of Biotechnology, Vietnam National University of Agriculture, Hanoi, Vietnam
*Corresponding Author's Email: huehuynh@ib.ac.vn

ABSTRACT

White Kwao Krua, also known as *Pueraria mirifica*, is a traditional medicinal plant in several Asian countries. This plant has a high content of essential phytoestrogens such as isoflavones and chromenes, particularly miroestrol. However, their biosynthetic pathway remains unclear and is currently under investigation. Three gene families Chalcone isomerases (*CHI*s), Chalcone synthases (*CHSs*), and UDP-dependent glycosyltransferases (*UGTs*) play a key role in the phytoestrogen biosynthesis in *P. mirifica*. They are large gene families with myriad members, involved in several plant functions. In the research, five *P. mirifica* cultivars (TLBYT, TLCNX, TLDB, NA, and SL) were cultivated under the same conditions and then their leaf, stem, and tuber tissues were collected. The transcriptomes of the cultivars were sequenced, assembled, and annotated in the research. By using RNA-seq, the transcriptome assembly yielded over 300,000 unigenes, of which approximately 32,000 were annotated across four major databases: NCBI-Nr (229569 unigenes), SwissProt (158667 unigenes), COG (112089 unigenes), and KEGG (61480 unigenes). Seventeen individuals from these gene families that may catalyze or influence miroestrol and isoflavone biosynthesis were detected. The RT-qPCR analysis revealed tissue-specific gene expression, with several genes showing preferential expression in either leaves or tubers. *CHS11*, *CHS13*, and *UGT74* gene were predominantly expressed in leaves, whereas *CHI4A*, *CHI3A2*, and *CHS14* showed higher expression in tubers, the primary site of phytoestrogen accumulation. These results provide transcriptomic data for different *P. mirifica* varieties and demonstrate tissue-specific expression patterns of key *CHI*, *CHS*, and *UGT* genes involved in the isoflavonoid biosynthetic pathway.

***Keywords*:** Chalcone isomerase, Chalcone synthase, phytoestrogens, *Pueraria mirifica,* UDP-dependent glycosyltransferases.

https://doi.org/10.36899/JAPS.2026.4.0100 Published first online May 21, 2026

INTRODUCTION

White Kwao Krua, *Pueraria candollei* var. *mirifica* (shortened as *Pueraria mirifica*) is a native herb widely distributed in the deciduous forests of Asia. Their organs have been used as traditional herbal medicine for centuries. Intharuksa *et al.* (2020) described the uses of *P. mirifica* as a rejuvenating medicine as well as for alleviating symptoms of estrogen deficiency (e.g., hair loss, wrinkles, and sagging breasts) (Intharuksa *et al.*, 2020). At least 17 phytoestrogen constituents that structurally or functionally mimic mammalian estrogen, 17β-estradiol, were reported in the *P. mirifica* roots (Malaivijitnond, 2012). In particular, miroestrol and its precursor deoxymiroestrol are the two chromenes types that accumulate mainly in *P. mirifica* and have the highest estrogenic capacity among phytoestrogens (Yusakul *et al*., 2011; Suntichaikamolkul *et al*., 2023). There has been an increasing interest in the medicinal properties of *P. mirifica*; however, the chromene synthesis route still is not illustrated in detail. In Vietnam, one research showed the increasing weight of reproductive organs in ovariectomized rats after treatment of *P. mirifica* extract (Dao *et al*., 2018). To date, extracted compounds from tuber of *P. mirifica* are used as important ingredients in dietary supplement products to reduce premenopausal symptoms.

Although several information regarding *P. mirifica* biological traits is currently available, aspects including population structure, nutritional requirements, reproduction, and studies involving NGS datasets are scarce. A few public transcriptomic resources containing raw RNA-seq data without open-access transcriptome assemblies are published for *P. mirifica (*Suntichaikamolkul *et al*., 2019) as well as the closely related *P. montana* species (Wang *et al*., 2015; Haynsen *et al*., 2018; He *et al*., 2019). The majority of the samples were leaf or root tissues from a single specimen. There were only two previous studies speculated on various *P. mirifica* tissues. Publicly available genome sequences and RNA-seq data for *Pueraria* sp. may facilitate a better understanding of *P. mirifica*'s molecular biology.

The concentration of phytoestrogens in the medicinal plants is remarkably influenced by various cultivars. Besides, Chalcone synthases (CHSs), Chalcone isomerases (CHIs), and 5'-diphosphate glycosyltransferases (UGTs), are the pivotal enzymes in phytoestrogens synthesis in medical plants (Austin and Noel, 2003). In legumes, CHS co-acts with other enzymes such as Chalcone reductase (CHR) to produce isoliquiritigenin chalcone, which is the first important process in the biosynthesis of flavonoid and miroestrol. Several researches have been determined on genes involved in phytoestrogen synthesis pathways. In soybean, 14 unique *CHS* genes in the *CHS* family were detected by genome-wide analysis. In terms of *CHIs*, this superfamily in Fabaceae is classified into four subfamilies with different functions based on their protein structure. The first two subfamilies (CHI1, CHI2) have the catalytic activity to convert chalcone to flavanone (Yin *et al*., 2019). In 2005, two other CHI subfamilies were found but their catalytic activity was concealed in *P. mirifica*. Besides, glycosylation is a crucial step in the downstream isoflavone biosynthesis (Szeja *et al*., 2017). The UGTs involves in transferring glycosyl moieties from UDP-sugar donors to a wide range of isoflavonoid acceptors (Egorova and Toukach, 2017; Wang *et al*., 2019). These genes are the key factor direct to isoflavone synthesis instead of miroestrol generation.

The *CHS*, *CHI*, and *UGT* gene families play important roles in the metabolic capacity of *P. mirifica*; however, their classification remains unclear. This study investigated the number, expression levels, and expression differences of these gene families among five *P. mirifica* cultivars from Thailand and Vietnam using comparative transcriptome analysis. The study of phytoestrogen biosynthetic genes in *P. mirifica* provides important insights into secondary metabolite production, thereby improving understanding of the impact of gene expression on secondary metabolism.

MATERIALS AND METHODS

Plant materials: The five *P. mirifica* cultivars were studied including TLBYT, TLCNX, TLDB originated from Thailand, and the NA, SL cultivars originated from Vietnam. The plants were cultivated and maintained for two years at Vinh University's experimental garden (18°39'34.8"N; 105°41'48.5"E). The plants were cultivated under nutrient-rich, well-aerated soil conditions with a pH range of 5.5–6.5. Irrigation was applied daily, and shading nets were used to protect the plants from excessive direct sunlight. The cultivars are vein plants with deltoid leaves and trifoliate leaf structures. Trichomes are distributed in both stem and leaves. The cultivars differ slightly in tuber shape, number of tubers per plant, and stem color ranges from gray to light purple. Samples were collected from two-year-old plants. Specimens of tuber, stem, and whole mature leaf tissues of each cultivar were harvested and preserved in the RNA-later Stabilization Reagent (Qiagen GmbH, Hilden, Germany) for 24 hours at 4°C, then stored at -80°C for total RNA extraction.

RNA extraction and RNA sequencing: For RNA extraction, 100 mg of plant tissues (leaves, stem) were homogenized using TRIzol buffer for each. Following the manufacturer's instructions, total RNA was extracted from leaf and stem samples using TRIzol (Invitrogen, ThermoFisher Scientific, Waltham, MA, USA). For tuberous roots, the RNA isolation procedure with an SDS-based extraction buffer was applied.

For each of five cultivars, one to three tissue types were collected. For each tissue type, three biological samples were collected and RNA was extracted independently, then the resulting total RNAs were pooled prior to the RNA-seq procedure. Details of sample collected are presented in Supplement Table 2. The Agilent TapeStation was used to evaluate the quality and integrity of the RNA (High sensitivity RNA ScreenTape, Agilent Technologies, Santa Clara, CA, USA). Extracted RNA quantities were measured using the Nanodrop 2000 (ThermoFisher Scientific, Waltham, MA, USA). Only samples with total RNA concentrations exceeding 50 ng μL^{-1} were considered acceptable for subsequent analyses. RNA integrity levels for each sample chosen for sequencing ranged from 5.6 to 7.1. PolyA-enrichment of total RNA was performed for cDNA library preparation. In total, nine libraries were eligible for sequencing on the Illumina NextSeq500 platform, generating 150 bp pair-end sequencing reads.

Raw sequence processing and Transcriptome assembly: The original image data was analyzed using Bcl2fastq v2.17.1.14 for base calling and demultiplexing, resulted in 10 paired-end RNA-seq libraries Sequence quality assessment of raw RNA-seq data, removal of Illumina adapter sequences, and quality trimming were performed using Cutadapt v1.9.1

(Martin, 2011). High-quality reads were De Bruijn graph-based de novo assembled using Trinity v2.2.0 (Haas *et al*., 2013). To remove foreign contamination, the assembly contigs were verified via NCBI-VecScreen. The clean assembly contigs were further connected via sequence clustering into long non-redundant unigene sequences (Cd-hit-est v4.8.1 (Li and Godzik, 2006). TransDecoder v5.5.0 (The Broad Institute, Cambridge, MA, USA) was used to detect open reading frames (GitHub).

Gene annotation of transcriptomes: To perform gene annotation for assembled transcriptomes, amino acid sequences of unigenes were blasted against up-to-date databases NCBI non-redundant protein database (NCBI-nr; https://www.ncbi.nlm.nih.gov/refseq/), Swiss-Prot (https://www.ebi.ac.uk/uniport), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.kegg.jp/kegg/) (Kanehisa and Goto, 2000), Clusters of Orthologous Groups (COG) (https://www.ncbi.nlm.nih.gov/COG/) (Adrian *et al*., 2013), and BLAST2GO v2.5.0 for Gene Ontology (GO) (Götz *et al*., 2008).

Differential expression analysis:Differential expression gene (DEG) analysis used the DESeq Bioconductor package, a model based on the negative binomial distribution (Love *et al*., 2014). After being adjusted by Benjamini and Hochberg's (José, 2007) approach for controlling the false discovery rate, the P-value of genes was set to less than 0.05 to detect differentially expressed ones.

***CHIs*, *CHSs*, *UGTs* genes screening and phylogenetic analysis:** From the annotation database, we classified unigenes of *CHIs*, *CHSs*, and *UGTs*, which are the essential gene families in the isoflavonoid biosynthesis pathways including miroestrol and isoflavone. To verify exactly all the members of *P. mirifica CHI*, *CHS*, and *UGT* families, these unigenes were analyzed by the BLAST tool (NCBI) (Altschul *et al*., 1990). All seventeen *CHI*, *CHS*, and *UGT* sequences were submitted to NCBI with AC number from OR588019 to OR588035.

The *CHIs*, *CHSs*, and *UGT*s coding sequences of other Fabaceae members were retrieved from Genbank (Dennis *et al*., 2013). Multiple sequence alignment was performed by MEGAv11 software (Kumar *et al*., 2018), then a maximum likelihood (ML) phylogenetic tree was conducted with the Hasegawa-Kishino-Yano model Hasegawa, and a bootstrap test was set to 1000 replicates.

Quantitative Real-Time PCR Analysis: The RT-qPCR was conducted to screen the expression of *CHSs*, *CHIs*, and *UGTs* in leaves and tubers among cultivars. First-strand cDNA was synthesized from total RNA using the Revert Aid Reversed Transcriptase Kit (ThermoFisher Scientific, Waltham, MA USA) for Real-Time PCR with the reaction set-up according to the product's manual. Ten specific primer pairs were designed to amplify ten genes related to flavonoid and miroestrol synthesis (Table S1).

The RT-qPCR analysis was conducted in the LightCycler® 96 System (Roche, Basel, Switzerland). Each 20 µL reaction mixture contained 10 µL Luna Universal qPCR Master Mix (New EngLand BioLab, Hitchin, Hertfordshire, UK), 0.5 µL of 10 µM of each primer, and 20 ng cDNA. Each reaction consisted of a cycle of 3 min at 95°C, followed by 45 cycles of 20 s at 95°C and 30 s at 60°C. Melting curves were obtained to verify primer efficiency by the default setup of the LightCycler® 96 system (Roche, Basel, Switzerland). The elongation Factor 1 alpha gene was chosen as endogenous control.

Data analysis and Availability: The qPCR assay included three technical and biological replicates for each gene. Statistical analysis was performed using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). The filtered and cleaned original RNA sequencing data were deposited at the NCBI Sequence Read Archive under the SRA project PRJNA666754.

RESULTS

Transcriptome sequencing output and assembly: From five two-year-old *P. mirifica* plants of Vietnamese origin (NA and SL) and Thai origin (TLBYT, TLCNX, and TLDB), nine high-throughput transcriptomic datasets were generated. To identify tissue-specific or tissue-preferential genes, transcriptomes from tubers, stems, and leaves were collected separately. A total of 80 GB was generated using Illumina RNA-seq technology and subsequently cleaned by removing adaptor sequences and low-quality reads. The average Q30 score and GC content were 95.21% and 47.4%, respectively (Table S3). Based on these high-quality data, 309,975 contigs comprising 109.3 Mbp of total transcript sequence were assembled and further filtered to remove contaminated sequences. Finally, 309,453 unigenes were identified from these contigs. Among the *Pueraria* transcriptomes, the non-redundant unigenes had an average length of 573 bp and an N50 value of 864 bp, constituting a highly comprehensive full-length sequence database with 93.4% BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness.

Functional annotation and classification of unigenes: To estimate the proportion of transcripts sharing homology with sequences descending from closely-related organisms, a taxonomic classification of all unigene sequences was performed using BLAST search (v2.2.28+) against the NCBI-Nr database. Among the total of 309453 unigenes had positive blast hits, *Glycine max* had the highest blast hits (63158 unigenes), followed by *G. soja* (36 384 unigenes) and *Mucuna pruriens* (21571 unigenes). These unigenes were annotated by NCBI-Nr (229569 unigenes), SwissProt (158667 unigenes), COG (112089 unigenes), and KEGG (61480 unigenes). Of these, 32023 unigenes were annotated in all four databases (Fig. 1a).

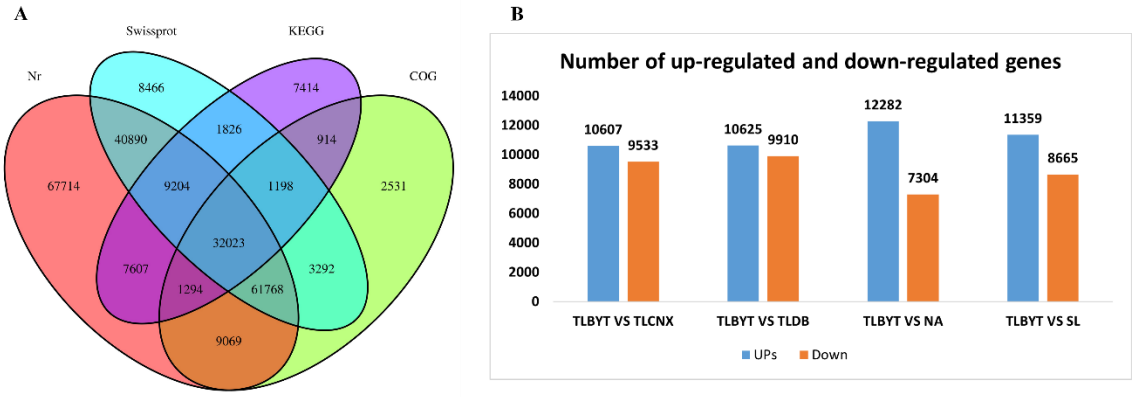

Fig. 1. Identification and functional analysis of unigenes from *Pueraria mirifica* transcriptome. (a) The Venn diagrams showing the number of unigenes annotated by different databases sources, (b) The number of up-regulated genes (blue) and down-regulated genes (red) in each comparison.

Gene Ontology (GO) analysis classified gene products from all *P. mirifica* transcriptomes into three major categories (Fig. S1). The majority of unigenes were assigned to the "cellular component" category (316,002 unigenes), followed by the "biological process" category (310,052 unigenes). Meanwhile, 217,407 unigenes were associated with the "molecular function" category. According to the COG database, the unigenes were classified into different functional groups (Fig S1). The two largest groups were "signal transduction mechanism" unigenes (17836 unigenes) and "General function prediction only" (14212 unigenes). KEGG pathway analysis was performed to identify gene products involved in metabolic pathways and their roles in cellular processes. A total of 61,480 unigenes were annotated across various pathways (Fig. S1). Aside from the "global and overview maps", the three most illustrated pathways were "translation" (15616 unigenes) belonging to the "Genetic information processing" group, while both "carbohydrate metabolism" (14771 unigenes) and "amino acid metabolism" (8672 unigenes) belonging to "Metabolism" group.

Comparison of Differentially Expression Genes (DEGs) among cultivars: Differentially expressed genes (DEGs) were identified among the transcriptome profiles of the cultivars. When the TLBYT dataset was compared with the TLCNX dataset, 20,140 unigenes were differentially expressed, including 10,607 upregulated and 9,533 downregulated genes. In the TLBYT versus TLDB and TLBYT versus SL comparisons, 10,625 genes were upregulated and 9,910 were downregulated, and 11,359 genes were upregulated and 8,665 were downregulated, respectively. The lowest number of DEGs was observed in the TLBYT versus NA comparison, with a total of 19,583 genes (12,282 upregulated and 7,304 downregulated) (Fig. 1b).

DEGs between each comparison were further annotated into KEGG and the top pathways with the most significant enrichment were listed in Fig. 2. Overall, the DEGs were annotated mainly in the "plant hormone signal transduction", "spliceosome", "starch and sucrose metabolism" groups with around 200 genes, and they enriched the biosynthesis and metabolism pathways. Fig. 2a illustrated the comparison between TLBYT and TLCNX cultivars in 21 pathways, DEGs enriched in "Stilbenoid, diarylheptanoid and gingerol biosynthesis", "Isoflavonoid biosynthesis", and "Zeatin biosynthesis". Twenty-one pathways were listed in Fig. 2b compared TLBYT versus TLDB, the most enriched parameters were recorded to "Stilbenoid, diarylheptanoid, and gingerol biosynthesis" and "Glycosylphosphatidylinositol (GPI)-anchor biosynthesis". Similar pathways were found in the two remaining comparisons (TLBYT versus NA (Fig. 2c) and TLBYT versus SL (Fig. 2d)).

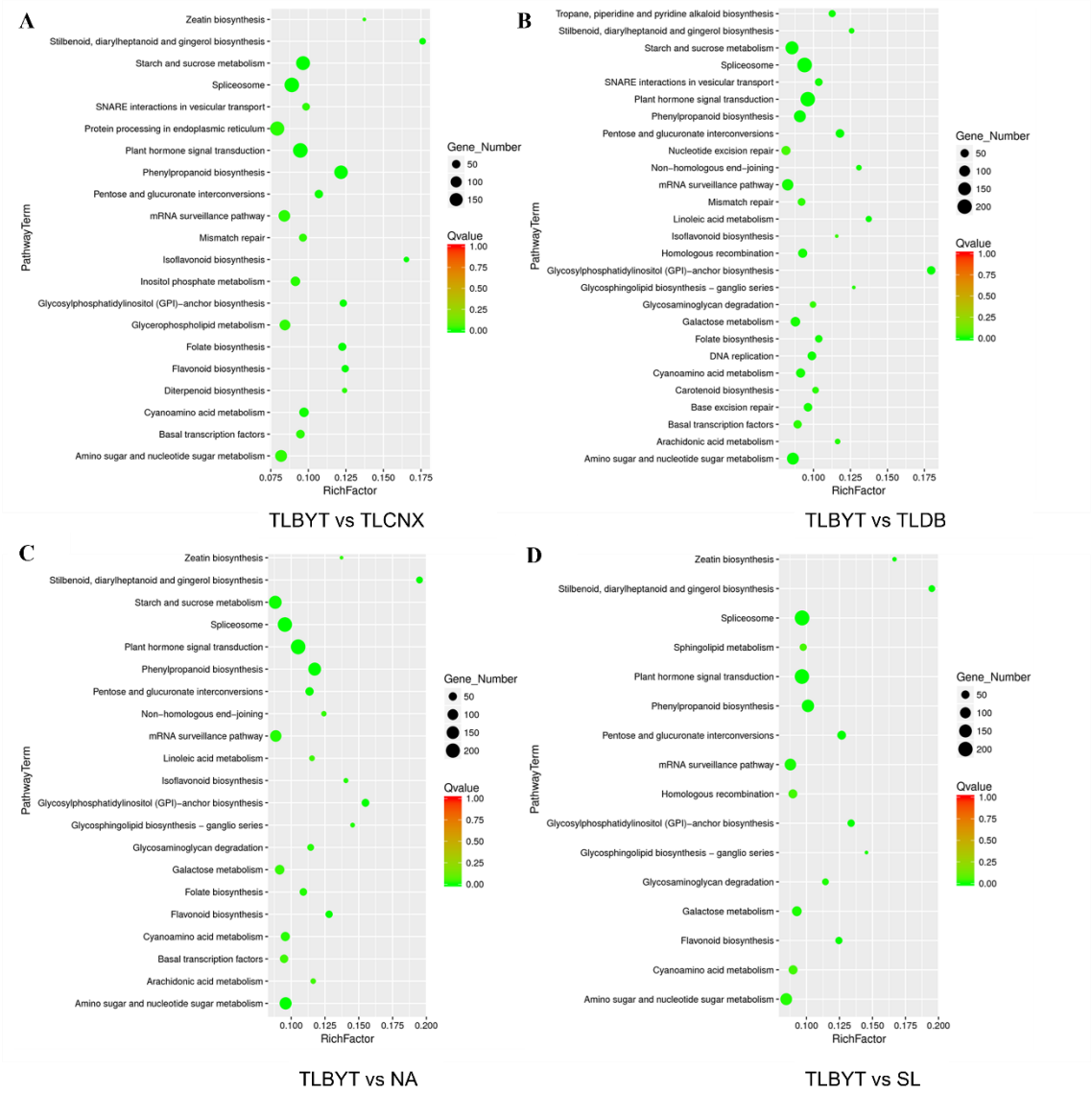

Fig. 2: Scatter plot of differential gene KEGG enrichment analysis between *Pueraria mirifica* cultivars. (a) TLBYT and TLNCX, (b) TLBYT and TLDB, (c) TLBYT and NA, (d) TLBYT and SL. The X-axis indicates the Rich factor. The Y-axis specifies the KEGG pathways. The dot size positively correlates with the number of differentially expressed genes in the pathway. Color coding indicates different Q-value ranges. The Rich factor indicates the ratio of the number of differentially expressed genes in the pathway to the total number of genes in the pathway. A greater Rich factor indicates a greater degree of enrichment. The Q-value is the P-value after multiple hypothesis testing and ranges between 0 and 1, the closer the Q-value is to zero, the more significant the enrichment is.

Phylogenetic analysis of P. *mirifica* CHIs, CHSs, and UGTs genes in the Fabaceae family: Unigenes were identified as *CHI*, *CHS*, and *UGT* genes based on high sequence similarity to GenBank reference sequences. For each gene family, the coding sequences (CDSs), together with CDSs from homologous genes identified in other Fabaceae species, were used to investigate genetic relationships.

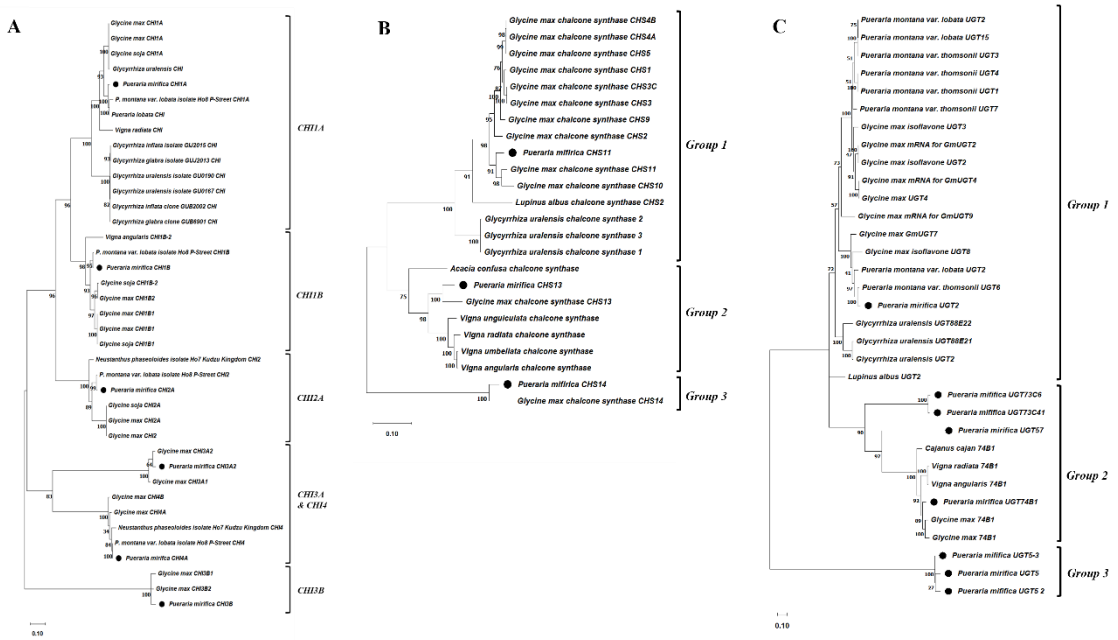

Fig. 3: Maximum likelihood phylogenetic tree based on the coding sequence of genes in each dataset. (a) *CHS* gene tree, (b) *CHI* gene tree, (c) *UGT* gene tree.

A Maximum likelihood phylogenetic tree was generated accordingly for each gene family by multiple sequence alignment (Fig. 3). As a result, six genes in the *CHI* family were detected among 27 unigenes with their full-length sequences: *CHI1A, CHI1B, CHI2A, CHI3A2, CHI3B,* and *CHI4A*. The *CHI* phylogenetic tree revealed five distinct clusters (Fig. 3a), which was similar to the recent study of the *GmCHI*. In more detail, the subfamilies *CHI1* and *CH2* were separated into different branches. *CHI3s* were divided into two clusters: *CHI3A2* and *CHI4A*. Notably, the *P. mirifica CH3A2* and *CHI4* groups (*PmCHI4A* and *PmCHI4B*) were grouped into two subclusters. This finding is congruent with the evolutionary history in which *CHI4* genes are derived from *CHI3* genes (Ngaki et al., 2012). Consistently, our study revealed that in *P. mirifica*, *CHI4* genes originated from *CHI3A*. The maximum-likelihood (ML) phylogenetic tree (Fig. 3a) showed that *PmCHI* genes have a closer genetic relationship with those of *P. lobata* and *Glycine max* than with genes from other Fabaceae members.

Using the same workflow, three full-length *CHS* genes - *CHS11*, *CHS13*, and *CHS14*-were identified from 33 *CHS* unigenes. As shown in Fig. 3b, *CHSs* were clustered into three divergent groups, with *PmCHS*s showing the closest relationships to *GmCHS*s. The largest group was further divided into three subgroups, comprising *GmCHS*s, Glycyrrhiza *CHSs*, and *PmCHS11*, which clustered with *GmCHS11* and *GmCHS10*. In the second group, *PmCHS13* was more closely related to *GmCHS13* than to *CHSs* from the genera *Vigna* and *Acacia confusa.* The third group consisted of *PmCHS14* and *GmCHS14* (Fig. 3b).

Eight full-length sequences were confirmed from 161 *UGT* unigenes, including *UGT2*, *UGT73C6, UGT73C41, UGT57, UGT74B1*, and three *UGT5* members. These genes were classified into three major groups (Fig. 3c). Group 1 comprised *PmUGT2* together with *UGTs* from *P. lobata*, *P. thomsonii*, *G. max*, and *Glycyrrhiza uralensis*. Within this group, *PmUGT2* showed the highest sequence similarity to *PtUGT6*, followed by *PiUGT2* and *GmUGTs*, forming a distinct subgroup. In Group 2, *PmUGT74B1* was more closely related to the two *G. max UGT74B1* genes than to those from Vigna species. Notably, the three *PmUGT5* genes formed a distinct and divergent clade in the *UGT* phylogenetic tree.

***CHIs*, *CHSs*, and *UGTs* gene verification by RT-qPCR**: To determine tissue-specific expression patterns of the gene families investigated *(CHIs, CHSs*, and *UGTs)* when compared with EF1α - an internal control, RT-qPCR was conducted to amplify gene members of these families. Ten specific primer pairs were developed and amplified successfully in both leaves and tuberous tissues of three cultivars (TLBYT, NA, SL) (Fig. 4). In general, members of each gene family expressed differently across leaf-tuber tissues in all three cultivars and some differences in the level of expression between cultivars. In detail, *CHI1A* gene had the lowest expression in all tested samples*, CHS13* gene was highest level expression especially in leaves*.* Among these genes, *CHI3B, CHS11, CHS13*, and *UGT74* were found to be substantially more highly

expressed in the leaves of three *P. mirifica* cultivars than other genes. Likewise, *UGT74* was nearly not expressed in TLBYT, NA, SL tubers, while *CHS11* and *CHI3B and UGT2* expression level in tuber were a quite high. The significant gene expression level in tuber were found in *CHI3A2*, *CHI4A*, *CHS14* genes.

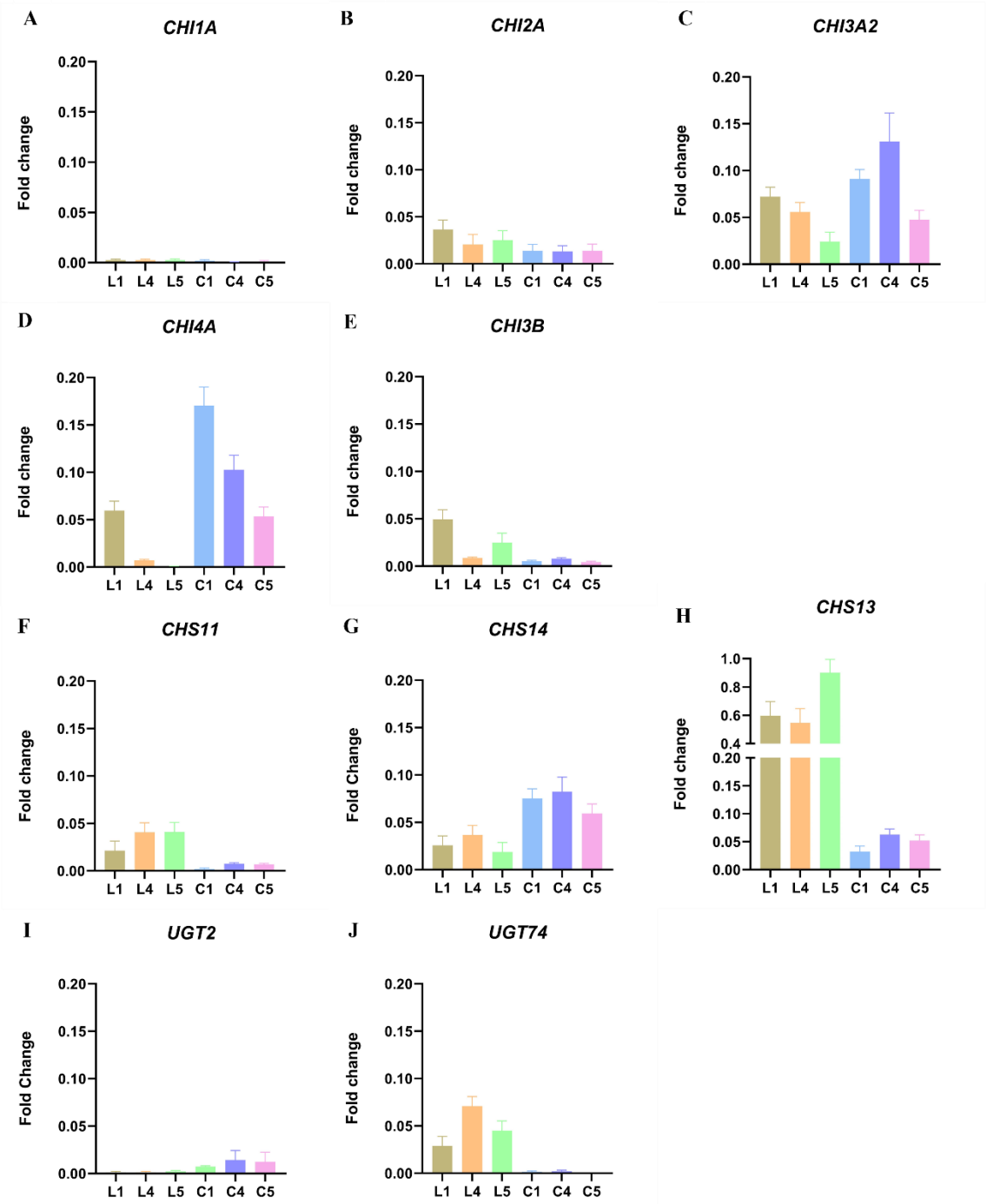

Fig. 4: Expression of the *CHIs*, *CHSs*, *UGTs* families involved in isoflavone synthesis in the leaf and tuberous tissues of *Pueraria mirifica* cultivars. (a) *CHI1A* gene, (b) *CHI2A* gene, (c) *CHI3A2* gene, (d) *CHI3B* gene, (e) *CHI4A* gene, (f) *CHS11* gene, (g) *CHS13* gene, (h) *CHS14* gene, (i) *UGT2* gene, (j) *UGT74* gene;

Cultivar TLBYT (1), NA (4), and SL (5); "L" denotes the leaf, and "C" denotes the tuber.

DISCUSSION

Advances in next-generation sequencing have enabled rapid transcriptome analysis of medicinal plants, providing comprehensive insights into gene regulation and expression networks, particularly in *P. mirifica*. BUSCO analysis indicated that 93.4% of complete and fragmented orthologs were represented in the assembled transcriptome, demonstrating high assembly completeness. Genomic variation accounted for 47.4% of the dataset, within the reported range for eukaryotes (20–60%) and exceeding values previously reported for *P. mirifica* (44.56%), *Arabidopsis* (44%), *G. max* (43%), and *P. lobata* (39.9%). Although *P. mirifica* and *P. lobata* belong to the same genus, BLAST analysis showed higher similarity to *G. max* and *G. soja*, likely due to the limited availability of full-length sequences for *P. lobata*. Differences in unigene distribution across OG, COG, and KEGG databases were also observed compared with earlier studies, with most unigenes associated with cellular components, signal transduction, and translation-related pathways.

DEGs among cultivars were annotated almost in plant hormone signal transduction with about 200 genes and enrichment analysis showed that DEGs enriched the most in "isoflavonoid biosynthesis", and "Zeatin biosynthesis". It is proven for the difference in miroestrol and deoxymiroestrol contents in five cultivars. Even though these plants were planted and harvested under the same conditions, these differences were influenced by the genes involved in miroestrol biosynthesis. The *CHSs*, *CHIs*, and *UGTs* gene families were annotated in *P. mirifica* and a large number of 17 full-length genes sequenced were detected. The *PmCHIs* are an abundant gene family in our study with six members distributed across four subfamilies. According to previous research in Fabaceae, *CHI* type I (*PmCHI2A*, *GmCHI2A*) converts 6'-hydroxychalcone to 5-hydroxyflavone, whilst *CHI* type II (*PmCHI1A*, *PmCHI1B*, *GmCHI1A*, *GmCHI1B*) is in charge of converting 6'-hydroxychalcone and 6'-deoxychalcone to 5-deoxyflavone and 5-hydroxyflavone in that order in the miroestrol biosynthesis pathway (Yin *et al*., 2019). Regarding *PmCHI* type III, *PmCHI3A2* and *PmCHI3B* catalyze fatty acid metabolism. Type IV (*PmCHI4A*) was named after a CHI-like protein and was hypothesized for playing the null function of *CHI*. Lin *et al.* (2021) noted that *CHIL* in Fabaceae may influence the *CHI-CHS* interaction through interaction between proteins to enhance phytoestrogen synthesis (Lin *et al*., 2021).

The previous attempt to understand the *CHSs* diversity found numerous members in the Fabaceae family. Six or seven genes were found in *P. lobata*, eight in *Pisum sativum*, six to eight in *P. vulgaris,* and up to 14 in *G. max* (Pandith *et al*., 2019; Vadivel *et al*., 2018*)*. However, the data for *P. mirifica* seemed less diverse with three divergent genes *CHS11*, *CHS13*, *CHS14* that have the closest genetic relationship with them in *G. max*. *CHS11* and *CHS13* in both *P. mirifica* and *G. max* are specific for leaf expression, while *CHS14* was specific for *P. mirifica* tuber and not recorded in *G. max* tuber. *PmCHS13* and *PmCHS14* formed distinct clades separate from each other and from the remaining *CHS* members. Given that *GmCHS13* and *GmCHS14* have been shown to undergo evolutionary changes resulting in the loss of key functional residues required for phytoestrogen biosynthesis (Vadivel *et al*., 2018), a similar functional divergence may also occur in *P. mirifica*.

UGTs constitute a large multigene family with hundreds of members in legume species. Recent studies have identified a total of 243, 212, 168, 94 authentic *UGTs* for *M. trunculanta*, *G. max*, *P. vulgaris*, and *L. japonicus* (Yin *et al*., 2019; Krishnamurthy *et al*., 2020). However, a small number of members participate in (iso)flavonoid biosynthesis. In *G. max*, Yin *et al.* (2019) indicated six *UGT* genes in subfamilies *UGT72, UGT73, UGT88, UGT92* exhibit activity in the synthesis of (iso)flavone while Wang *et al.* (2016, 2019) detected six *PlUGTs* from subfamilies *UGT2, UGT4, UGT15, UGT20, UGT45, UGT57* (Yin *et al*., 2019; Wang *et al*., 2019; Wang *et al*., 2016). In this study, eight *UGTs* were identified from the *P. mirifica* transcriptome data including *UGT2, UGT57, UGT73* that have been mentioned. Meanwhile, *UGT5* is involved in the biosynthesis of neoandrographolide, a colorless crystalline compound in green chiretta (*Andrographis paniculata*) (Li *et al*., 2018), and *UGT74* catalyzes terpenoid biosynthesis in *Arabidopsis thaliana* and *Panax ginseng* (Kang *et al*., 2018).

Distinct clades in the phylogenetic tree revealed close genetic relationships among homologous sequences across Fabaceae species and enabled gene classification within defined groups. These gene families provide insights into the evolutionary history of *CHI*, *CHS*, and *UGT* families in *P. mirifica*, which shows close genetic relationships with *G. max*, *P. thomsonii*, and *P. lobata*. Gene family evolution is likely driven by selective gene duplication and nucleotide substitution, facilitating functional diversification and adaptation to environmental changes (Pandith *et al*., 2019).

Miroestrol and isoflavone were demonstrated that are highly active phytoestrogens derived from the tuberous roots of *P. mirifica* (Udomsuk *et al*., 2012). In this study, several genes involved in synthesizing these compounds were expressed more highly in leaves (*CHS11, CHS13, UGT74*), some in tuber. Notably, *CHS14* was demonstrated not directly

affect to phytoestrogen synthesis while its gene expressed much more in the tuber. Besides that, *UGT2* was proved to have phytoestrogen catalytic activity and its significant gene expression in the *P. mirifica* tubers showed the correlation.

Nevertheless, gene expression does not always have the same trend as phytoestrogen accumulation. A recent study in miroestrol synthesis by Suntichailamolkul *et al.,* (2019) demonstrated that cytochrome P450 81E*,* Isoflavone reductase*,* and Prenyltransferase genes expressed mainly in leaves, which is not the accumulation location of miroestrol (Suntichaikamolkul *et al*., 2019). The authors assumed that those compounds were generated in mature leaves, transformed to a soluble form of glycoside, and then carried via organs before being deposited in tubers. In 2015, Dastmalchi and Dhaybhadel pointed out isoflavone concentrates are highest in soybeans leaves but its synthesis Isoflavone synthase gene is not expressed highly there (Dastmalchiand Dhaubhadel, 2015). Furthermore, Winkel BS (2004) observed that metabolite accumulation can be regulated by various factors such as rate of synthesis, turnover, transport, and conjugation (Winkel, 2004). In tobaco, nicotines were accumulated in leaves, whereas these compounds are generated in roots (Yazaki *et al*., 2008. In the protein levels, these CHI, CHS, UGT enzymes are the super-molecular complexes so their metabolism is governed by the expression of their constituents, environmental conditions, protein association, and dissociation. Therefore, the transport mechanism of miroestrol intermediates after biosynthesis is yet unknown and should be further studied. Additionally, the comprehensive function of *CHIs*, *CHSs*, *UGTs,* and their characterization in terms of sub-organ and cellular position should be the subject of further investigation.

Conclusions: Overall, the transcriptome of *P. mirifica* from five cultivars have sequenced, assembled and annotated. Accordingly, *CHIs*, *CHSs*, *UGTs* family genes in *P. mirifica* had been annotated, resulting in seventeen full-length sequences along with their characterization. The qRT-PCR results revealed distinct tissue-specific expression patterns: *CHS11*, *CHS13*, and *UGT74* were predominantly expressed in leaves, while *CHI4A*, *CHI3A2*, and *CHS14* showed higher expression in tubers suggesting tissue-specific expression patterns of *CHI*, *CHS*, and *UGT* genes in *P. mirifica.*

Acknowledgment:: This work was funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 106.02-2019.13.

Conflict of interests: The authors declare there were no competing interests.

Author contribution: H.T.T. Huynh, conceived, initiated the project, and designed the experiments; H.T.T. Huynh, C.X. Nguyen, analyzed the data; N.T.B. Nguyen, H.H. Ha, O.T.K. Pham, performed the experiments; H.T.T. Huynh, H.H. Nguyen wrote the article. All authors interpreted the data, critically revised the manuscript for important intellectual contents, and approved the final version.

REFERENCES

Adrian, M.A., G. Manuel, H.G. Gaston and D. Christophe (2013). Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8(1):e53786. https://doi.org/10.1371/journal.pone.0053786.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215(3):403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Austin, M.B. and J.P. Noel. (2003). The Chalcone synthase superfamily of type III polyketide synthases. Nat. Prod. Rep. 20(1):79-110. https://doi.org/10.1039/b100917f.

Dao, T.V., T.H. Nguyen, T.N. Mai, T.V.A. Dang and Q.H. Nguyen. (2018). Evaluation of the estrogenic activity of *Pueraria candollei* var. *mirifica* in ovariectomized rats using vaginal cornification assay. J. Pharm. Res. Drug Inf. 9(2). http://dx.doi.org/10.18203/2319-2003.ijbcp20194255.

Dastmalchi, M. and S. Dhaubhadel (2015). Soybean chalcone isomerase: evolution of the fold, and the differential expression and localization of the gene family. Planta 241(2):507-23. https://doi.org/10.1007/s00425-014-2200-5.

Dennis, A. B., C. Mark, C. Karen, K.M. Ilene, J.L. David, O.l. James and W.S. Eric. (2013). GenBank. Nucleic Acids Res. 41(D1):D36-42. https://doi.org/10.1093/nar/gks1195.

Egorova, K.S. and P.V. Toukach. (2017). CSDB GT: a new curated database on glycosyltransferases. Glycobiology 27(4):285-290. https://doi.org/10.1093/glycob/cww137.

Götz, S., J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talón, J. Dopazo and A. Conesa. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420-3435. https://doi.org/10.1093/nar/gkn176 .

Haas, B.J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R.

Henschel, R.D. LeDuc, N. Friedman and A. Regev (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8(8):1494-1512. https://doi.org/10.1038/nprot.2013.084.

Hasegawa, M., H.Kishino and T. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22(2):160-74. https://doi.org/10.1007/BF02101694.

Haynsen, M. S., M. Vatanparast, G. Mahadwar, D. Zhu, R.Z. Moger-Reischer, J.J. Doyle, K.A. Crandall and A.N. Egan. (2018). De novo transcriptome assembly of *Pueraria montana* var. *lobata* and *Neustanthus phaseoloides* for the development of eSSR and SNP markers: narrowing the US origin(s) of the invasive kudzu. BMC Gen. 19(1):439. https://doi.org/10.1186/s12864-018-4798-3.

He, M., Y. Yao, Y. Li, M. Yang, Y. Li, B. Wu and D. Yu (2019). Comprehensive transcriptome analysis reveals genes potentially involved in isoflavone biosynthesis in *Pueraria thomsonii* Benth. PLoS One 14(6):e0217593. https://doi.org/10.1371/journal.pone.0217593.

Intharuksa, A., M. Kitamura, N. Peerakam, W. Charoensup, H. Ando, Y. Sasaki, and P. Sirisa-Ard (2020). Evaluation of white Kwao Krua (*Pueraria candollei* Grah. ex Benth.) products sold in Thailand by molecular, chemical, and microscopic analyses. J Nat Med. 74(1):106-118. https://doi.org/10.1007/s11418-019-01351-2.

José, A.F. (2007). The Benjamini-Hochberg method in the case of discrete test statistics. Int J Biostat. 3(1):11. https://doi.org/10.2202/1557-4679.1065.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1):27-30. https://doi.org/10.1093/nar/28.1.27.

Kang, K.B., M. Jayakodi, Y.S. Lee, V.B. Nguyen, H.S. Park, H.J. Koo, Y. Choi, D.H. Kim, Y.J. Chung, B. Ryu, D.Y. Lee, S.H. Sung and T.J. Yang (2018). Identification of candidate UDP-glycosyltransferases involved in protopanaxadiol-type ginsenoside biosynthesis in *Panax ginseng*. Sci Rep. 8:1-10. https://doi.org/10.1038/s41598-018-30262-7.

Krishnamurthy, P., C. Tsukamoto and M. Ishimoto (2020). Reconstruction of the evolutionary histories of *UGT* gene superfamily in legumes clarifies the functional divergence of duplicates in specialized metabolism. Int. J. Mol. Sci. 21(5):1855. https://doi.org/10.3390/ijms21051855.

Kumar, S., G. Stecher, M. Li, C. Knyaz and K. Tamura (2018). Mega X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35(6):1547-1549. https://doi.org/10.1093/molbev/msy096.

Li, W. and A. Godzik (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658-1659. https://doi.org/10.1093/bioinformatics/btl158.

Li, Y., H.X. Lin, J. Wan, J. Yang, C.J.S. Lai, X. Wang, B.W. Ma, J.F. Tang, Y. Li, X.L. Li, J. Guo, W. Gao and L.Q. Huang (2018). Glucosyltransferase capable of catalyzing the last step in neoandrographolide biosynthesis. Org. Lett. 20(19):5999-6002. https://doi.org/10.1021/acs.orglett.8b02146.

Lin, L.M., H.Y. Guo, X. Song, D.D. Zhang, Y.H. Long and Z.B. Xing (2021). Adaptive evolution of chalcone isomerase superfamily in fagaceae. Biochem. Genet. 59(2):491-505. https://doi.org/10.1007/s10528-020-10012-z.

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

Love, M.I., W. Huber and S. Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

Malaivijitnond, S. (2012). Medical applications of phytoestrogens from the Thai herb *Pueraria mirifica*. Front Med. 6(1):8-21. https://doi.org/10.1007/s11684-012-0184-8.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17(1):10. https://doi.org/10.14806/ej.17.1.200.

Ngaki, M.N., G.V. Louie, R.N. Philippe, G. Manning, F. Pojer, M.E. Bowman, L. Li, E. Larsen, E.S. Wurtele and J.P. Noel. (2012). Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis. Nature 485:530-533. https://doi.org/10.1038/nature11009.

Pandith, S.A., S.Ramazan, M. I.Khan, Z.A. Reshi and M.A. Shah (2019). Chalcone synthases (*CHSs*): the symbolic type III polyketide synthases. Planta 251(1):15. https://doi.org/10.1007/s00425-019-03307-y.

Suntichaikamolkul, N., K. Tantisuwanichkul, P. Prombutara, K. Kobtrakul, J. Zumsteg, S. Wannachart, H. Schaller, M. Yamazaki, K. Saito, W. De-eknamkul, S. Vimolmangkang and S. Sirikantaramas (2019). Transcriptome analysis of *Pueraria candollei* var. *mirifica* for gene discovery in the biosynthesis of isoflavones and miroestrol. BMC Plant Biol. 19(1):581. https://doi.org/10.1186/s12870-019-2205-0.

Suntichaikamolkul, N., T. Akashi, P. Mahalapbutr, K. Sanachai, T. Rungrotmongkol, J.E. Bassard, H. Schaller, W. De-Eknamkul, S. Vimolmangkang, M. Yamazaki and S. Sirikantaramas (2023). Daidzein hydroxylation by

CYP81E63 is involved in the biosynthesis of miroestrol in *Pueraria mirifica*. Plant Cell Physiol. 64(1): 64-79. https://doi.org/10.1093/pcp/pcac140.

Szeja, W., G. Grynkiewicz and A. Rusin (2017). Isoflavones, their glycosides and glycoconjugates. synthesis and biological activity. Curr Org Chem. 21(3):218-235. https://doi.org/10.2174/1385272820666160928120822.

Udomsuk, L., T. Juengwattanatrakul, K. Jarukamjorn and W. Putalun (2012). Increased miroestrol, deoxymiroestrol and isoflavonoid accumulation in callus and cell suspension cultures of *Pueraria candollei* var. *mirifica*. Acta Physiol. Plant. 34:1093-1100. https://doi.org/10.1007/s11738-011-0906-6.

Vadivel, A.K.A., K. Krysiak, G. Tian and S. Dhaubhadel (2018). Genome-wide identification and localization of chalcone synthase family in soybean (*Glycine max* [L]Merr). BMC Plant Biol. 18(1):325. https://doi.org/10.1186/s12870-018-1569-x.

Wang, X., C. Li, Z. Zhou and Y. Zhang (2019). Identification of three (Iso)flavonoid glucosyltransferases from *Pueraria lobata*. Front. Plant Sci. 10:28. https://doi.org/10.3389/fpls.2019.00028.

Wang, X., R. Fan, J. Li, C. Li and Y. Zhang (2016). Molecular cloning and functional characterization of a novel (Iso)flavone 4′,7-O-diglucoside glucosyltransferase from *Pueraria lobata*. Front. Plant Sci. 7:387. https://doi.org/10.3389/fpls.2016.00387.

Wang, X., S. Li, J. Li, C. Li and Y. Zhang (2015). *De novo* transcriptome sequencing in *Pueraria lobata* to identify putative genes involved in isoflavones biosynthesis. Plant Cell Rep. 34(5):733-743. https://doi.org/10.1007/s00299-014-1733-1.

Winkel, B.S. (2004). Metabolic channelling in plants. Annu. Rev. Plant Biology. 55:85-107. https://doi.org/10.1146/annurev.arplant.55.031903.141714.

Yazaki, K., A. Sugiyama, M. Morita and N. Shitan (2008). Secondary transport as an efficient membrane transport mechanism for plant secondary metabolites. Phytochem. Rev. 7:513-524. https://doi.org/10.1007/s11101-007-9079-8.

Yin, Y., X. Zhang, Z. Gao, T. Hu and Y. Liu (2019). The research progress of chalcone isomerase (*CHI*) in plants. Mol. Biotechnol. 61(1):32-52. https://doi.org/10.1007/s12033-018-0130-3.

Yusakul, G., W. Putalun, O. Udomsin, T. Juengwatanatrakul and C. Chaichantipyuth (2011). Comparative analysis of the chemical constituents of two varieties of *Pueraria candollei*. Fitoterapia 82(2):203-207. https://doi.org/10.1016/j.fitote.2010.09.009.