

De novo transcriptome assembly reveals putative biosynthetic genes involved in the biosyntheses of isoflavones and miroestrol in *Pueraria candollei* var. *mirifica*

Nithiwat Suntichaikamolkul

Chulalongkorn University Faculty of Science

Kittiya Tantisuwanichkul

Chulalongkorn University Faculty of Science

Pinidphon Prombutara

Chulalongkorn University Faculty of Science

Khwanlada Kobtrakul

Chulalongkorn University Faculty of Science

Julie Zumsteg

Institut de Biologie Moleculaire des Plantes

Siriporn Wannachart

Kasetsart University Kamphaeng Saen Campus

Hubert Schaller

Institut de Biologie Moleculaire des Plantes

Mami Yamazaki

Chiba University

Kazuki Saito

Chiba University

Wanchai De-eknamkul

Chulalongkorn University Faculty of Pharmaceutical Sciences

Sornkanok Vimolmangkang

Chulalongkorn University Faculty of Pharmaceutical Sciences

Supaart Sirikantaramas (✉ supaart.s@chula.ac.th)

<https://orcid.org/0000-0003-0330-0845>

Research article

Keywords: Relative expression of the candidate MYB TFs involved in isoflavone biosynthesis across four tissues of *P. mirifica* by qRT-PCR. Tissue abbreviations: Y; young leaves, M; mature leaves, T; cortex-

excised tubers, C; tuberous cortices. Significantly different expression levels of putative genes (ANOVA, post-hoc Duncan test, $P < 0.05$) have different lowercase letters

Posted Date: July 26th, 2019

DOI: <https://doi.org/10.21203/rs.2.12015/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published on December 26th, 2019. See the published version at <https://doi.org/10.1186/s12870-019-2205-0>.

Abstract

Background *Pueraria candollei* var. *mirifica*, a Thai medicinal plant used traditionally as a rejuvenating herb, is known as a rich source of phytoestrogens, including isoflavonoids and the highly estrogenic miroestrol and deoxymiroestrol. Although these active constituents in *P. mirifica* have been known for some time, actual knowledge regarding their biosynthetic genes remains unknown. **Results** A de novo transcriptome analysis was conducted using combined *P. candollei* var. *mirifica* tissues of young leaves, mature leaves, tuberous cortices, and cortex-excised tubers. The assembly of 166,923 contigs was used for a functional annotation against protein databases, and as a library for the identification of genes possibly involved in the biosynthesis of isoflavonoid and miroestrol. Twenty-one differential expressed genes (DEGs) from four separate libraries were identified as candidates involved in the biosynthesis pathways and had their expressions validated by qRT-PCR. Notably, isoflavonoid profiling generated by LC-MS/MS was positively correlated to the expression levels of isoflavonoid biosynthetic genes across the four types of tissues. In addition, R2R3 MYB transcription factors were also identified that may be involved in the regulation of isoflavonoid biosynthesis in *P. candollei* var. *mirifica*. To confirm the function of the key-isoflavone biosynthetic gene, *P. candollei* var. *mirifica* isoflavone synthase (PmIFS) identified in our library was transiently co-expressed with the Arabidopsis MYB12 transcription factor (AtMYB12) in *Nicotiana benthamiana* leaves. Remarkably, the combined expression of those proteins led to the production of isoflavone genistein. **Conclusions** Our results provide compelling evidence regarding the integration of transcriptome and metabolome as a powerful tool for unraveling biosynthetic pathways in plants.

Background

White Kwao Krua (*Pueraria candollei* var. *mirifica*, hereafter shortened to *P. mirifica*), a Thai medicinal plant, has been extensively used in Thai traditional medicine as a rejuvenating herb because of its numerous phytoestrogenic constituents [1]. Phytoestrogens are plant-derived compounds that structurally or functionally mimic mammalian estrogen and play an important role in the prevention of cancers, heart disease, menopausal symptoms, and osteoporosis [2]. Due to the numerous effects phytoestrogens have on human health, it is not surprising to consider using *P. mirifica* for treating various diseases and for increasing the development of commercial products.

Chemically, there are three major types of phytoestrogens found in *P. mirifica*: isoflavones, coumestans, and chromenes [3–5]. Isoflavones are an important type of phytoestrogens that are derived from the phenylpropanoid pathway and biosynthesized predominantly in leguminous plants [6]. Seven isoflavones that have been identified in *P. mirifica* tubers are puerarin, daidzin, genistin, daidzein, genistein, kwakhurin, and mirificin [3, 5]. Four coumestans also reported in the tuber of this plant are coumestrol, mirificoumestan, mirificoumestan hydrate, and mirificoumestan [7]. The structures of the chromenes have received a large amount of attention because of their low toxicity and broad pharmacological uses as anticancer, antimicrobial, and anti-inflammatory agents [8]. Despite being accumulated at very small levels [9], miroestrol and its precursor deoxymiroestrol are the major chromenes specifically found in the

tuberous cortex of *P. mirifica* [10]. Both exhibit the highest estrogenic activity among other isolated phytoestrogens [11]. Since both chromenes have not been reported in other plant species, a unique evolution on its biosynthesis in *P. mirifica* can be suggested. Interestingly, although miroestrol has been known for almost six decades, its biosynthetic pathway and enzymes are still unknown.

In terms of biosynthesis of these phytoestrogens, isoflavonoids generally share the early steps with flavonoid biosynthesis, starting with phenylalanine ammonia lyase (PAL) removing amides from the first substrate, phenylalanine, to produce cinnamic acid that is hydroxylated by cinnamate 4-hydroxylase (C4H), to form *p*-coumarate. The enzyme 4-coumarate-CoA ligase (4CL) further activates the *p*-coumarate by attaching a CoA. Next, chalcone synthase (CHS) carries out the condensation of *p*-coumaroyl-CoA with three molecules of malonyl-CoA to form the chalcone skeleton. The chalcone can be converted to flavanone by the enzyme chalcone isomerase (CHI). Liquiritigenin is one of the shared substrates between the flavonoid and isoflavonoid pathway. Isoflavonoids are generally synthesized from common intermediates (either liquiritigenin or naringenin) within the recognized flavonoid biosynthetic pathway by aryl migration, which is catalyzed by isoflavone synthase (IFS). This pathway leads to the production of the intermediate, 2-hydroxyisoflavanone, which is then dehydrated to daidzein through catalysis of 2-hydroxyisoflavanone dehydratase (HID) (Fig. 1) [12]. Although CHS and IFS have been identified in *P. mirifica* several years ago [13, 14], their enzymatic functions have not been characterized. Udomsuk et al. (2011) [15] have proposed miroestrol biosynthesis as possibly sharing a common pathway with the isoflavonoid biosynthesis due to their similar backbone structure. Currently, biosynthetic genes and enzymes involved in miroestrol biosynthesis have not been reported or identified. These include the transcription factors responsible for regulating the expression of these biosynthetic genes. The V-myb myeloblastosis viral oncogene homolog transcription factors (MYB TFs), which are the largest plant transcription factor families, have been reported to possess key functions in regulating the synthesis of phenylpropanoid-derived compounds in plants [16]. These proteins have generated a great attention for increasing phytoestrogen biosynthesis in plants.

Recently, the construction of a transcriptome using next generation sequencing technology in various plants has demonstrated a potential source for successfully identifying genes involved in the biosynthesis of various secondary metabolites. The transcriptome of *Pueraria lobata*, a closely related species of *P. mirifica*, has been reported [17, 18]. Although these two plants produce similar types of isoflavones, miroestrol can only be found in *P. mirifica*. In those two studies, *P. lobata* genes encoding the core isoflavone biosynthetic enzymes were identified and their expression levels in various tissues were profiled.

In the current study, we carried out high-throughput Illumina sequencing to obtain *P. mirifica* transcriptome libraries from young leaves, mature leaves, cortex-excised tubers, and tuberous cortices, and used *de novo* transcriptome assembly to propose a new miroestrol biosynthetic pathway. Additionally, MYB TFs involved in isoflavonoid biosynthesis were also identified. Integration of transcriptomics and metabolomics provides new insights into the prediction and identification of putative

biosynthetic genes and transcription factors possibly involved in *P. mirifica* isoflavonoid and miroestrol biosynthetic pathways.

Results

De novo transcriptome assembly of *P. mirifica*

To obtain nucleotide sequences of expressed genes in various tissues of *P. mirifica*, we constructed a cDNA library from pooled tissues including, young leaves, mature leaves, cortex-excised tubers, and tuberous cortices (Fig. 2a–2d). The library was processed using Illumina Hiseq 2000 platform, yielding approximately 8.2 Gbps. Unknown nucleotides in clean reads were not detected, confirming that our sequencing was efficient for *de novo* assembly. The clean reads were assembled *de novo*, resulting in 166,923 contigs (62,567,517 nt) and 104,283 unigenes (81,810,584 nt), which are summarized in Supplementary Table S1.

Protein functional annotation and classification

Functional annotation of the assembled transcripts provides insights into the diverse molecular functions and biological processes within an organism. The functional annotation of *P. mirifica* assembled transcripts was created based on the similarities with proteins/transcripts in various public databases. The statistics of functional annotation were summarized in Supplementary Table S2. Aligned unigenes showed a significant homology against an NCBI non-redundant (nr) protein database (Fig. 3a). Based on the BLAST similarity distribution, over 68% of unigenes exhibited similarities greater than 80% (Fig. 3b). A top-hit species distribution analysis showed unigenes with BLAST hits sharing high sequence similarities to *Glycine max* (85.3%), *Medicago truncatula* (6.2%), and *Vitis vinifera* (1.6%; Fig. 3c).

The clusters of orthologous groups (COG) indicated the functional classification of each unigene at the cellular level. There were 1,091 unigenes predicted to be involved in secondary metabolite biosynthesis (Fig. 4). Gene ontology (GO), which classifies standardized gene function, is useful for annotating gene functions and gene products in any organism. GO contains three main dependent categories: biological processes, molecular functions, and cellular components. The 37,058 unigenes yielded a corresponding GO term that can be further classified into 56 sub-categories: 23 related to biological processes, 17 to cellular components and 16 to molecular functions. Supplementary Figure S1 shows the large number of transcripts related to cellular components and metabolic processes. The remaining 67,225 unigenes harbored no BLAST hits. These unannotated unigenes may be uncharacterized genes, or assembled sequences that were too small to produce hits.

KEGG provides a robust instrument for biological pathway assignment, as well as the functional annotation of gene products. This pathway-based annotation provides an overview of the different metabolic processes active within an organism, and enables further understanding of the biological functions of the unigenes. All unigenes were analyzed in the KEGG pathway with an *E*-value cutoff of less

than 10^{-5} . We obtained 33,317 unigenes, where 3,997 unigenes were related to biosynthesis of secondary metabolite and 8,103 to the general metabolic pathways. The top-five ranking pathways were plant hormone signal transduction (2,061 unigenes), endocytosis (1,254 unigenes), RNA transportation (1,232 unigenes), glycerophospholipid metabolism (1,131 unigenes), and purine metabolism (1,113 unigenes). Regarding the significant capacity of leguminous plants to accumulate functional flavonoids, 476 flavonoid biosynthetic unigenes and 167 isoflavonoid biosynthetic unigenes are shown in Supplementary Figure S2. These functional annotations were used for the identification of genes involved in isoflavonoid biosynthesis in *P. mirifica*.

Proposed miroestrol biosynthetic pathway, differential accumulation of transcripts associated with isoflavonoids, and miroestrol biosynthesis as determined by RNA-Seq and qRT-PCR

Miroestrol biosynthesis possibly shares a common pathway with isoflavonoid biosynthesis [15]. We propose that daidzein, a common isoflavone aglycone, would be hydroxylated by at least two cytochrome P450 enzymes at the 2' and 3' of the B-ring to yield 2',5'-dihydroxydaidzein, possibly into a CYP81E subfamily which has been known to use isoflavones as substrates [19–22]. The 2',5'-dihydroxydaidzein would then be reduced by the enzyme isoflavone reductase (IFR) and subsequently prenylated by a prenyltransferase (PT) using dimethylallyl diphosphate (DMAPP) as a co-substrate, to obtain deoxymiroestrol and miroestrol, respectively (Fig. 1).

Based on a functional annotation of each unigene found in the *P. mirifica de novo* transcriptome assembly and phylogenetic analysis, a total of 14 putative genes involved in isoflavone biosynthesis were predicted as follow; two *PALs*, one *C4H*, three *4CLs*, two *CHSs*, one *CHR*, two *CHIs*, one *IFS*, and two *HIDs*. In addition, a total of seven putative genes were also identified in our proposed miroestrol biosynthetic pathway; three *CYP81Es*, two *IFRs*, and two *PTs*. To investigate the gene expression levels of those unigenes across four tissues of *P. mirifica*, reads per kilobase million (RPKM) of transcripts were calculated from the NextSeq500 generated RNA-Seq. The relative differential expressed gene (DEG) analysis of those unigenes ($q = 0.9$) was represented in a heat map in our proposed miroestrol biosynthetic pathway (Fig. 1).

To validate the DEG analysis obtained from our RNA-Seq, candidate unigenes were selected and analyzed at the expression level using the qRT-PCR technique in young leaves, mature leaves, cortex-excised tubers, and tuberous cortices of *P. mirifica*. As expected, the expression levels of the selected unigenes obtained from qRT-PCR were positively correlated to the DEGs performed by RNA-Seq (Fig. 5). Indeed, most of the isoflavone biosynthetic genes were highly expressed in tuberous cortices in comparison with the young leaves, mature leaves, and cortex-excised tuber. However, *PmCYP81Es*, *PmIFRs*, and *PmPT* were highly expressed in mature leaves compared to the other tissues.

Phytoestrogens and annotated constituents in *P. mirifica*

HPLC coupled to tandem MS is a powerful platform that offers high selectivity and sensitivity. Seven phytoestrogens in *P. mirifica* were identified by comparing the standard compounds based on their retention times and MS fragmentation patterns. These included daidzein (RT = 16.33 min, m/z = 254.05791), daidzin (RT = 10.16 min, m/z = 416.11073), genistein (RT = 18.78 min, m/z = 270.05282), genistin (RT = 13.70 min, m/z = 432.10565), 2'-hydroxydaidzein (RT = 14.74 min, m/z = 270.05282), 3'-hydroxydaidzein (RT = 15.18 min, m/z = 270.05282), and puerarin (RT = 10.40 min, m/z = 416.11073). Due to the lack of standard compounds, mirificin, 2',5'-dihydroxydaidzein, deoxymiroestrol, and miroestrol were predicted on the basis of mass accuracy ranges and MS fragmentation patterns searches conducted on the METLIN metabolomics database or/and comparisons with published data [23]. The relative abundance of the major phytoestrogens found in young leaves, mature leaves, cortex-excised tubers, and tuberous cortices of *P. mirifica* are shown in a heat map in Figure 2e. Genistein, genistin, and the annotated 2', 5'-dihydroxydaidzein were highly accumulated in mature leaves. Whereas, daidzein, puerarin, the annotated mirificin, and the annotated deoxymiroestrol were highly accumulated in tuberous cortices compared to the other tissues. Although miroestrol was not detected in any tissues, its precursor (the annotated deoxymiroestrol; RT = 15.03 min, m/z = 342.14672) was found only in the *P. mirifica* tuberous cortex, the main accumulation site of miroestrol [10].

Isoflavone production in *PmIFS* overexpressing *N. benthamiana* leaves

Recently, *N. benthamiana* has been widely used to identify transient expressions of several plant genes to confirm gene functions [24]. To demonstrate that our transcriptome libraries contain candidate genes involved in the *P. mirifica* phytoestrogen biosynthesis, we cloned *P. mirifica* isoflavone synthase (*PmIFS*) and performed a functional characterization using a transient (co)expression in *N. benthamiana*, which does not produce any isoflavones. The transient expression of the green fluorescent protein (GFP) in five week old *N. benthamiana* leaves was used as a negative control. The co-expression of the *PmIFS* and *Arabidopsis* R2R3 MYB12 transcription factor (*AtMYB12*), a regulator enhancing metabolic flux to flavonoid biosynthesis [25], generated two novel major peaks identified as genistein (Supplementary Figure S4), suggesting that *PmIFS* is involved in isoflavone biosynthesis.

Identification of MYB transcription factors involved in isoflavone biosynthesis

We identified 85 putative genes encoded as MYB transcription factors from the *P. mirifica* transcriptome. All putative *P. mirifica* MYB transcription factors (PmMYB) were aligned to characterized MYB transcription factors from other plants such as *Arabidopsis thaliana*, *Glycine max*, plus other plant species, and are shown in a phylogenetic tree (Fig. 6). Based on this phylogenetic analysis, we found six candidate PmMYBs possibly involved in regulating the biosynthesis of isoflavonoid. The PmMYB18 was closely related to GmMYB29, which activates the expression of *IFS* and *CHS* in soybean [26]. PmMYB23

and PmMYB24 were grouped together with GmMYB12B2 and regulate expression of *CHS* in soybean [27, 28]. The last three interesting putative MYB transcription factors, PmMYB75, PmMYB76, and PmMYB77, were clustered with GmMYB176, which is also involved in controlling *CHS* and isoflavonoid synthesis in soybean [29, 30]. We also validated *PmMYB24* and *PmMYB77* as being highly expressed in mature leaves whereas, *PmMYB18*, *PmMYB23*, *PmMYB75*, and *PmMYB76*, were highly expressed in tuberous cortices (Fig. 7). These PmMYBs may also play a role in regulating isoflavonoid biosynthesis in different tissues.

Discussion

De novo transcriptome assembly provides an excellent platform to generate a comprehensive gene resource about an organism without whole genome sequencing. In addition, this technique allows for the discovery of novel genes, molecular markers, and tissue-specific expression patterns. Lacking the genome database of *P. mirifica*, RNA-seq was used to profile the *P. mirifica* transcriptome. Although a typical HiSeq platform exhibits various advantages in large-scale transcriptomics research, it requires more time to process the output data. In contrast, a NextSeq platform is designed as a faster and easier operating benchtop that reduces cost and data processing times. Furthermore, the overall error rates (<1%) are quite similar in both platforms [31]. Thus, we employed the Illumina HiSeq 2000 platform to construct the whole transcriptome library from four tissues, and NextSeq 500 to determine the expression patterns of genes in each library constructed from four tissues of *P. mirifica*. Notably, the candidate genes involved in the biosynthesis of isoflavonoid and miroestrol were successfully identified in the libraries sequenced by both platforms. We obtained a large dataset of unigenes in our *de novo* transcriptome assembly generated from the HiSeq platform. The mean GC content of those unigenes was 44.56%. In eukaryotes, average GC content ranges from approximately 20 to 60% [32]. The *P. mirifica* GC content is in the middle of this range, slightly higher than those reported for *P. lobata* (39.9%) [17] and *Medicago truncatula* (40%) [33]. In addition, these values are very close to those reported for *Glycine max* (43%) [33]. Furthermore, most of annotated unigenes were matched to the *Glycine max* protein database, indicating that the relative functions are conserved within these two species. Although *P. lobata* is considered a closely related species of *P. mirifica*, only 0.12% of annotated unigenes were matched to *P. mirifica*. This is due to the limited number of full-length sequences of *P. lobata* being deposited in the protein database, and the available *P. lobata* transcriptome data was deposited as raw sequencing data in NCBI [17, 18]. In addition, the gene or protein names, the descriptions of COG and GO terms, and possible metabolic pathways, were also annotated. Notably, all candidate genes generated from the HiSeq platform were found in the four libraries generated from NextSeq platform. Their expressed levels across the four tissues of *P. mirifica* can be used to reduce the candidates for further functional characterization.

Isoflavones are a class of flavonoid that are produced almost exclusively in leguminous plants [6]. In the *P. mirifica* transcriptome dataset, 21 putative genes involved in isoflavonoid biosynthesis were identified. The DEG values of these unigenes that were validated by quantitative RT-PCR, exhibited high expression levels in tuberous cortices compared to those of the other three tissues. Consequently, almost all isoflavones, such as puerarin and daidzin, also showed the highest accumulation in the tuberous

cortices. When comparing *P. mirifica* and *P. lobata*, gene expression and the isoflavones accumulation profile in each tissue were similar [17, 18], except for *P. lobata* not expressing *IFS* or accumulating isoflavonoids in leaves [17]. This observation suggests a divergent evolution pattern in these two closely related species that are found in different regions (temperate zones for *P. lobata* and tropical zones for *P. mirifica*). It is possible that *P. mirifica* has evolved a mechanism to accumulate isoflavonoids in leaves and miroestrol in tuberous cortices possibly for defense purposes. To demonstrate the existence of candidate genes in our transcriptome data, we cloned *PmlFS* into *N. benthamiana*, which does not naturally produce isoflavonoids, and transiently co-expressed it with *AtMYB12*, an activator in flavonoid biosynthesis [25]. The cloning led to an accumulation of genistein and genistin. The demonstrated *in planta* function of *PmlFS* indicated that our transcriptome libraries are usable and suitable for the discovery of other genes involved in miroestrol biosynthesis. Although a tentative biosynthetic pathway of miroestrol has been mentioned elsewhere [15], genes required for miroestrol biosynthesis remain completely unknown. Herein, the biosynthetic pathway of miroestrol was reconsidered and the most plausible scenario was proposed (Fig. 1). Seven putative genes encoding three biosynthetic enzymes involved in our proposed miroestrol biosynthetic pathway were also identified in our transcriptome dataset. As previously reported, the accumulation of deoxymiroestrol and miroestrol was higher in the tuberous cortices than in the cortex-excised tubers of *P. mirifica* [10]. Similarly, we detected deoxymiroestrol specifically in the tuberous cortices of *P. mirifica*. However, miroestrol was not detected. It has been reported that miroestrol can be non-enzymatically converted from deoxymiroestrol by high temperature or within acidic/alkaline solutions during storage and extraction [34, 35]. Indeed, in our experiment, *P. mirifica* tissues were harvested and rapidly frozen in liquid nitrogen that would preserve deoxymiroestrol without any conversion to miroestrol. In addition, the proposed intermediates 2'-hydroxydaidzein and 3'-hydroxydaidzein, were not detected in any tissues. We hypothesized that an organized multienzyme cluster, known as metabolon, might be involved in miroestrol biosynthesis and be responsible for the absence of those cryptic biosynthetic intermediates. In fact, an increasing number of publications describe the organization of biosynthetic pathways as metabolons. Particularly, a recent study showed the efficient and specific substrate transformation in a metabolon complex between *IFS* and *CHR* in soybean [36]. In addition, the formation of a metabolon channeling substrate towards a product without the leakage of labile intermediates has been known [37].

Regarding the putative genes involved in miroestrol biosynthesis in *P. mirifica*, we found that *CYP81E*, *IFR*, and *PT* were highly expressed in mature leaves, which were not the accumulation sites of miroestrol and deoxymiroestrol. Perhaps, those intermediates or deoxymiroestrol would be initially synthesized in mature leaves and converted to a soluble form of glycoside, readily transported across organs, and stored mainly in tuberous cortices. In fact, a glycosylated form of miroestrol has been identified in *P. mirifica* tuber [38]. Examples of such transports are known: in tobacco, nicotine is transported from the roots, which is the true biosynthetic site, to the leaves where it accumulates. It is generally known that secondary metabolites are translocated between plant cells through a secondary transporter and accumulated in the appropriate tissues or organs [39]. However, the transport of intermediates of deoxymiroestrol is still unclear and could be the subject of further investigations. Nevertheless, a deeply functional

characterization of biosynthetic genes and isotopic labeling of any intermediates must be accomplished to confirm this hypothesis. Alternatively, we do not exclude the possibility that the intermediates and deoxymiroestrol could also be biosynthesized in a tuberous cortex, because the corresponding biosynthetic genes were detected at levels lower than found in mature leaves. The lower expression levels of the putative miroestrol biosynthetic genes could be partially responsible for the low accumulation levels of deoxymiroestrol and miroestrol. In addition, we found extremely high amounts of puerarin in the tuberous cortices. Since puerarin, plus and deoxymiroestrol and miroestrol, share the same early biosynthetic pathways (Fig. 1), daidzein may be primarily fluxed into the puerarin biosynthetic pathway, leaving lower daidzein levels for miroestrol biosynthesis. Concurrently, we phylogenetically identified a total of five putative genes annotated as a *C*-glycosyltransferase, which plays a key role in producing 8-*C*-glycosylation [40] for the puerarin biosynthetic pathway. One of the putative genes, CL7002, exhibits 92.32% similarity to the characterized *P. lobata* *C*-glycosyltransferase (CGT43). These putative genes were highly expressed in tuberous cortices as compared to the other tissues (Supplementary Figure S3). Additionally, several glycosyltransferases have been shown to form a metabolon complex for facilitating efficient production of bioactive compounds [41–43]. These observations could contribute predominantly to daidzein utilization for producing puerarin. Suppression of *PmCGT* might be an alternative strategy to enhance miroestrol production in tubers.

Another strategy to enhance the production of isoflavonoids and their valuable derivatives is to manipulate TF regulation. In this study, we identified MYB TF, which generally acts either as a transcriptional activator or repressor for various groups of plant secondary metabolites, including isoflavonoids. Amino acid alignment and expression patterns are informative for determining the function of individual MYB TF. Six candidate MYBs (highlighted in red in the phylogenetic tree; Fig. 6) have been proposed due to their phylogenetic clustering with characterized MYB TFs that are involved in the regulation of isoflavonoid biosynthetic genes. Interestingly, four were highly expressed in the tuberous cortex, which had the highest accumulation of isoflavone (Fig. 7). Moreover, two of these four, *PmMYB18* and *PmMYB75*, connectedly shared the same phylogenetic branch with the MYBs reported as activators for CHR and IFS transcription and increased isoflavonoid production[26, 30]. These two MYBs were eligible candidates for further functional characterization. The candidate *PmMYB* TFs could possibly be involved in the regulation of isoflavonoid biosynthesis and subsequently influence miroestrol production.

Conclusion

In summary, we constructed *de novo* transcriptome libraries that represented a valuable gene resource of *P. mirifica* and a powerful tool for unraveling biosynthetic pathways in plants. We identified several candidate genes encoding key enzymes or TFs involved in the biosynthesis of isoflavonoid and miroestrol. Integration of transcriptomics and metabolomics revealed the complexity of gene expression and metabolite profiles across tissues, confirming the hypothesis of cortical tuber tissue being a major miroestrol biosynthetic site. Further molecular and biochemical studies on candidate genes involved in miroestrol biosynthesis will be necessary to confirm their function(s).

Methods

Plant materials, chemicals and reagents

Leaves and tubers from approximately two-year old *P. mirifica* cultivar SARDI190 (Fig. 2a–2d) were harvested from a demonstration *P. mirifica* farm at Kasetsart University, Kamphaeng Saen Campus, Nakhonpathom, Thailand (141N, 9958E). The samples were authenticated as *P. mirifica* by comparing with the voucher specimen numbers BCU010250 and BCU011045, kept at the Professor Kasin Suvatabhandhu Herbarium, Department of Botany, Faculty of Science, Chulalongkorn University, Thailand. Research on *P. mirifica* has been approved from Department of Thai Traditional and Alternative Medicine, Ministry of Public Health, Thailand (DTAM1–2/2561). For *Nicotianabenthamiana*, seeds were germinated on sterilized soil in a culture room at 25 °C with a 16/8 h light/dark cycle. After germinating for five days, the *N. benthamiana* seedlings were watered twice a week until harvested. Standard compounds of 2'-hydroxydaidzein [19] and 3'-hydroxydaidzein [19] were gifts from Dr. Tomoyoshi Akashi, Nihon University, Japan. The daidzein and genistein were purchased from CALBIOCHEM (Merck, Germany), and daidzin, genistin, and puerarin were purchased from Sigma-Aldrich (USA). All other chemicals were commercially available products of analytical-reagent grade.

Construction of *de novo* transcriptome assembly from the combined RNA of *P. mirifica* using Illumina Hiseq 2000

- Total RNA from young leaves, mature leaves, cortex-excised tubers, and tuberous cortices were isolated using an RNeasy plant mini kit (Qiagen, USA) and on-column DNA were removed with an RNase-free DNase set (Qiagen, USA). The total RNA quality was checked by running the samples on agarose gel electrophoresis and the quantity was measured using a biospectrometer (Eppendorf, Germany). Equal quantities of high-quality RNA from the four tissues were combined and used for sequencing by Illumina Hiseq 2000 (BGI, Hongkong). Filtering of raw reads were performed using filter fq, an internal software of BGI, resulting in paired end and clean reads, and unpaired reads that lost corresponding sequence partners due to quality control. The Illumina reads were then assembled into contigs and unigenes using Trinity software [44]. The raw transcriptome data have been deposited at the sequence read archive (SRA) of the National Center for Biotechnology Information (accession number: SRR6917866)
- Analysis of gene expression levels across the four tissues of *P. mirifica* generated using Illumina NextSeq500

For RNA-seq analysis, the adjusted quantities of high-quality RNA from young leaves, mature leaves, cortex-excised tubers, and tuberous cortices of *P. mirifica* were sequenced using Illumina NextSeq500. Reads were assembled into contigs using Trinity software [44]. To estimate contig expression levels, we mapped all fastq format reads back to the contigs and calculated the reads per kilobase of the transcript per million mapped reads (RPKM) values. The non-parametric approach for the identification of

differentially expressed genes, NOISeq-sim [45], was applied to analyze four independent pairwise sample comparisons.

- Functional annotation of assembled sequences

The assembled sequences were first compared against the NCBI non-redundant (nr) protein database (<https://www.ncbi.nlm.nih.gov/refseq/>), Swiss-Prot (<https://www.ebi.ac.uk/uniprot>), Kyoto Encyclopedia of Genes and Genomes (KEGG; <http://www.kegg.jp/kegg/>), and clusters of orthologous groups (COG; E -value < $1E-5$) (<https://www.ncbi.nlm.nih.gov/COG/>) by BLASTx, and nucleotide database NT (E -value < $1E-5$) (<https://www.ncbi.nlm.nih.gov/nucleotide/>) by BLASTn. Then, the number of annotated unigenes with each database were counted. The unigenes were mapped to the COG database, predicted the possible functions and statistics. BLAST2GO program v2.5.0 [46] was applied to assign GO categories based on the BLAST results with default settings. Visualization of all unigene GO functional categories and the distribution of gene functions in the different species was prepared with the WEGO program [47].

Phylogenetic analysis

All putative genes were translated to amino acid sequences with the Expasy translate tool (<http://web.expasy.org/translate/>). All full-length putative sequences involved in isoflavonoid and chromene biosyntheses were aligned using the BioEdit software (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). For the analysis of evolutionary relationships, phylogenetic trees using the Neighbor-Joining method with 1000 bootstrap replicates were constructed using the MEGA7 software [48].

Validation of candidate gene expression using qRT-PCR

The transcript levels of candidate genes were evaluated using quantitative real-time reverse transcription polymerase chain reaction (qRT-PCR). Total RNA from *P. mirifica* young leaves, mature leaves, cortex-excised tubers, and tuberous cortices were used (Fig. 2a–2d) to synthesize single stranded cDNA using the RevertAid First Strand cDNA synthesis Kit (Thermo Fisher, USA). The qRT-PCR was performed using an equal amount of cDNA in a CFX96 Touch™ Real-Time PCR Detection system using iTaq Universal SYBR Green Supermix (Biorad, USA) with the following conditions: 95 °C for 8 min, 40 cycles of 95 °C for 10 s, 60 °C for 15 s and 72 °C for 30 s. The specific primers used for amplification of the candidate genes are shown in the Supplementary Table S3. The *P. mirifica* unigene encoding EF1α was amplified as an internal control. Each reaction was completed with three biological replicates. The results were calculated as the relative expression ratios of the candidate genes expressed in leaves versus tuberous roots, which were normalized to a reference gene (EF1α) transcript level in the same cDNA sample [49].

Metabolite extraction of *P. mirifica* samples

The *P. mirifica* samples were frozen, ground to a fine powder, and lyophilized overnight. A 30 mg sample of the dried powder was dissolved in 500 μ L of 85% methanol to extract metabolites by vigorous shaking with a Mixer Mill MM400 (Retsch, Germany) at 25 hertz for 7 min. Crude metabolic extracts were filtered through an Acrodisc[®] Syringe Filter with 0.2 μ m Supor[®] membrane (Pall, USA) and stored at -20 °C for further metabolite analysis.

Phytoestrogen profile by high performance liquid chromatography quadrupole time of flight mass spectrometry (HPLC-QTOF-MS/MS)

The HPLC-QTOF-MS/MS analysis was performed on an Agilent HPLC 1260 series coupled with a QTOF 6540 UHD accurate mass (Agilent Technologies, Germany). The separation of the sample solution was performed on a Luna C18 (2) 150 $\times$ 4.6 mm, 5 μ m (Phenomenex, USA). The solvent flow rate was 0.5 mL/min, with 5 μ L of the sample solution being injected into the LC system. The binary gradient elution system was composed of water as solvent A and acetonitrile as solvent B, with both solvents containing 0.1% of formic acid (v/v). The linear gradient elution was 5–95% for solvent B at 35 min and a post run for 5 min. The column temperature was set at 40 °C. The conditions for the negative ESI source were a drying gas (N_2) flow rate of 10 L min⁻¹, drying gas temperature at 350 °C, nebulizer of 30 psig; fragmentor set to 100 V, capillary voltage at 3500 V, and scan spectra from m/z 100–1500. The auto MS/MS for the fragmentation was set with collision energies of 10, 20, and 40 V. All data analyses were controlled using Agilent MassHunter Qualitative Analysis Software B06.0 (Agilent Technologies, USA).

Molecular cloning and transient expression of *P. mirifica* isoflavone synthase (*PmIFS*) and *Arabidopsis* MYB12 (*AtMYB12*) in *N. benthamiana*

The unigene encoding isoflavone synthase (*PmIFS*; accession number MK524721) was amplified from *P. mirifica* cDNA using the following gene specific primers; forward 5'-ATGTTGCTGGAAGCTTGCAATTG-3' and reverse 5'-TCAAGAAGGAGGTTTAGATGC-3'. For *Arabidopsis* MYB12 (*AtMYB12*; accession number NM130314), the full-length gene was amplified from *Arabidopsis* cDNA using the following gene specific primers; forward 5'-ATGGGAAGAGCGCCATGT-3' and reverse 5'-TCATGACAGAAGCCAAGCG-3'. The PCR was done in a 50 μ L reaction using a high fidelity DNA polymerase (Phusion[®] HF, Finland).

Thermocycling involved initial denaturation at 98 °C for 1 min, followed by 35 cycles of 10 s at 98 °C, 30 s at 60 °C, 1 min at 72 °C, with a final extension at 72 °C for 5 min. The PCR products were checked by running in an agarose gel electrophoresis. Amplicons of the desirable size were ligated into a pJET1.2 vector (Fermentas, USA) for sequence verification. The verified genes were sub-cloned into pDONOR207 and pEAQ-HT-DEST1 [50], respectively, using a gateway cloning system according to manufacturer's instructions (Invitrogen, USA). The recombinant pEAQ-HT-DEST1 harboring *PmIFS* and pEAQ-HT-DEST1 harboring *AtMYB12* were electro-transformed into *Agrobacterium tumefaciens* LBA4404. The positive

clones were chosen by colony PCR. A single colony of each transformant was grown overnight (28 °C, 250 rpm) in 5 mL YEB medium amended with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin. The overnight cultures were washed three times with sterilized distilled water. The resulting pellets were placed in a resuspension solution (10 mM MgCl₂, 10 mM MES-K, 100 µM acetosyringone, pH 5.6). The Agrobacterium solution was adjusted to an A₆₀₀ of approximately 0.4 and placed on a bench at room temperature for 2–3 h before infiltration. The Agrobacterium solution was infiltrated into the third and the fourth leaves (from the shoot tip) of approximately four-week old *N. benthamiana*. After five days post infiltration, the infiltrated leaves were rapidly frozen in liquid nitrogen and ground into a powder. The powder was immediately lyophilized and vacuum stored at room temperature. A 30 mg sample of the dried powder was dissolved in 500 µL of 85% methanol and metabolically extracted by vigorous shaking using a Mixer Mill MM400 (Retsch, Germany) at 25 Hertz for 7 min. The crude extracts were filtered through an Acrodisc[®] Syringe Filter with 0.2 µm Supor[®] membrane (Pall, USA).

Abbreviations

4CL: 4-coumarate CoA ligase

C: tuberous cortices

C4H: cinnamate 4-hydroxylase

CGT: C-glycosyltransferase

CHI: chalcone isomerase

CHR: chalcone reductase

CHS: chalcone synthase

COG: clusters of orthologous groups

CYP81E: cytochrome P450 subfamily 81E

DEG: differential expressed gene

DMAPP: dimethylallyl pyrophosphate

EF1α: elongation factor 1 alpha

GFP: green fluorescent protein

GO: gene ontology

HID: 2-hydroxyisoflavanone dehydratase

HPLC-QTOF-MS/MS: high performance liquid chromatography quadrupole time of flight tandem mass spectrometry

IFR: isoflavone reductase

IFS: isoflavone synthase

KEGG: kyoto encyclopedia of genes and genomes

M: mature leaves

MYB TF: v-myb myeloblastosis viral oncogene homolog transcription factor

nr: NCBI non-redundant protein database

PAL: phenylalanine ammonia lyase

PT: prenyltransferase

RPKM: Read Per Kilobase of Transcript Per Million Mapped Reads

SRA: sequence read archive

T: cortex-excised tubers

Y: young leaves

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Our gratitude is extended to Dr. Tomoyoshi Akashi, Nihon University, for providing standard compounds. HS is thankful to Frédéric Tournay at the Botanical Garden of the University of Strasbourg for horticultural work and taking care of the plants.

Funding

The authors express their gratitude to Chulalongkorn University for providing research funds (CU–56–508-HR and GRU 6101133003–1). N. S. is the recipient of a student fellowship from the 100th Anniversary of Chulalongkorn University Fund for Doctoral Scholarship, the Overseas Research Experience Scholarship for Graduate Students from the Graduate School, Chulalongkorn University, and the FY2018 Thesis Grant for Ph.D. Student, National Research Council of Thailand. The research was also partly supported by the Research Unit for Natural Product Biotechnology, Chulalongkorn University, the Chulalongkorn Academic Advancement into Its 2nd Century Project, and the Franco-Thai Cooperation Program in Higher Education and Research together with the french PHC programme SIAM2014–2015. HS and WD are thankful to the French Ministry of Foreign Affairs, Campus France, and the French embassy in Bangkok for supporting the *BIOACTIVE NATURAL PRODUCTS FROM PLANTS* project (project n°31827TJ). The funding bodies were not involved in the design of the study and collection, data analysis, data interpretation, and the writing of the manuscript.

Author Contributions

NS designed and performed the experiments, including analyzing the data and writing the manuscript. KT, PP, JZ, HS, KK, and SS analyzed the data. SS, SV, HS, MY, KS, and WD conceived and designed the study and revised the manuscript. SW provided *P. mirifica* materials. All authors read and approved the manuscript.

References

- 1.Malaivijitnond S: *Medical applications of phytoestrogens from the Thai herb Pueraria mirifica*. *Frontiers in Medicine* 2012, 6(1):8–21.
- 2.Ososki AL, Kennelly EJ: *Phytoestrogens: a review of the present state of research*. *Phytotherapy Research* 2003, 17(8):845–869.
- 3.Ingham JL, Tahara S, Dziedzic SZ: *A chemical investigation of Pueraria mirifica roots*. *Zeitschrift für Naturforschung C* 1986, 41(4):403–408.

4. Cherdshewasart W, Subtang S, Dahlan W: *Major isoflavonoid contents of the phytoestrogen rich-herb Pueraria mirifica in comparison with Pueraria lobata. Journal of Pharmaceutical and Biomedical Analysis* 2007, 43(2):428–434.
5. Cherdshewasart W, Sriwatcharakul S: *Major isoflavonoid contents of the 1-year-cultivated phytoestrogen-rich herb, Pueraria mirifica. Bioscience, Biotechnology, and Biochemistry* 2007, 71(10):2527–2533.
6. Veitch NC: *Isoflavonoids of the Leguminosae. Natural Product Reports* 2013, 30(7):988–1027.
7. Ingham JL, Tahara S, Dziedzic SZ: *Coumestans from the roots of Pueraria mirifica. Zeitschrift für Naturforschung C* 1988, 43(1–2):5–10.
8. Costa M, Dias TA, Brito A, Proença F: *Biological importance of structurally diversified chromenes. European Journal of Medicinal Chemistry* 2016, 123:487–507.
9. Jones HE, Pope GS: *A method for the isolation of miroestrol from Pueraria mirifica. Journal of Endocrinology* 1961, 22(3):303–312.
10. Yusakul G, Putalun W, Udomsin O, Juengwatanatrakul T, Chaichantipyuth C: *Comparative analysis of the chemical constituents of two varieties of Pueraria candollei. Fitoterapia* 2011, 82(2):203–207.
11. Cain JC: *Miroestrol: an oestrogen from the plant Pueraria mirifica. Nature* 1960, 188:774–777.
12. Dixon RA, Achnine L, Kota P, Liu C-J, Reddy MSS, Wang L: *The phenylpropanoid pathway and plant defence—a genomics perspective. Molecular Plant Pathology* 2002, 3(5):371–390.
13. Wiriyaampaiwong P, Thanonkeo S, Thanonkeo P: *Molecular characterization of isoflavone synthase gene from Pueraria candollei var. mirifica. African Journal of Agricultural Research* 2012, 7(32):4489–4498.
14. Wiriyaampaiwong P, Thanonkeo S, Thanonkeo P: *Cloning and characterization of chalcone synthase gene from Pueraria candollei var. mirifica. Journal of Medicinal Plants Research* 2012, 6(42):5469–5479.
15. Udomsuk L, Juengwattanatrakul T, Jarukamjorn K, Putalun W: *Increased miroestrol, deoxymiroestrol and isoflavonoid accumulation in callus and cell suspension cultures of Pueraria candollei var. mirifica. Acta Physiologiae Plantarum* 2011, 34(3):1093–1100.
16. Liu J, Osbourn A, Ma P: *MYB transcription factors as regulators of phenylpropanoid metabolism in plants. Molecular Plant* 2015, 8(5):689–708.
17. Han R, Takahashi H, Nakamura M, Yoshimoto N, Suzuki H, Shibata D, Yamazaki M, Saito K: *Transcriptomic landscape of Pueraria lobata demonstrates potential for phytochemical study. Frontiers in Plant Science* 2015, 6.

- 18.Wang X, Li S, Li J, Li C, Zhang Y: *De novo transcriptome sequencing in Pueraria lobata to identify putative genes involved in isoflavones biosynthesis. Plant Cell Reports* 2015, 34(5):733–743.
- 19.Akashi T, Aoki T, Ayabe S-i: *CYP81E1, a cytochrome P450 cDNA of licorice (Glycyrrhiza echinata L.), encodes isoflavone 2'-hydroxylase. Biochemical and Biophysical Research Communications* 1998, 251(1):67–70.
- 20.Liu CJ, Huhman D, Sumner LW, Dixon RA: *Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from Medicago truncatula. Plant Journal* 2003, 36(4):471–484.
- 21.Overkamp S, Hein F, Barz W: *Cloning and characterization of eight cytochrome P450 cDNAs from chickpea (Cicer arietinum L.) cell suspension cultures. Plant Science* 2000, 155(1):101–108.
- 22.Shimada N, Akashi T, Aoki T, Ayabe S-i: *Induction of isoflavonoid pathway in the model legume Lotus japonicus: molecular characterization of enzymes involved in phytoalexin biosynthesis. Plant Science* 2000, 160(1):37–47.
- 23.Lee JH, Kim JY, Cho S-H, Jeong JH, Cho S, Park HJ, Baek SY: *Determination of miroestrol and isomiroestrol from Pueraria mirifica (White Kwao Krua) in dietary supplements by LC–MS-MS and LC–Q-Orbitrap/MS. Journal of Chromatographic Science* 2017, 55(3):214–221.
- 24.Sainsbury F, Lomonossoff GP: *Transient expressions of synthetic biology in plants. Current Opinion in Plant Biology* 2014, 19:1–7.
- 25.Pandey A, Misra P, Khan Mohd P, Swarnkar G, Tewari Mahesh C, Bhambhani S, Trivedi R, Chattopadhyay N, Trivedi Prabodh K: *Co-expression of Arabidopsis transcription factor, AtMYB12, and soybean isoflavone synthase, GmIFS1, genes in tobacco leads to enhanced biosynthesis of isoflavones and flavonols resulting in osteoprotective activity. Plant Biotechnology Journal* 2013, 12(1):69–80.
- 26.Chu S, Wang J, Zhu Y, Liu S, Zhou X, Zhang H, Wang C-e, Yang W, Tian Z, Cheng H et al: *An R2R3-type MYB transcription factor, GmMYB29, regulates isoflavone biosynthesis in soybean. PLOS Genetics* 2017, 13(5):e1006770.
- 27.Yi J, Derynck MR, Chen L, Dhaubhadel S: *Differential expression of CHS7 and CHS8 genes in soybean. Planta* 2010, 231(3):741–753.
- 28.Li X-W, Li J-W, Zhai Y, Zhao Y, Zhao X, Zhang H-J, Su L-T, Wang Y, Wang Q-Y: *A R2R3-MYB transcription factor, GmMYB12B2, affects the expression levels of flavonoid biosynthesis genes encoding key enzymes in transgenic Arabidopsis plants. Gene* 2013, 532(1):72–79.
- 29.Yi J, Derynck MR, Li X, Telmer P, Marsolais F, Dhaubhadel S: *A single-repeat MYB transcription factor, GmMYB176, regulates CHS8 gene expression and affects isoflavonoid biosynthesis in soybean. Plant Journal* 2010, 62(6):1019–1034.

30. Yu O, Shi J, Hession AO, Maxwell CA, McGonigle B, Odell JT: *Metabolic engineering to increase isoflavone biosynthesis in soybean seed. Phytochemistry* 2003, 63(7):753–763.
31. Reuter Jason A, Spacek DV, Snyder Michael P: *High-throughput sequencing technologies. Molecular Cell* 2015, 58(4):586–597.
32. Serres-Giardi L, Belkhir K, David J, Glémin S: *Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell* 2012, 24(4):1379–1397.
33. Tian A-G, Wang J, Cui P, Han Y-J, Xu H, Cong L-J, Huang X-G, Wang X-L, Jiao Y-Z, Wang B-J *et al*: *Characterization of soybean genomic features by analysis of its expressed sequence tags. Theoretical and Applied Genetics* 2004, 108(5):903–913.
34. Chansakaow S, Ishikawa T, Sekine K, Okada M, Higuchi Y, Kudo M, Chaichantipyuth C: *Isoflavonoids from Pueraria mirifica and their estrogenic activity. Planta Medica* 2000, 66(06):572–575.
35. Udomsin O, Juengwatanatrakul T, Yusakul G, Putalun W: *Chromene stability: The most potent estrogenic compounds in White Kwao Krua (Pueraria candollei var mirifica) crude extract. Journal of Functional Foods* 2015, 19:269–277.
36. Mameda R, Waki T, Kawai Y, Takahashi S, Nakayama T: *Involvement of chalcone reductase in the soybean isoflavone metabolon: identification of GmCHR5, which interacts with 2-hydroxyisoflavanone synthase. Plant Journal* 2018, 96(1):56–74.
37. Laursen T, Møller BL, Bassard J-E: *Plasticity of specialized metabolism as mediated by dynamic metabolons. Trends in Plant Science* 2015, 20(1):20–32.
38. Bang M-H, Lee D-G, Baek Y-S, Cho J-G, Han M-w, Choi K-S, Chung D-K, Ko S-k, Oh C-H, Cho S-Y *et al*: *A new miroestrol glycoside from the roots of Pueraria mirifica. Chemistry of Natural Compounds* 2013, 49:443.
39. Yazaki K, Sugiyama A, Morita M, Shitan N: *Secondary transport as an efficient membrane transport mechanism for plant secondary metabolites. Phytochemistry Reviews* 2008, 7(3):513–524.
40. Wang X, Li C, Zhou C, Li J, Zhang Y: *Molecular characterization of the C-glucosylation for puerarin biosynthesis in Pueraria lobata. Plant Journal* 2017, 90(3):535–546.
41. Nielsen KA, Tattersall DB, Jones PR, Møller BL: *Metabolon formation in dhurrin biosynthesis. Phytochemistry* 2008, 69(1):88–98.
42. Jørgensen K, Rasmussen AV, Morant M, Nielsen AH, Bjarnholt N, Zagrobelny M, Bak S, Møller BL: *Metabolon formation and metabolic channeling in the biosynthesis of plant natural products. Current Opinion in Plant Biology* 2005, 8(3):280–291.

- 43.Laursen T, Borch J, Knudsen C, Bavishi K, Torta F, Martens HJ, Silvestro D, Hatzakis NS, Wenk MR, Dafforn TR *et al*: *Characterization of a dynamic metabolon producing the defense compound dhurrin in sorghum*. *Science* 2016, *354*(6314):890.
- 44.Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M *et al*: *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis*. *Nature Protocols* 2013, *8*(8):1494–1512.
- 45.Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A: *Differential expression in RNA-seq: A matter of depth*. *Genome Research* 2011, *21*(12):2213–2223.
- 46.Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: *High-throughput functional annotation and data mining with the Blast2GO suite*. *Nucleic acids research* 2008, *36*(10):3420–3435.
- 47.Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L *et al*: *WEGO: a web tool for plotting GO annotations*. *Nucleic Acids Research* 2006, *34*:W293-W297.
- 48.Kumar S, Stecher G, Tamura K: *MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets*. *Molecular Biology and Evolution* 2016, *33*(7):1870–1874.
- 49.Livak KJ, Schmittgen TD: *Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method*. *Methods* 2001, *25*(4):402–408.
- 50.Sainsbury F, Thuenemann EC, Lomonossoff GP: *pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants*. *Plant Biotechnology Journal* 2009, *7*(7):682–693.

Figures

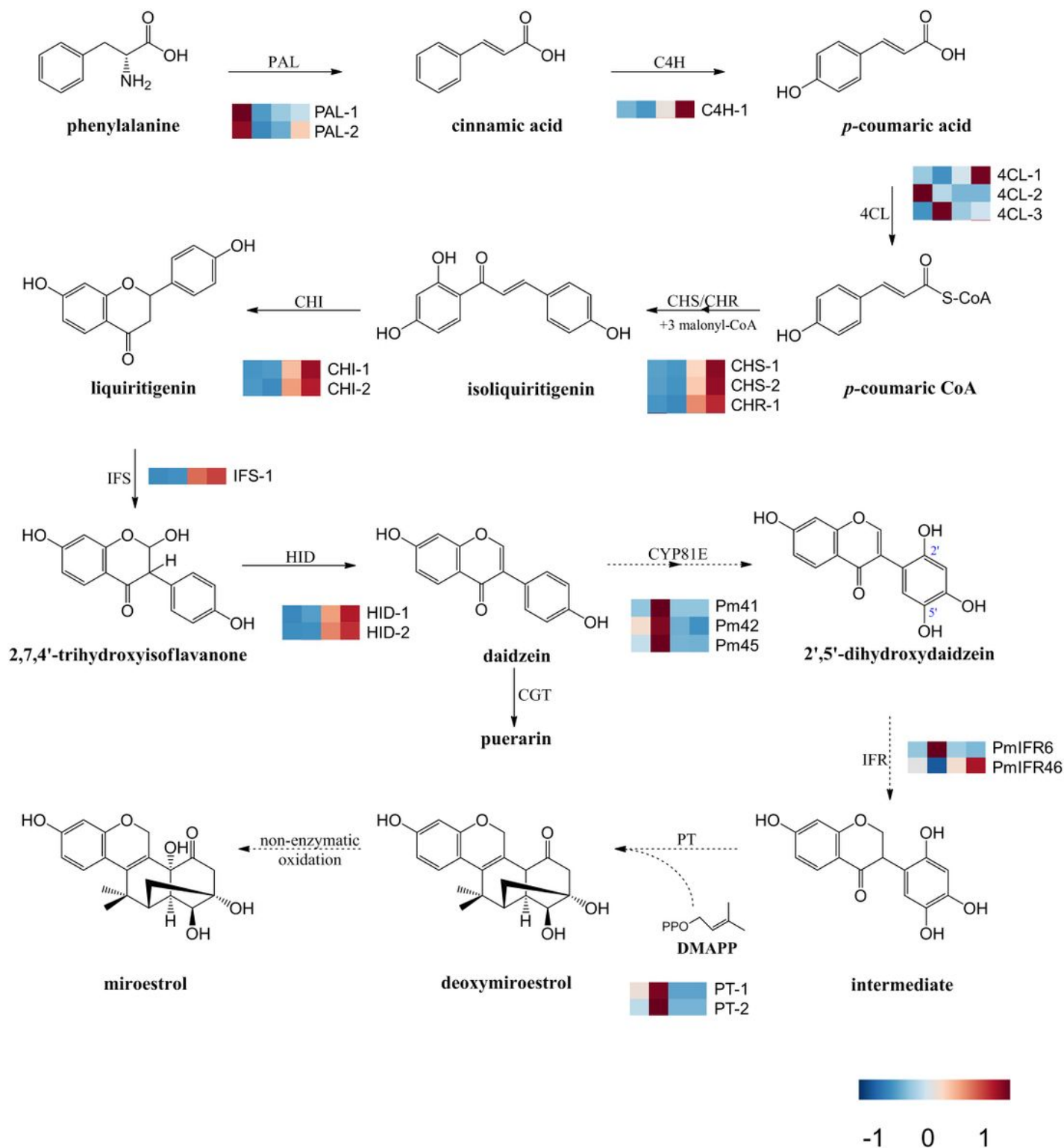


Figure 1

Proposed miroestrol biosynthetic pathway in *P. mirifica*. The heatmap following each gene name indicates the relative differential expression of genes in young leaves, mature leaves, tubers without cortices, and cortices of *P. mirifica*, respectively. Enzyme abbreviation: PAL; Phenylalanine ammonia-lyase, C4H; Cinnamate 4-monooxygenase, 4CL; Coumarate-CoA ligase, CHS; Chalcone synthase, CHI; Chalcone isomerase, IFS; 2-hydroxyisoflavone synthase, HID; 2-hydroxyisoflavone dehydratase, CGT; C-

glycosyltransferase, CYP81E; Cytochrome P450 subfamily 81E, IFR; Isoflavone reductase, PT; Prenyltransferase. Although the assembled unigenes are hypothetical candidate genes, a comparative transcriptome analysis using *P. lobata* is a promising way to study and identify species-specific and evolutionary conserved pathways involved in the highly complex biosynthesis of miroestrol.

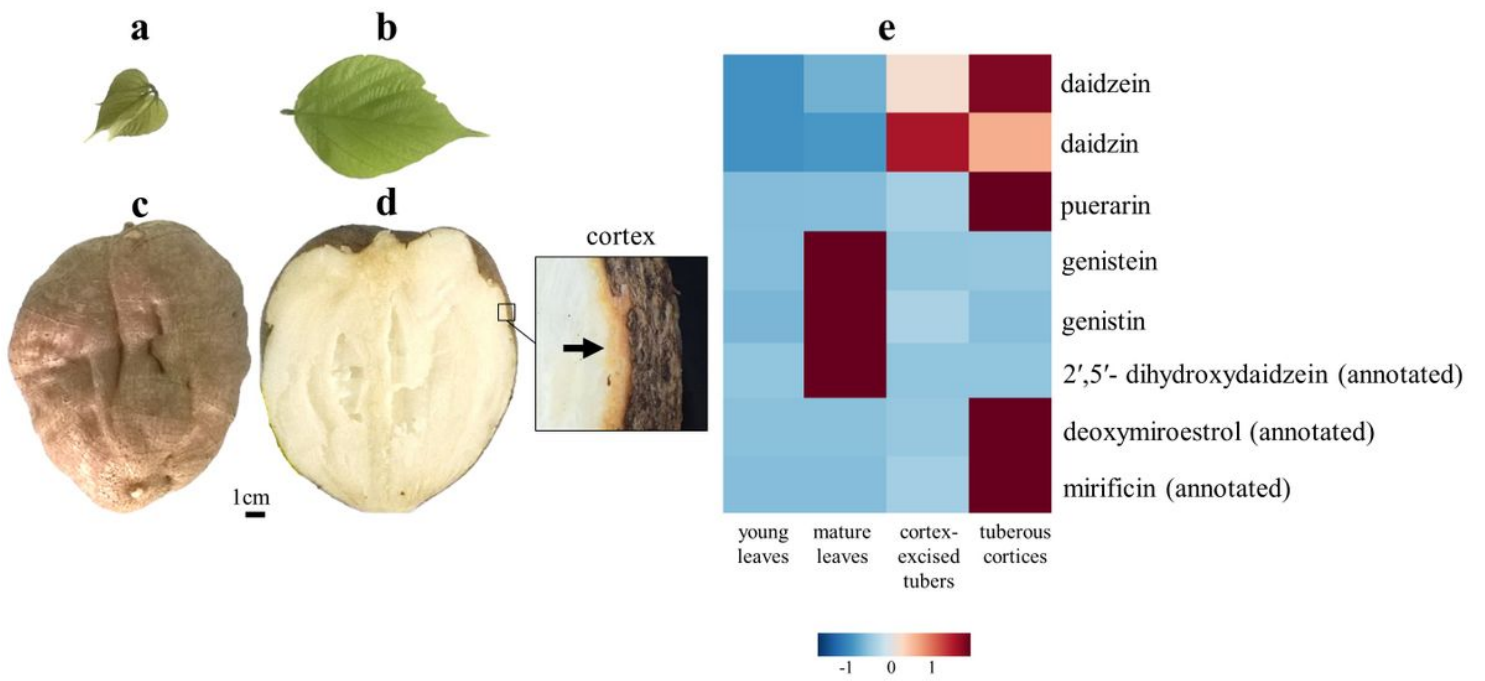


Figure 2

Young leaves (a), mature leaves (b) and tuber (c) morphology of *P. mirifica* cultivar SARDI19 used for transcriptome assembly, RNA-Seq, qRT-PCR, and HPLC-QTOF-MS/MS analysis (left). Cross section of three year old whole tuber (d). Accumulation of isoflavonoids and chromenes in four different tissues of *P. mirifica* as measured with HPLC-QTOF-MS/MS (e). The analysis was performed with three biological replicates.

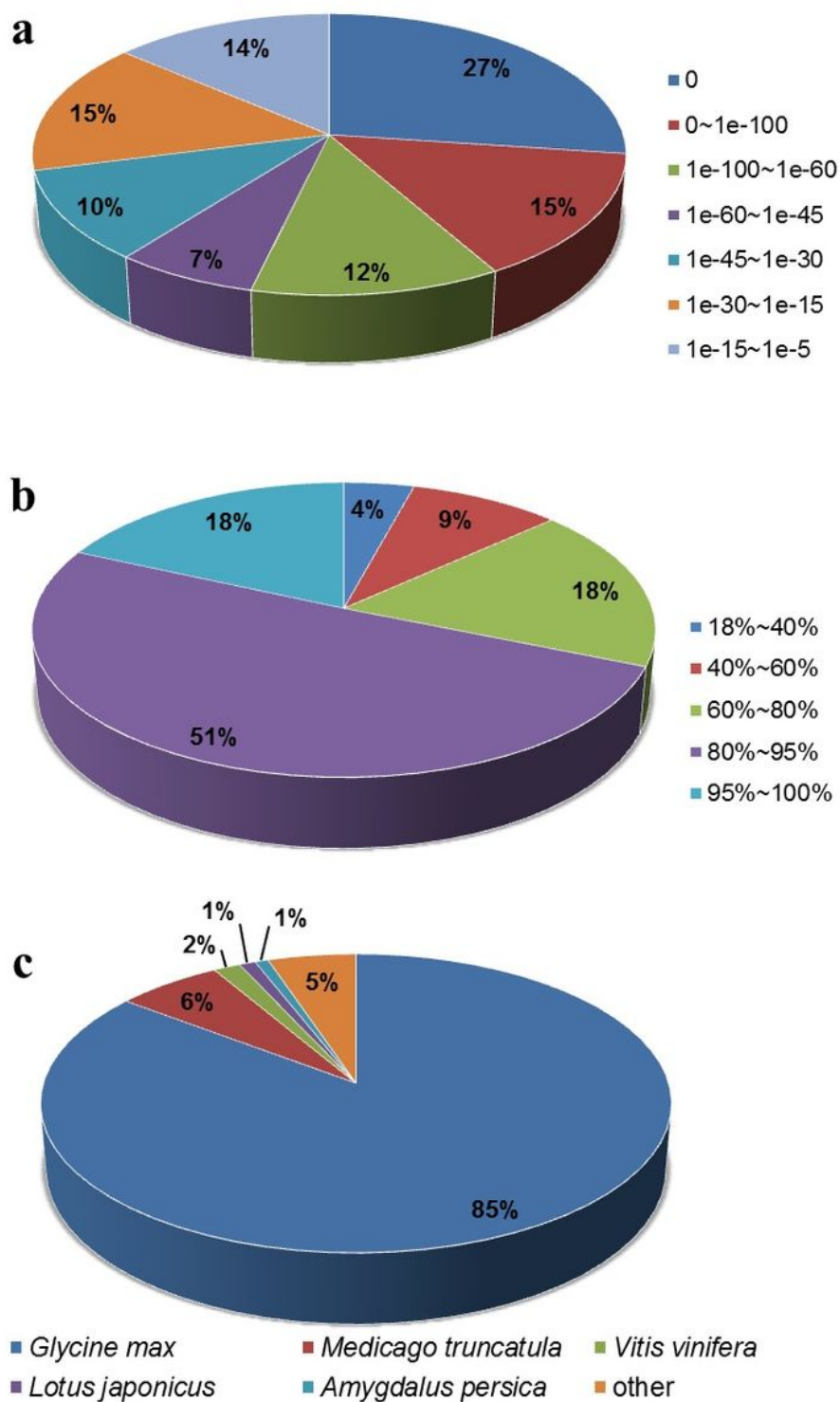


Figure 3

Characterization of *P. mirifica* unigenes based on NCBI non-redundant (nr) protein database search. (a) E-value distribution of BLAST hits for the *P. mirifica* assembled transcriptome with a cutoff of E-value < 10⁻⁵. b Similarity score distribution of top BLAST hits for the assembled unigenes. c The top five species distribution of BLAST hits for the assembled unigenes.

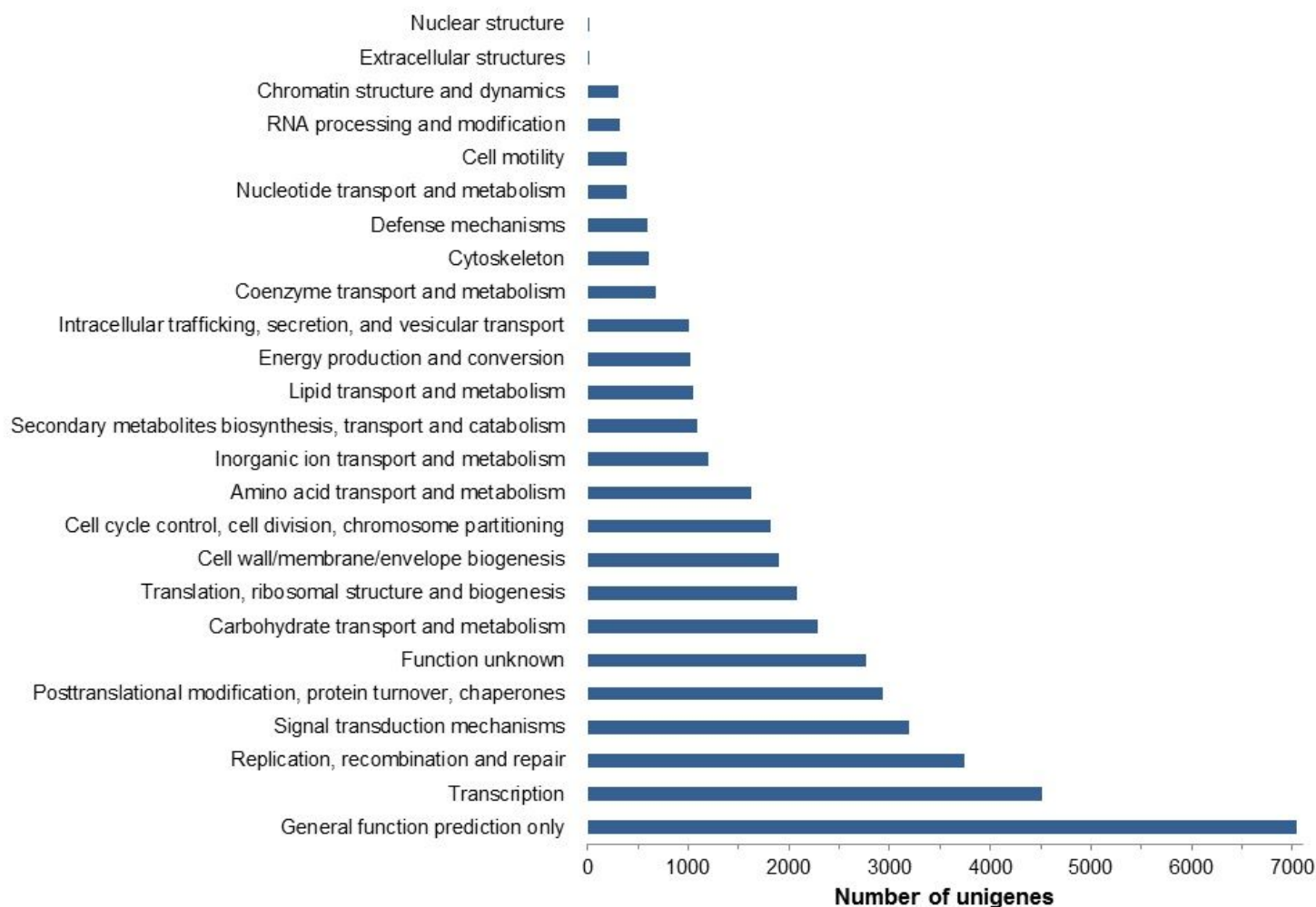


Figure 4

Clusters of orthologous groups (COG) functional classification for all assembled unigenes in *P. mirifica*. The vertical coordinates are function classes of COG, and the horizontal coordinates are numbers of unigenes.

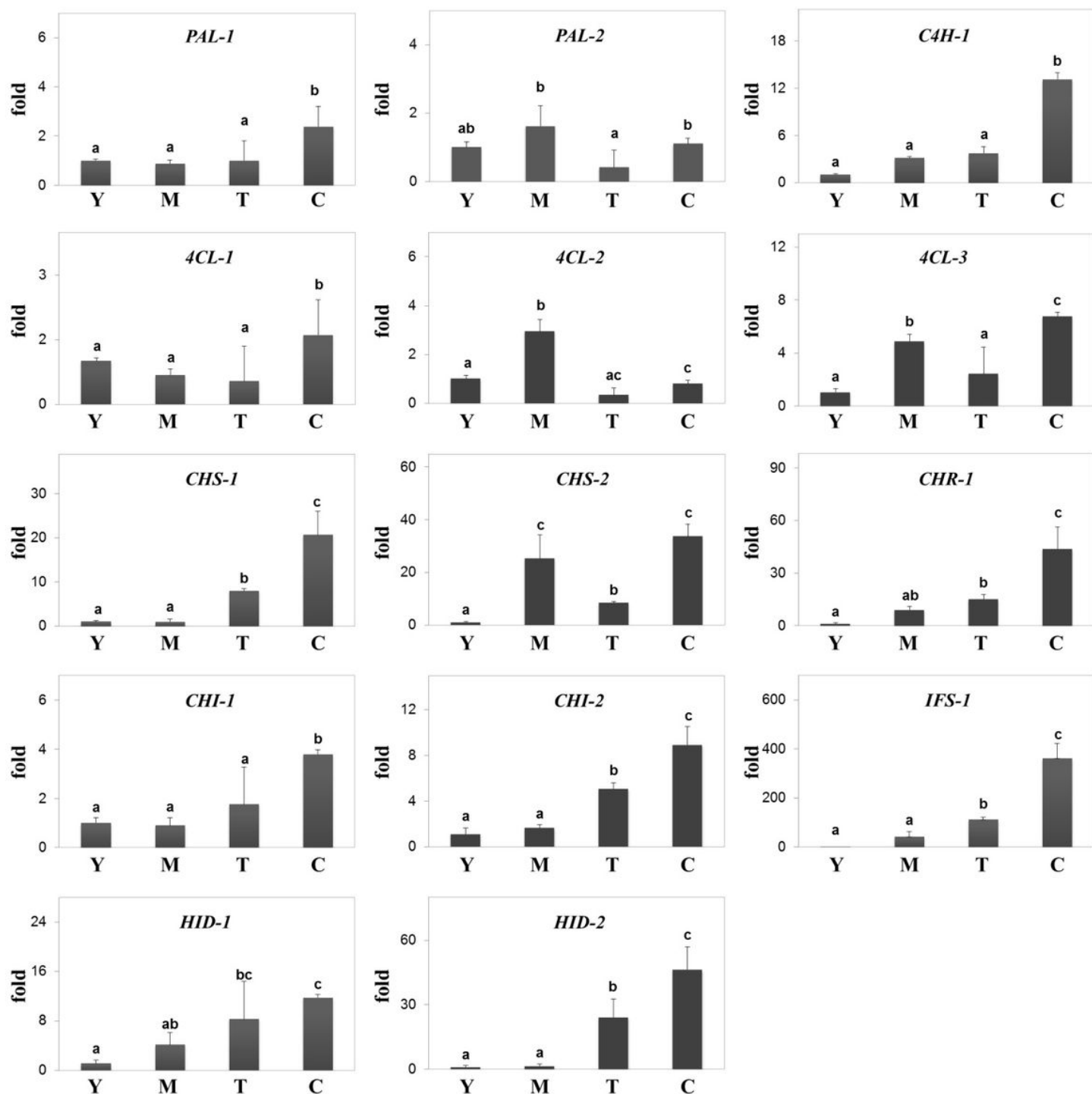


Figure 5

Relative expression of the candidate unigenes involved in isoflavone biosynthesis across four tissues of *P. mirifica* by qRT-PCR. Tissue abbreviations: Y; young leaves, M; mature leaves, T; cortex-excised tubers, C; tuberous cortices. Gene abbreviations: PAL; Phenylalanine ammonia-lyase, C4H; Cinnamate 4-monooxygenase, 4CL; Coumarate-CoA ligase, CHS; Chalcone synthase, CHI; Chalcone isomerase, IFS; 2-hydroxyisoflavone synthase, HID. Significantly different expression levels of putative genes (ANOVA, post-hoc Duncan test, $P < 0.05$) have different lowercase letters.

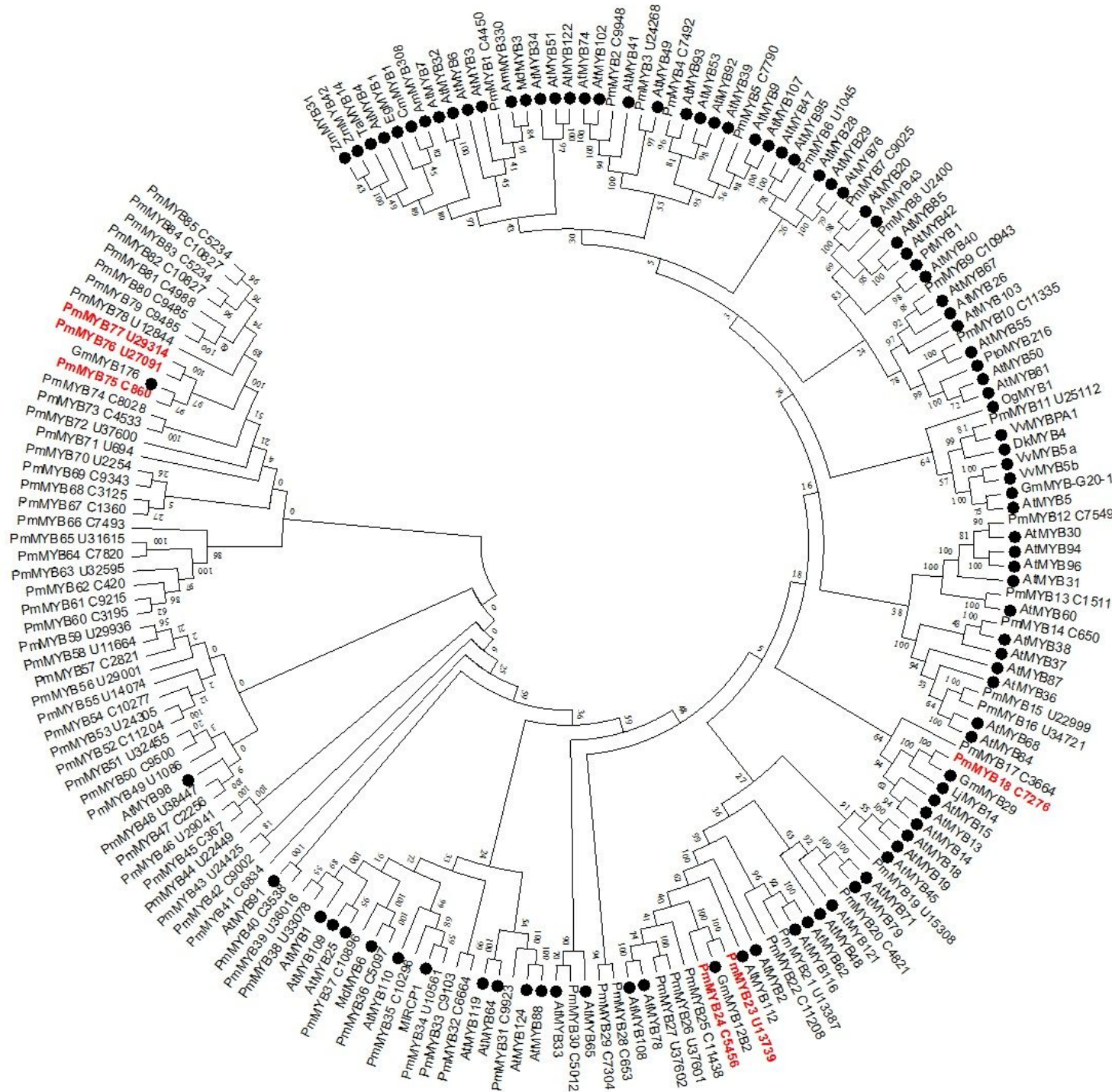


Figure 6

Phylogeny of MYB transcription factors associated with the phenylpropanoid pathway. The tree was constructed using the neighbor-joining method with putative amino acid full-length MYB sequences with MEGA7. Bootstrap values are shown as percentage (1000 replicates) when greater than 50%.

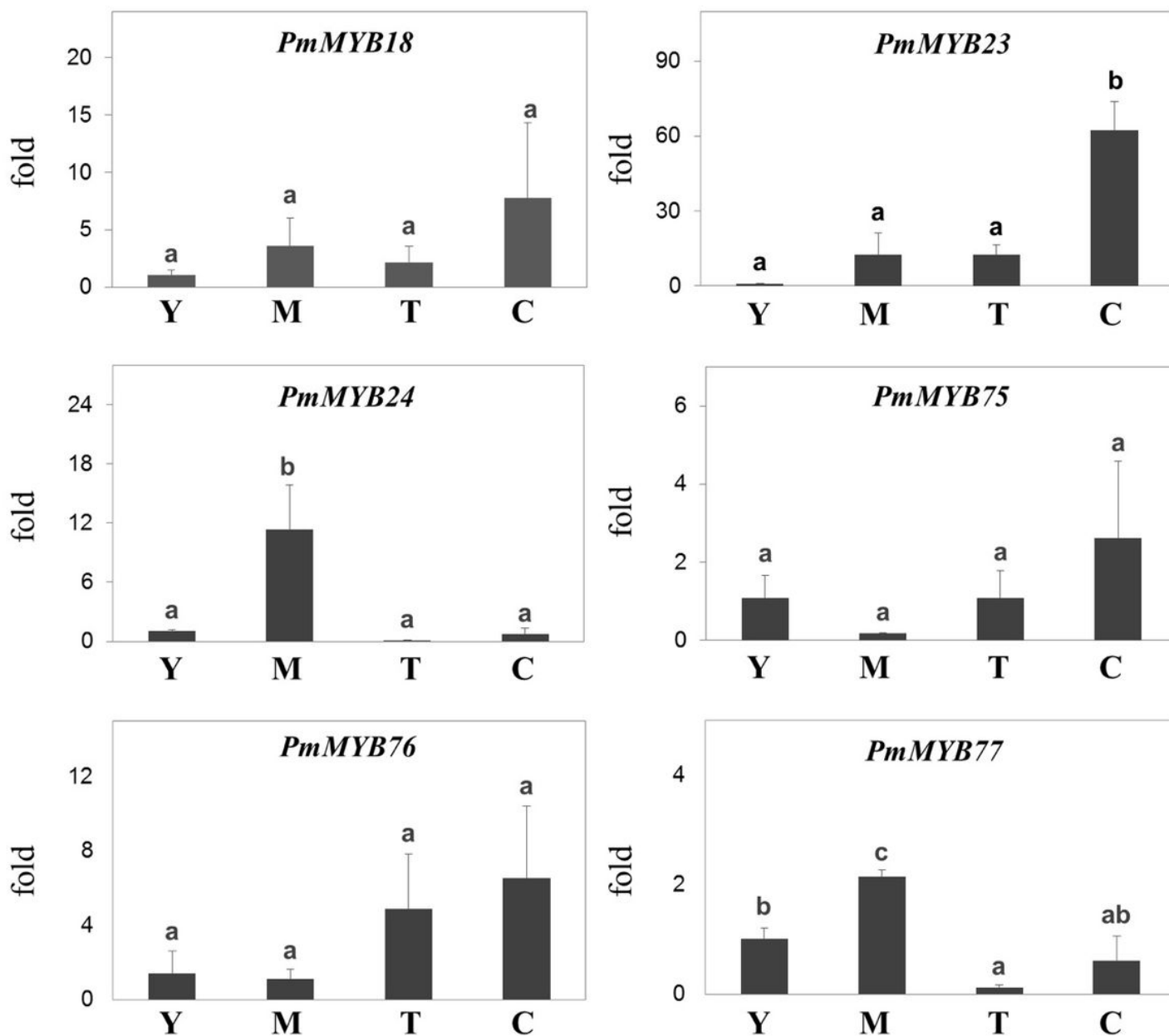


Figure 7

Relative expression of the candidate MYB TFs involved in isoflavone biosynthesis across four tissues of *P. mirifica* by qRT-PCR. Tissue abbreviations: Y; young leaves, M; mature leaves, T; cortex-excised tubers, C; tuberous cortices. Significantly different expression levels of putative genes (ANOVA, post-hoc Duncan test, $P < 0.05$) have different lowercase letters.

Supplementary Files

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

- [supplement1.docx](#)

- [supplement2.docx](#)
- [supplement3.docx](#)
- [supplement4.jpg](#)
- [supplement5.jpg](#)
- [supplement6.jpg](#)
- [supplement7.jpg](#)