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: Pueraria candollei var. mirifica, White Kwao Krua, miroestrol, isoflavones, transcriptome

Posted Date: October 8th, 2019

DOI: https://doi.org/10.21203/rs.2.12015/v2

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. candollei var. 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.A total of 166,923 contigs was assembled for functional annotation using protein databases and as a library for identification of genes that are potentially involved in the biosynthesis of isoflavonoids and miroestrol. Twenty-one differentially expressed genes from four separate libraries were identified as candidatesinvolved in these biosyntheticpathways, and their respective expressions were validated by quantitative real-time reverse transcription polymerase chain reaction. Notably, isoflavonoid profiling generated by LC-MS/MS was positively correlated with expression levels of isoflavonoid biosynthetic genes across the four types of tissues. Moreover, we identified R2R3 MYB transcription factors that may be involved in the regulation of isoflavonoid biosynthesis in P. candollei var. mirifica . To confirm the function of a key-isoflavone biosynthetic gene, P. candollei var. mirifica isoflavone synthase identified in our library was transiently co-expressed with an Arabidopsis MYB12 transcription factor (At MYB12) in Nicotiana benthamiana leaves. Remarkably, the combined expression of these proteins ledto the production of the 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*, shown in Supplementary Fig. S1), 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 they have been applied to treat different forms of cancer, heart disease, menopausal symptoms, and osteoporosis [2]. Considering the numerous effects of phytoestrogens on human health, *P. mirifica* may be a promising candidate for treating various diseases and for developing novel medicinal products.

Three major types of phytoestrogens occur in *P. mirifica*: isoflavones, coumestans, and chromenes [3-5]. Isoflavones are an important type of phytoestrogens, which are biosynthesized via the phenylpropanoid pathway and occur predominantly in leguminous plants [6]. Seven isoflavones that have been identified in *P. mirifica* tubers: puerarin, daidzin, genistin, daidzein, genistein, kwakhurin, and mirificin [3, 5]. Four coumestans that also occur in the tuber of this plant are coumestrol, mirificoumestan, mirificoumestan hydrate, and mirificoumestan [7]. The chemical structure of chromenes has received considerable attention because of their low toxicity and broad pharmacological application as anticancer, antimicrobial, and anti-inflammatory agents [8]. Miroestrol and its precursor deoxymiroestrol are typically

accumulated at very low levels [9], however, these compounds are the predominant chromenes in the tuberous cortex of *P. mirifica* [10], and both compounds exhibit considerably highest estrogenic activity [11]. Since both chromenes have not been reported in other plant species, their biosynthesis is likely unique to *P. mirifica*. Interestingly, although miroestrol has been identified almost six decades ago, its biosynthetic pathway and the enzymes involved are still unknown.

Regarding biosynthesis of phytoestrogens, the early steps of isoflavonoid and flavonoid are generally similar, starting with phenylalanine ammonia lyase removing amides from the first substrate, phenylalanine, to produce cinnamic acid that is hydroxylated by cinnamate 4-hydroxylase to produce *p*-coumarate. The enzyme 4-coumarate-CoA ligase then activates *p*-coumarate by attaching a co-enzyme A (coA), and subsequently, chalcone synthase (CHS) binds *p*-coumaroyl-CoA to three molecules of malonyl-CoA to form a chalcone skeleton. Chalcone can be converted to flavanone by chalcone isomerase. Liquiritigenin is a substrate for both the flavonoid and the 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 formation of an intermediate product, 2-hydroxyisoflavanone, which is then dehydrated to daidzein through catalysis by 2-hydroxyisoflavanone dehydratase (Fig. 1) [12]. CHS and IFS have been identified in *P. mirifica* several years ago [13, 14], however, their enzymatic functions have not been elucidated. Udomsuk et al. [15] suggested that miroestrol biosynthesis may share a common pathway with isoflavonoid biosynthesis due to their similar backbone structure. So far, no biosynthetic genes or enzymes involved in miroestrol biosynthesis have been reported, thus also the transcription factors responsible for regulating the expression of these biosynthetic genes remain to be identified. V-myb myeloblastosis viral oncogene homolog transcription factors (MYB TFs), which are the largest plant transcription factor family, have been reported to possess key functions in regulating the synthesis of phenylpropanoid-derived compounds in plants [16]. These proteins have attracted substantial interest regarding phytoestrogen biosynthesis in plants.

Transcriptomes produced from high-throughput sequencing of various plants are a potential source for identifying genes involved in the biosynthesis of different secondary metabolites. The transcriptome of *Pueraria lobata*, a species closely related to *P. mirifica,* has been published previously [17, 18]. Although these two plants produce similar types of isoflavones, miroestrol occurs only *in P. mirifica*. In the afore-mentioned studies, *P. lobata* genes that encode core isoflavone biosynthetic enzymes were identified, and their expression levels in various tissues were examined.

In the current study, we high-throughput sequenced *P. mirifica* to produce transcriptome libraries of young leaves, mature leaves, cortex-excised tubers, and tuberous cortices, and we *de novo* assembled the transcriptomes to characterize the biosynthetic pathway of miroestrol. Additionally, MYB TFs involved in isoflavonoid biosynthesis were also identified. This integrative approach of using transcriptomics and metabolomics provides new insights for the prediction and identification of putative biosynthetic genes and transcription factors that are potentially 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 an Illumina Hiseq 2000 platform, yielding approximately 8.2 giga base pairs and a total of 7,386,137,640 clean nucleotides (nt). The *de novo* assembly resulted in 166,923 contigs (62,567,517 nt) and 104,283 unigenes (81,810,584 nt), the mean length/N50 for contigs and unigenes was 375/734 nt and 785/1558 nt, respectively. To assess the completeness of transcriptome data, we performed a BUSCO analysis compared to the 2,121 single-copy orthologs of the eudicot lineage. The *de novo* transcriptome assembly was complete to 87.4%, 6.9% of contigs were fragmented, and 5.7% of the transcriptome was missing (Supplementary Table S1). In addition, the assembly produced here was compared to that of other leguminous plants (summarized in Table 1).

Table 1 Comparison of *P. mirifica de novo* assembly to other leguminous plants.

Description	*P. mirifica*	*P. lobata* [18]	*Ammopiptanthus mongolicus* [19]	*Millettia pinnata* [20]	*Medicago sativa* [21]
Total Clean Reads	82068196	73360286	67287120	80212402	28790610
Total Clean Nucleotides (nt)	7386137640	6973955470	6055840800	7219116180	5642959560
Contig					
Total Number	166923	335582	148797	108731	81277
Total Length (nt)	62567517	390188024	51308749	39658128	70966536
Mean Length (nt)	375	1162	345	365	873
N50 (nt)	734	1988	619	682	1323
Unigene					
Total Number	104283	163625	84583	53586	40433
Total Length (nt)	81810584	111776902	57108594	42186440	32482946
Mean Length (nt)	785	683	675	787	803
N50 (nt)	1558	1153	1191	1204	1300

Functional annotation and classification of protein

Functional annotation of the assembled transcripts provides insights in potential metabolic functions and biological processes within an organism. The functional annotation of *P. mirifica* assembled transcripts was performed based on similarities with proteins or transcripts according to information that is available in various public databases. The statistics of functional annotation are summarized in Supplementary Table S2. Aligned unigenes showed significant homologies using the National Center for Biotechnology Information (NCBI) non-redundant protein database. Based on the BLAST similarity distribution, over 68% of unigenes exhibited similarities greater than 80%. A top-hit species distribution analysis showed unigenes with BLAST hits sharing high sequence similarities with *Glycine max* (85.3%), *Medicago truncatula* (6.2%), and *Vitis vinifera* (1.6%).

The clusters of orthologous groups (COG) indicated the functional classification of each unigene at the cellular level. In total, 1,091 unigenes were predicted to be involved in secondary metabolite biosynthesis (Supplementary Fig. S2). Gene ontology (GO), which classifies standardized gene function, is useful for annotating gene functions and gene products in various organisms. GO is based on three major dependent factor 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 categories related to biological processes, 17 to cellular components, and 16 to molecular functions. Supplementary Figure S3 shows the substantial number of transcripts related to cellular components and metabolic processes. The remaining 67,225 unigenes produced no BLAST hits. These unannotated unigenes may be uncharacterized genes or assembled sequences that were too short to produce hits.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive database for identifying biological pathways and for functional annotation of gene products. Pathway-based annotation helps produce an overview of the different metabolic processes that are active within an organism, and it helps improve our understanding of biological functions of unigenes. All unigenes were analyzed using the KEGG pathway with an *e*-value cutoff of $< 10^{-5}$. We obtained 33,317 unigenes, 3,997 of which were related to biosynthesis of secondary metabolites and 8,103 to 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 crucial capacity of leguminous plants to accumulate functional flavonoids, 476 flavonoid biosynthetic unigenes and 167 isoflavonoid biosynthetic unigenes are shown in Supplementary Figure S4. These functional annotations were used for identifying genes involved in isoflavonoid biosynthesis in *P. mirifica*.

Proposed miroestrol biosynthetic pathway, differential accumulation of transcripts associated with isoflavonoids, and miroestrol biosynthesis

Miroestrol biosynthesis potentially shares a pathway with isoflavonoid biosynthesis [15]. We propose that daidzein, a common isoflavone aglycone, is hydroxylated by at least two cytochrome P450 enzymes at

the 2' and 3' carbon of the B-ring to produce 2',5'-dihydroxydaidzein; these enzyme may be members of the CYP81E subfamily which is known to use isoflavones as substrates [22-25]. Then, 2',5'-dihydroxydaidzein would be reduced by isoflavone reductase and subsequently would be prenylated by a prenyltransferase using dimethylallyl diphosphate as a co-substrate to produce deoxymiroestrol and miroestrol (Fig. 1).

Based on a functional annotation of each unigene found in the *P. mirifica de novo* transcriptome assembly and phylogenetic analyses, a total of 14 putative genes involved in isoflavone biosynthesis were predicted as follows: two *PAL* genes (encoding phenylalanine ammonia lyase), one *C4H* gene (encoding cinnamate 4-hydroxylase), three *4CL* genes (encoding 4-coumarate-CoA ligase), two *CHS* genes (encoding chalcone synthase), one *CHR* gene (encoding chalcone reductase), two *CHI* genes (encoding chalcone isomerase), one *IFS* gene (encoding IFS), and two *HID* genes (encoding 2-hydroxyisoflavanone dehydratase). In addition, a total of seven putative genes were identified in our proposed miroestrol biosynthetic pathway: three *CYP81E* genes, two *IFR* genes (encoding isoflavone reductase), and two *PT* genes (encoding prenyltransferase). To investigate gene expression levels of these unigenes across four tissues of *P. mirifica*, reads per kilobase million of transcripts were calculated from the RNA sequencing data generated using a NextSeq500 platform. An analysis of differentially expressed genes (DEG) on these unigenes (probability threshold q = 0.9) was visualized as a heat map based on our proposed miroestrol biosynthetic pathway (Fig. 1). Additionally, a total of 16 putative genes encoding UDP-glycosyltransferase that are potentially involved in this pathway were phylogenetically identified from the transcriptome data. Their DEGs across the four tissue types are shown as a heatmap in Supplementary Fig. S5.

To validate the DEG analysis obtained from RNA sequencing, candidate unigenes were selected and analyzed based on their expression levels in the four tissue types using quantitative real-time reverse transcription polymerase chain reaction (qRT-PCR). As expected, expression levels of the selected unigenes were positively correlated with DEGs identified by RNA sequencing (Fig. 3). Indeed, most of the isoflavone biosynthetic genes were highly expressed in tuberous cortices, compared to young leaves, mature leaves, and cortex-excised tubers; however, *Pm*CYP81Es, *Pm*IFRs, and *Pm*PT higher expressed in mature leaves than in the other tissues.

Phytoestrogens and annotated constituents in *P. mirifica*

High-performance liquid chromatography (HPLC) coupled to tandem mass spectroscopy (MS) is a powerful tool with high selectivity and sensitivity. Seven phytoestrogens were identified in *P. mirifica* by comparison with standard compounds based on their retention times and MS fragmentation patterns. These phytoestrogens 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 [26] or/and comparisons with published data [27]. The relative abundance of major phytoestrogens found in young leaves, mature leaves, cortex-excised tubers, and tuberous cortices of *P. mirifica* are shown as a heat map in Figure 2e. Genistein, genistin, and annotated 2', 5'-dihydroxydaidzein were highly accumulated in mature leaves, whereas daidzein, puerarin, annotated mirificin, and annotated deoxymiroestrol were highly accumulated in tuberous cortices, compared to the other tissues. Although miroestrol was not detected in any tissue, its precursor (annotated deoxymiroestrol; RT = 15.03 min, *m/z* = 342.14672) was observed only in *P. mirifica* tuberous cortex, which is the main accumulation site of miroestrol [10].

Isoflavone production in *N. benthamiana* leaves over-expressing *P. mirifica* isoflavone synthase

Recently, *N. benthamiana* has been used to identify transient expression of several plant genes to confirm gene functions [28]. To demonstrate that our transcriptome libraries contained candidate genes involved in *P. mirifica* phytoestrogen biosynthesis, we cloned *P. mirifica* isoflavone synthase (*PmIFS*) and conducted a functional characterization using transient (co-)expression in *N. benthamiana*, which does not produce any isoflavones. Transient expression of green fluorescent protein in five-week-old *N. benthamiana* leaves was used as a negative control. Co-expression of *PmIFS* and *Arabidopsis* R2R3 MYB12 transcription factor (*AtMYB12*), a regulator enhancing metabolic flux to flavonoid biosynthesis [29], generated two novel major peaks that were identified as genistein (Supplementary Figure S6), suggesting that *PmIFS* is involved in isoflavone biosynthesis.

Identification of MYB transcription factors involved in isoflavone biosynthesis

We identified 85 putative genes encoding MYB transcription factors in the *P. mirifica* transcriptome. All putative *P. mirifica* MYB transcription factors (PmMYB) were aligned to known MYB transcription factors of other plants such as *Arabidopsis thaliana*, *Glycine max*, and other species, and a phylogenetic tree was produced from this alignment (Fig. 4). Based on these analyses, we found six candidate PmMYBs that are potentially involved in the regulation of isoflavonoid biosynthesis. PmMYB18 was closely related to GmMYB29, which activates expression of *IFS* and *CHS* in soybean [30]. PmMYB23 and PmMYB24 clustered with GmMYB12B2, which regulates expression of *CHS* in soybean [31, 32]. PmMYB75, PmMYB76, and PmMYB77 clustered with GmMYB176, which is also involved in controlling *CHS* expression and isoflavonoid synthesis in soybean [33, 34]. Furthermore, also *PmMYB24* and *PmMYB77* were highly expressed in mature leaves, whereas *PmMYB18*, *PmMYB23*, *PmMYB7*5, and *PmMYB76* were highly expressed in tuberous cortices (Fig. 5).

Discussion

De novo transcriptome assembly is a useful method for gathering comprehensive information on genetic resources without the need for whole genome sequencing. In addition, this technique facilitates discovery of novel genes, molecular markers, and tissue-specific expression patterns. In the absence of comprehensive genomic data of *P. mirifica*, RNA sequencing was used to explore the *P. mirifica*

transcriptome. Although the Illumina HiSeq platform has various advantages over other methods in large-scale transcriptomics research, the output data requires more time for processing and analyses. The Illumina NextSeq platform was thus designed as a faster and easier operating benchtop device to reduce costs and data processing time. Overall error rates (<1%) are comparable across these two platforms [35]. Here, we employed an Illumina HiSeq 2000 platform to produce a transcriptome library from four tissues, and an Illumina NextSeq 500 platform to determine expression patterns of genes in each transcriptome library produced from four tissues of *P. mirifica*. The BUSCO analysis confirmed that 94.3% of the complete and fragmented sequences were included in the assembly (Supplementary Table S1), suggesting that the vast majority of orthologs were covered by our *de novo* transcriptome assembly. We produced a large dataset of unigenes from the *de novo* transcriptome assembly generated on the HiSeq platform. The average GC content of those unigenes was 44.56%. In eukaryotes, GC content averages approximately 20-60% [36]. The observed *P. mirifica* GC content was within this range, although slightly above those in *P. lobata* (39.9%) [17] and *Medicago truncatula* (40%) [37] but similar to that in *Glycine max* (43%) [37]. Furthermore, most of the annotated unigenes produced matches in the *Glycine max* protein database, indicating that the respective functions are conserved in these two species. Although *P. lobata* is considered closely related to *P. mirifica*, only 0.12% of the annotated unigenes matched *P. mirifica*. This is due to the limited number of full-length sequences of *P. lobata* available in the protein database, and the currently available *P. lobata* NCBI reference transcriptome is raw sequencing data [17, 18]. In addition, gene or protein names, descriptions of COG and GO terms, and possible metabolic pathways were also annotated. Notably, all candidate genes generated from the HiSeq platform data were found in the four libraries generated from the NextSeq platform data. The respective expression levels across the four tissues of *P. mirifica* can be used to reduce the number of candidates for further functional characterization.

Isoflavones are a class of flavonoids 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 qRT-PCR, exhibited high expression levels in tuberous cortices compared to those in 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 isoflavone accumulation profile in each tissue were similar [17, 18], apart from the lack of *IFS* expression and isoflavonoid accumulation in the leaves of *P. lobata* [17]. This observation suggests a divergent evolution pattern in these two closely related species which occur in different regions (*P. lobata* in temperate zones and *P. mirifica* in tropical zones). It is possible that *P. mirifica* has evolved a mechanism to accumulate isoflavonoids in leaves and miroestrol in tuberous cortices as a means of defense. To assess whether our transcriptome data contained candidate genes, we cloned *Pm*IFS into *N. benthamiana*, which does not naturally produced isoflavonoids, for transient co-expression of *Pm*IFS and *At*MYB12, an activator of flavonoid biosynthesis [29]. This led to accumulation of genistein and genistin (Supplementary Fig.S6). The demonstrated *in planta* function of *Pm*IFS suggests that our transcriptome libraries were suitable for identifying other genes involved in miroestrol biosynthesis. Although in a previous study, a biosynthetic pathway of

miroestrol was tentatively suggested [15], the genes required for miroestrol biosynthesis remained unknown so far. Here, the biosynthetic pathway of miroestrol was reconsidered and the most plausible mechanism was proposed (Fig. 1). Seven putative genes encoding three biosynthetic enzymes involved in our proposed miroestrol biosynthetic pathway were also identified in our transcriptome data. Accumulation of deoxymiroestrol and miroestrol was reported to be higher in tuberous cortices than in cortex-excised tubers of *P. mirifica* [10]. Similarly, we detected deoxymiroestrol specifically in the tuberous cortices of *P. mirifica*; miroestrol, however, was not detected. Miroestrol can be non-enzymatically converted from deoxymiroestrol at high temperatures or in acidic or alkaline solutions during storage and extraction [38, 39]. In our experiment, *P. mirifica* tissues were harvested and rapidly frozen in liquid nitrogen, which would prevent conversion of deoxymiroestrol to miroestrol. In addition, the proposed intermediates 2'-hydroxydaidzein and 3'-hydroxydaidzein were not detected in any tissue. We hypothesized that an organized multienzyme cluster, known as metabolon, may be involved in miroestrol biosynthesis and be responsible for the absence of those cryptic biosynthetic intermediates. In fact, an increasing number of studies describe the organization of biosynthetic pathways as metabolons. Particularly, one recent study showed efficient and specific substrate transformation in a metabolon complex between IFS and chalcone reductase in soybean plants [40]. In addition, the formation of a metabolon channeling substrate towards a product without the leakage of labile intermediates has been shown [41].

Regarding putative genes involved in miroestrol biosynthesis in *P. mirifica*, we found that *CYP81E*, *IFR*, and *PT* genes were highly expressed in mature leaves, which were not the accumulation sites of miroestrol and deoxymiroestrol. Perhaps those intermediates or deoxymiroestrol are initially synthesized in mature leaves and are converted to a soluble form of glycoside that is readily transported across organs and is then stored predominantly in tuberous cortices. In fact, a glycosylated form of miroestrol has been identified in the *P. mirifica* tuber [42]. Examples of such transports are also known in tobacco, where nicotine is transported from the roots where it is biosynthesized site to the leaves where it is accumulated. Secondary metabolites are translocated between plant cells through secondary transporters and are then accumulated in the appropriate tissues or organ [43]. However, the transport of intermediates of deoxymiroestrol is still unclear and should be further investigated. Nevertheless, comprehensive functional characterization of biosynthetic genes and isotopic labeling of any intermediates is required to test this hypothesis. Alternatively, we do not exclude the possibility that deoxymiroestrol and intermediates could also be biosynthesized in the tuberous cortex, as the corresponding biosynthetic gene expression was lower there than in mature leaves. These lower expression levels of putative miroestrol biosynthetic genes could be partially responsible for the low accumulation levels of deoxymiroestrol and miroestrol. In addition, we found considerably high amounts of puerarin in tuberous cortices. Since puerarin, deoxymiroestrol, and miroestrol share an early biosynthetic pathway (Fig. 1), daidzein may be primarily diverted into the puerarin biosynthetic pathway, producing lower daidzein levels for miroestrol biosynthesis. Moreover, we phylogenetically identified a total of five putative genes annotated as *C*-glycosyltransferase, which plays a key role in producing 8-*C*-glycosylation [44] in the puerarin biosynthetic pathway. One of these putative genes, CL7002, showed

92.32% similarity to previously identified *P. lobata C*-glycosyltransferase (CGT43). These putative genes were highly expressed in tuberous cortices as compared to the other tissues (Supplementary Figure S5). Additionally, several glycosyltransferases have been shown to form a metabolon complex for facilitating efficient production of bioactive compounds [45-47]. These observations could contribute predominantly to daidzein utilization for producing puerarin. Suppression of *Pm*CGT may be an alternative strategy to enhance miroestrol production in tubers.

A different strategy to enhance the production of isoflavonoids and their valuable derivatives is to manipulate transcription factor 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 alignments and expression patterns are informative for determining the functions of individual MYB TFs. Six candidate MYBs (indicated in red in the phylogenetic tree in Fig. 4) are proposed due to their phylogenetic clustering with characterized MYB TFs that are involved in the regulation of isoflavonoid biosynthetic genes. Interestingly, four of these transcription factor genes were highly expressed in the tuberous cortex, which showed the highest accumulation of isoflavones (Fig. 5). Moreover, two of these transcription factors, *Pm*MYB18 and *Pm*MYB75, shared one branch with those MYBs reported as activators for transcription of chalcone reductase and IFS and increased isoflavonoid production[30, 34]. These two MYBs are thus important candidates for further functional characterization, and *Pm*MYB TFs may be associated with the regulation of isoflavonoid biosynthesis and miroestrol production.

Conclusion

We identified several candidate genes encoding key enzymes or transcription factors involved in the biosynthesis of isoflavonoid and miroestrol. Integrative analyses of transcriptomics and metabolomics indicated the complexity of gene expression and metabolite profiles across tissues, suggesting that cortical tuber tissue is a major site of isoflavonoid biosynthesis. Further molecular and biochemical studies on candidate genes involved in miroestrol biosynthesis are required to identify the functions of these candidate genes.

Methods

Plant material, chemicals, and reagents

Leaves and tubers of an approximately two-year-old *P. mirifica* cultivar SARDI190 (Fig. 2a-2d) were obtained from a *P. mirifica* farm at Kasetsart University, Kamphaeng Saen Campus, Nakhonpathom, Thailand (14°1¢N, 99°58¢E). The species was confirmed to be *P. mirifica* by comparison with the voucher specimens no. 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 by the Department of Thai Traditional and Alternative Medicine, Ministry of Public Health, Thailand (DTAM1-2/2561). Seeds of N. benthamiana were germinated on sterilized soil in a culture room

at 25 °C under a 16/8 h light/dark cycle. After germinating for five days, N. benthamiana seedlings were watered twice per week until harvest. Standard compounds of 2'-hydroxydaidzein [22] and 3'-hydroxydaidzein [22] were obtained from Dr. Tomoyoshi Akashi, Nihon University, Japan. 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 was isolated from young leaves, mature leaves, cortex-excised tubers, and tuberous cortices using the Rneasy Plant Mini Kit (Qiagen, USA), and on-column DNA was removed using the Rnase-free Dnase Set (Qiagen, USA). RNA quality was examined by agarose gel electrophoresis, and quantity was measured using a biospectrometer (Eppendorf, Germany). Aliquots of high-quality RNA of all four tissue types were pooled and subjected to Illumina Hiseq 2000 sequencing at BGI, Hongkong. Raw reads were filtered using an in-house software of BGI, filter fq, to produce clean paired-end reads and unpaired reads. Contigs and unigenes were then assembled using Trinity software [48]. We used a BUSCO (Benchmarking Single-Copy Universal Orthologs) version 3.0.2 analysis against the eudicotyledons_odb10 dataset [49, 50] to assess completeness. Raw transcriptome data were deposited in the NCBI sequence read archive (accession number: SRR6917866).

Analysis of gene expression levels across the four tissues of *mirifica* generated using Illumina NextSeq500

Total RNA was extracted from each tissue type using three biological replicates. Based on spectrophotometry, high-quality RNA aliquots of the replicates were pooled per tissue type at equal quantities. The four RNA pools were then subjected to pair-end sequencing using an Illumina NextSeq500 platform. Contigs were assembled using Trinity software [48]. To estimate gene expression levels, we mapped all reads in *fastq* format to the contigs and calculated the reads per kilobase of the transcript per million mapped reads values. A non-parametric approach for identifying DEGs was used to produce four independent pairwise sample comparisons with the softawre NOISeq-sim [51]. Raw transcriptome data were deposited in the NCBI sequence read archive (accession numbers SRR10177497, SRR10177499, SRR10177500, and SRR10177498 for sequences produced from young leaves, mature leaves, cortex-excised tubers, and tuberous cortices, respectively.)

Functional annotation of assembled sequences

The assembled sequences were first compared against the NCBI non-redundant 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 < 1*E*-5; https://www.ncbi.nlm.nih.gov/COG/) by BLASTx, and nucleotide database NT (*e*-value < 1*E*-5; https://www.ncbi.nlm.nih.gov/nucleotide/) using BLASTn. Then, the numbers of unigenes annotated

from each database were counted. Unigenes were mapped to the COG database, potential functions were predicted, and statistical analyses were performed. BLAST2GO software v2.5.0 [52] was used to assign GO categories based on BLAST results, using default settings. All unigene GO functional categories and distribution of gene functions in different species were visualized using WEGO software [53].

Phylogenetic analyses

All putative genes were translated to amino acid sequences using the Expasy translate tool (http://web.expasy.org/translate/). All full-length sequences of putative genes involved in isoflavonoid and chromene synthesis were aligned using BioEdit software (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). We applied a maximum likelihood method with 100 bootstrap replicates using MEGA7 software [54] to produce phylogenetic trees.

Assessment of candidate gene expression using qRT-PCR

Transcription levels of candidate genes were assessed using qRT-PCR. Total RNA isolated from the four tissue types was 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 equal amounts of template DNA in a CFX96 TouchTM Real-Time PCR Detection system with iTaq Universal SYBR Green Supermix (Biorad, USA) and under the following conditions: 95 °C for 8 min followed by 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 candidate gene sequences are shown in Supplementary Table S3. The *P. mirifica* unigene encoding EF1α was amplified as an internal control. Each reaction was performed using three biological replicates. The relative expression ratios of the candidate genes expressed in leaves and tuberous roots were calculated and normalized to a reference gene (EF1α) transcript level in the same sample [55].

Metabolite extraction from *P. mirifica* samples

P. mirifica samples were frozen, ground to a fine powder, and lyophilized overnight. Thirty milligrams of dried powder were dissolved in 500 µL methanol (85%) to extract metabolites under vigorous shaking using a Mixer Mill MM400 (Retsch, Germany) at 25 Hz for 7 min. Crude metabolite extracts were filtered through an Acrodisc® Syringe Filter with a 0.2-µm Supor® membrane (Pall, USA) and stored at -20 °C until analysis.

Phytoestrogen profiling by HPLC quadrupole time of flight mass spectrometry (HPLC-QTOF-MS/MS)

The HPLC-QTOF-MS/MS analysis was performed using an Agilent HPLC 1260 series device coupled with a QTOF 6540 UHD Accurate-Mass system (Agilent Technologies, Germany). The separation of the sample solution was performed on a Luna C18 column (2) 150 × 4.6 mm, 5 µm (Phenomenex, USA). The solvent flow rate was 0.5 mL/min, with 5 µL 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% formic acid (v/v). The linear gradient elution was 5–95% for solvent B at 35 min

with a post-run for 5 min. The column temperature was set at 40 °C. The conditions for the negative ESI source were as follows: drying gas (N_2) flow rate 10 L/min, drying gas temperature 350 °C, nebulizer 30 psig, the fragmentor set to 100 V, capillary voltage 3500 V, and scan spectra from *m/z* 100-1500. The auto MS/MS for the fragmentation was set at collision energies of 10, 20, and 40 V. All data analyses were performed using Agilent MassHunter Qualitative Analysis Software B06.0 (Agilent Technologies, USA).

Molecular cloning and transient expression of *Pm*IFS and *At*MYB12 in *N. benthamiana*

The unigene encoding isoflavone synthase (*Pm*IFS; NCBI sequence accession number MK524721) was amplified from *P. mirifica* cDNA using the following specific primers: forward 5'-ATGTTGCTGGAACTTGCAATTG-3' and reverse 5'-TCAAGAAGGAGGTTTAGATGC-3'. For Arabidopsis MYB12 (*At*MYB12; accession number NM130314), the full-length gene sequence was amplified from *Arabidopsis* cDNA using the following gene-specific primers: forward 5'- ATGGGAAGAGCGCCATGT-3' and reverse 5'- TCATGACAGAAGCCAAGCG-3'. The PCR was performed using 50 µL reaction volumes and a Phusion® HF high-fidelity DNA polymerase (Thermo Fisher Scientific, Finland) with the following thermocycling protocol: 98 °C for 1 min followed by 35 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 1min, and a final extension step at 72 °C for 5 min. PCR products were visualized by agarose gel electrophoresis. Amplicons of target size were ligated into a pJET1.2 vector (Fermentas, USA) for sequencing. The genes were then sub-cloned into pDONOR207 and pEAQ-HT-DEST1 [56] vectors, respectively, using a gateway cloning system according to manufacturer's instructions (Invitrogen, USA). The recombinant pEAQ-HT-DEST1 vectors harboring *PmIFS* and pEAQ-HT-DEST1 vectors harboring *At*MYB12 were electro-transformed into *Agrobacterium tumefaciens* LBA4404. Positive clones were subjected to colony PCR. A single colony of each transformant was grown overnight (at 28 °C and under rotation at 250 rpm) in 5 mL YEB medium supplemented with 100 µg/mL rifampicin, 50 µg/mL kanamycin, and 100 µg/mL streptomycin. The cultures were then washed three times using sterilized distilled water. The resulting pellets were placed in a resuspension solution (10 mM $MgCl_2$, 10 mM MES-K, and 100 µM acetosyringone, at pH 5.6). The *Agrobacterium* solution was adjusted to an A_{600} 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 leave (from the shoot tip) of approximately four-week old *N. benthamiana* plants. After five days, the infiltrated leaves were rapidly frozen using liquid nitrogen and were subsequently ground to a powder. The powder was immediately lyophilized and was then vacuum-stored at room temperature. Thirty milligrams of powder were dissolved in 500 µL of 85% methanol, and metabolites were extracted by vigorous shaking using a Mixer Mill MM400 (Retsch, Germany) at 25 Hz for 7 min. The crude extracts were filtered through an Acrodisc® Syringe Filter with a 0.2 µm Supor® membrane (Pall, USA).

Abbreviations

4CL: 4-coumarate CoA ligase

BUSCO: Benchmarking Single-Copy Universal Orthologs

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 analyzed 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

We would like to thank Dr. Tomoyoshi Akashi (Nihon University, Japan) for providing standard compounds. We extend our gratitude to Dr. George Lomonossoff (John Innes Centre, UK) and Plant Bioscience Limited (UK) for supplying pEAQ vector. HS thanks 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. received a student fellowship from the 100th Anniversary of Chulalongkorn University Fund for Doctoral Scholarship, an Overseas Research Experience Scholarship for Graduate Students from the Graduate School, Chulalongkorn University, and the FY2018 Thesis Grant for Ph.D. Students, National Research Council of Thailand. This 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 program SIAM2014-2015. HS and WD thank 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 analyses, data interpretation, and writing of the manuscript.

Authors' contributions

NS designed and performed the experiments, including data analysis, and wrote 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. 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 Reseearch* 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. Pang T, Ye C-Y, Xia X, Yin W: **De novo sequencing and transcriptome analysis of the desert shrub, *Ammopiptanthus mongolicus*, during cold acclimation using Illumina/Solexa**. *BMC genomics* 2013, **14**(1):488.
20. Huang J, Guo X, Hao X, Zhang W, Chen S, Huang R, Gresshoff PM, Zheng Y: **De novo sequencing and characterization of seed transcriptome of the tree legume *Millettia pinnata* for gene discovery and SSR marker development**. *Molecular Breeding* 2016, **36**(6):75.
21. Liu Z, Chen T, Ma L, Zhao Z, Zhao PX, Nan Z, Wang Y: **Global Transcriptome Sequencing Using the Illumina Platform and the Development of EST-SSR Markers in Autotetraploid Alfalfa**. *PloS one* 2013, **8**(12):e83549.
22. 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.
23. Liu CJ, Huhman D, Sumner LW, Dixon RA: **Regiospecific hydroxylation of isoflavones by cytochrome P450 81E enzymes from *Medicago truncatula***. *The Plant Journal* 2003, **36**(4):471-484.
24. 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.
25. 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.
26. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, Koellensperger G, Huan T, Uritboonthai W, Aisporna AE *et al*: **METLIN: A Technology Platform for Identifying Knowns and Unknowns**. *Analytical chemistry* 2018, **90**(5):3156-3164.
27. 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.

28. Sainsbury F, Lomonossoff GP: **Transient expressions of synthetic biology in plants**. *Current Opinion in Plant Biology* 2014, **19**:1-7.
29. 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.
30. 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.
31. Yi J, Derynck MR, Chen L, Dhaubhadel S: **Differential expression of CHS7 and CHS8 genes in soybean**. *Planta* 2010, **231**(3):741-753.
32. 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.
33. 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**. *The Plant Journal* 2010, **62**(6):1019-1034.
34. 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.
35. Reuter Jason A, Spacek DV, Snyder Michael P: **High-throughput sequencing technologies**. *Molecular Cell* 2015, **58**(4):586-597.
36. Serres-Giardi L, Belkhir K, David J, Glémin S: **Patterns and evolution of nucleotide landscapes in seed plants**. *The Plant cell* 2012, **24**(4):1379-1397.
37. 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.
38. 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.
39. 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.
40. 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**. *The Plant Journal* 2018, **96**(1):56-74.
41. 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.

42. 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.
43. 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.
44. Wang X, Li C, Zhou C, Li J, Zhang Y: **Molecular characterization of the *C*-glucosylation for puerarin biosynthesis in *Pueraria lobata***. *The Plant Journal* 2017, **90**(3):535-546.
45. Nielsen KA, Tattersall DB, Jones PR, Møller BL: **Metabolon formation in dhurrin biosynthesis**. *Phytochemistry* 2008, **69**(1):88-98.
46. 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.
47. 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.
48. 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.
49. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**(19):3210-3212.
50. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics**. *Molecular biology and evolution* 2017, **35**(3):543-548.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.

Supplemental Files

Figure S1. Whole plant of the approximately 3-year-old *P. mirifica*. (bar = 5 cm)

Figure S2. 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 S3. Gene ontology (GO) annotation for all assembled unigenes in *P. mirifica*. The 56 subcategories are affiliated to three main domains: biological process, cellular component, and molecular function. The GO categories were created using WEGO software.

Figure S4. KEGG pathway enrichment analysis of assembled unigenes in *P. mirifica*. **a** The Number of unigenes in 19 sub-categories of metabolic pathway category. **b** The 21 sub-categories of metabolism of terpenoids and polyketides, and other secondary metabolites.

Figure S5. Differential expressed genes (DEGs) predicted as UDP-glycosyltransferases that might be involved in isoflavone biosynthetic pathway across the four tissues of *P. mirifica*.

Figure S6. Ion extract chromatogram of the annotated genistein ([M-H] = 269, RT = 18.78) of methanolic extracts of GFP transiently overexpressed in *N. benthamiana* (**a**), *P. mirifica* Isoflavone synthase (*Pm*IFS) transiently expressed in *N. benthamiana* (**b**), and *P. mirifica* Isoflavone synthase (*Pm*IFS) and Arabidopsis MYB12 transcription factor transiently co-expressed in *N. benthamiana* (**c**), as detected by HPLC-QTOF-MS/MS in negative mode.

Additional file 1.docx - Table S1. Summary of the *P. mirifica* transcriptome assembly.

Additional file 2.docx - Table S2. Annotation statistics.

Additional file 3.docx - Table S3. DNA primer list for qRT-PCR validation.

Figures

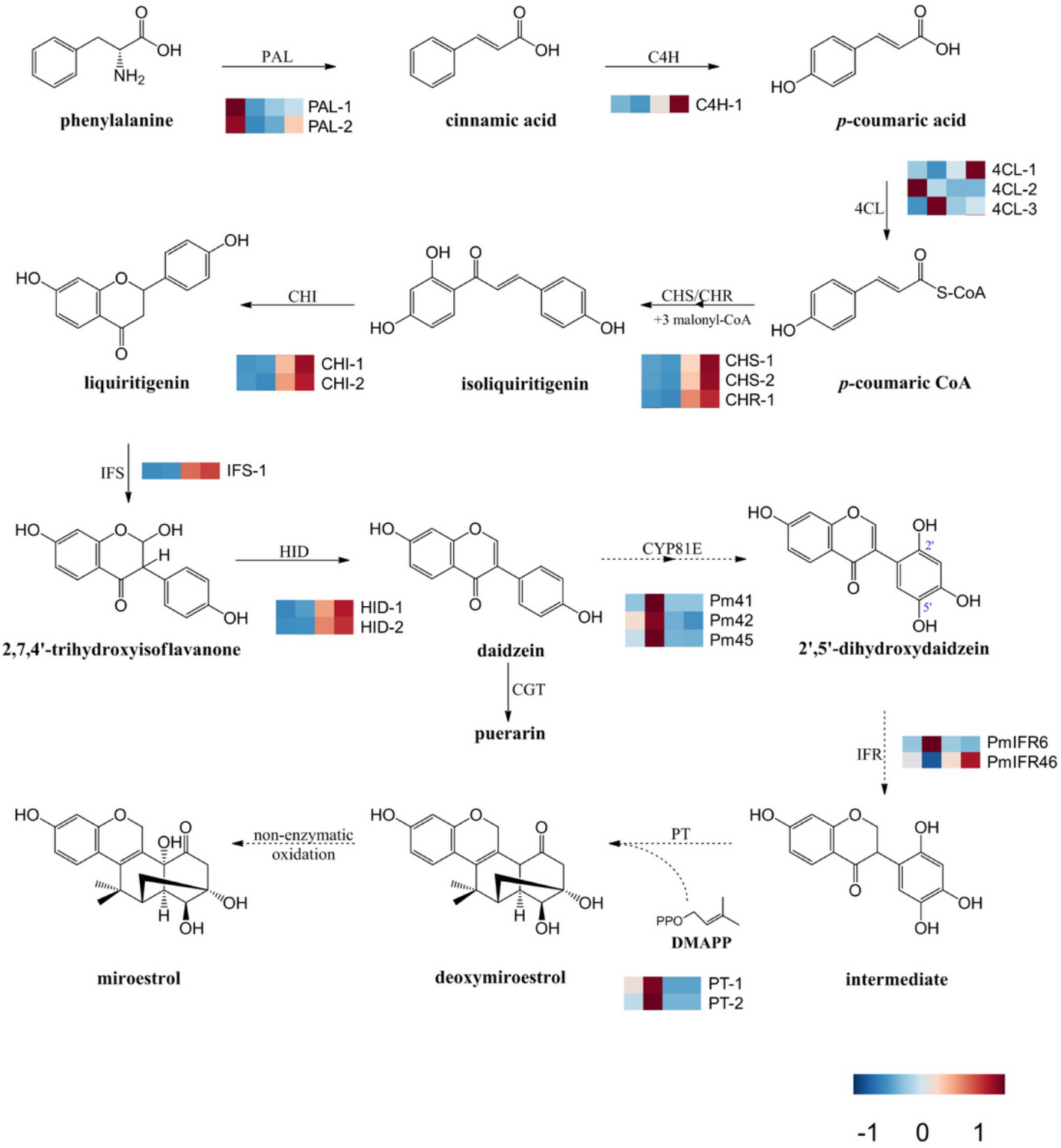

Figure 1

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.

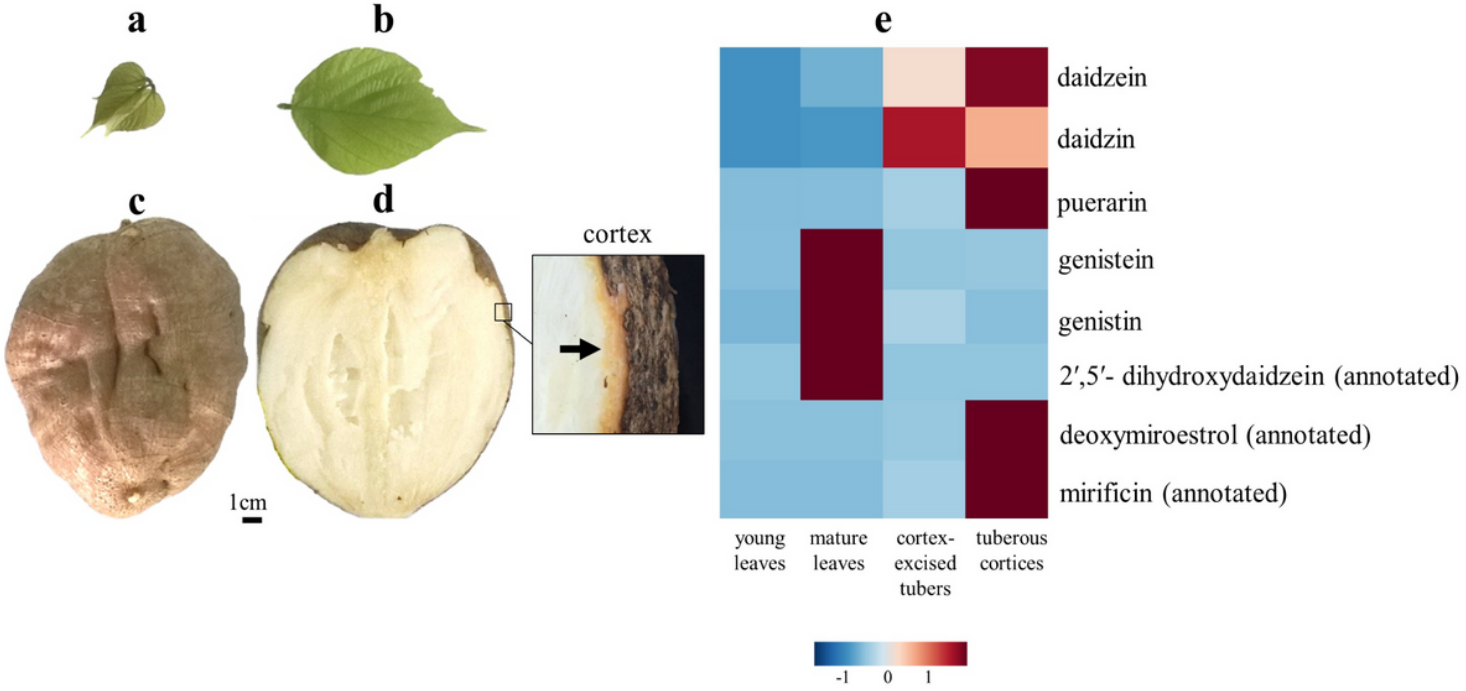

Figure 2

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.

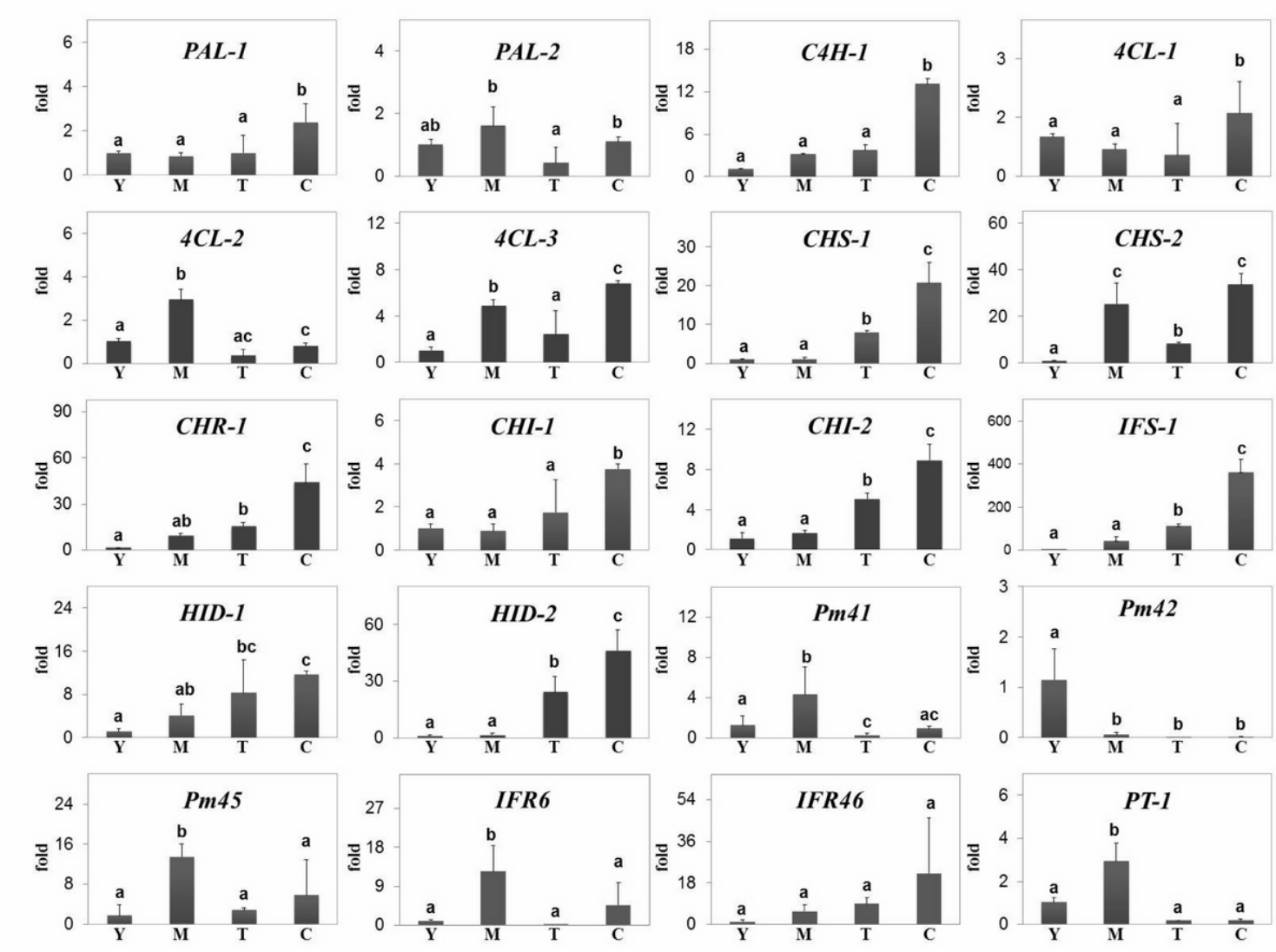

Figure 3

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.

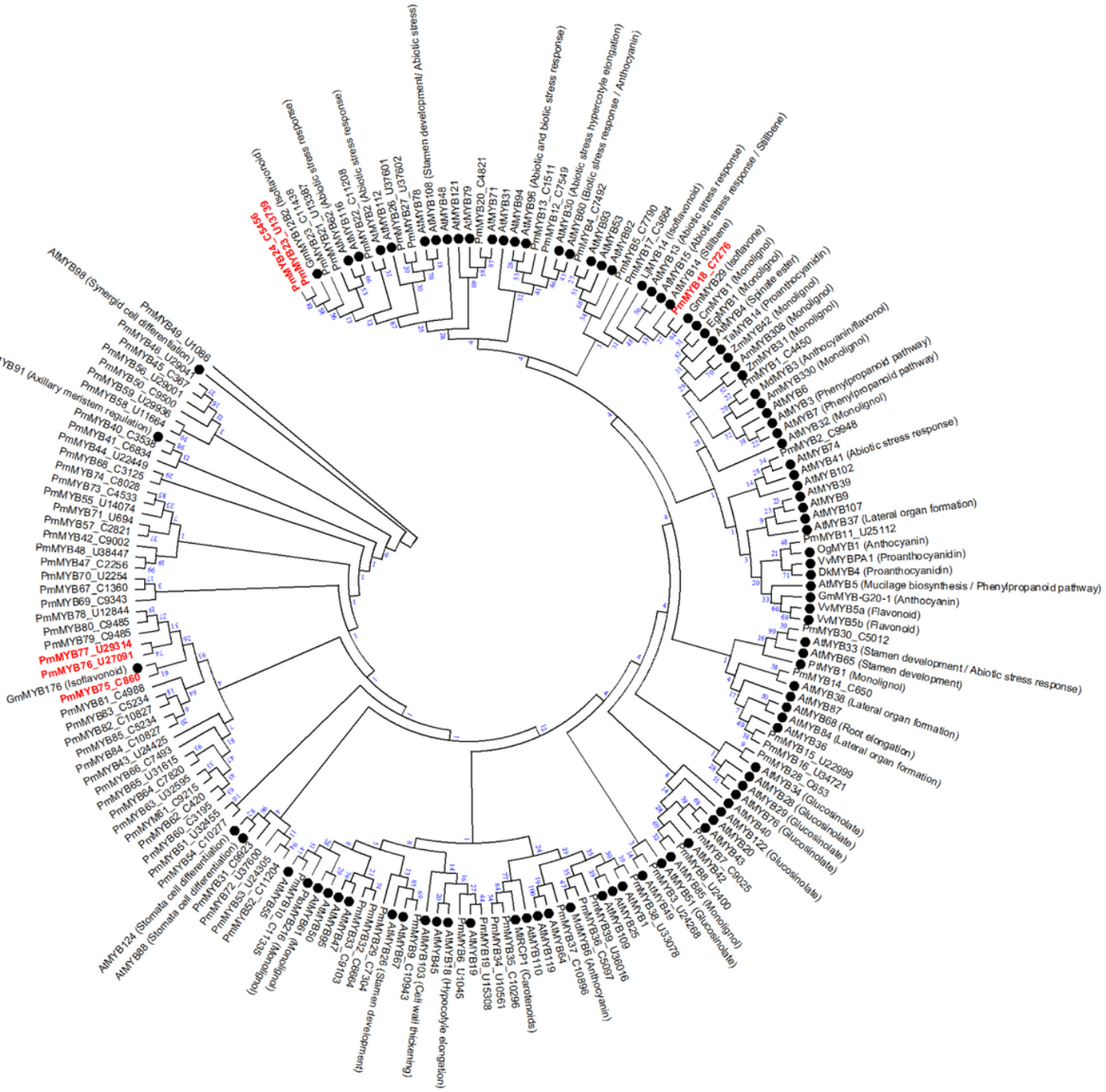

Figure 4

Phylogeny of MYB transcription factors associated with the phenylpropanoid pathway. The tree was constructed using the Maximum likelihood method with putative amino acid full-length MYB sequences with MEGA7. Bootstrap values are shown as percentage (100 replicates). Six P. mirifica candidate MYB transcription factors are indicated in red.

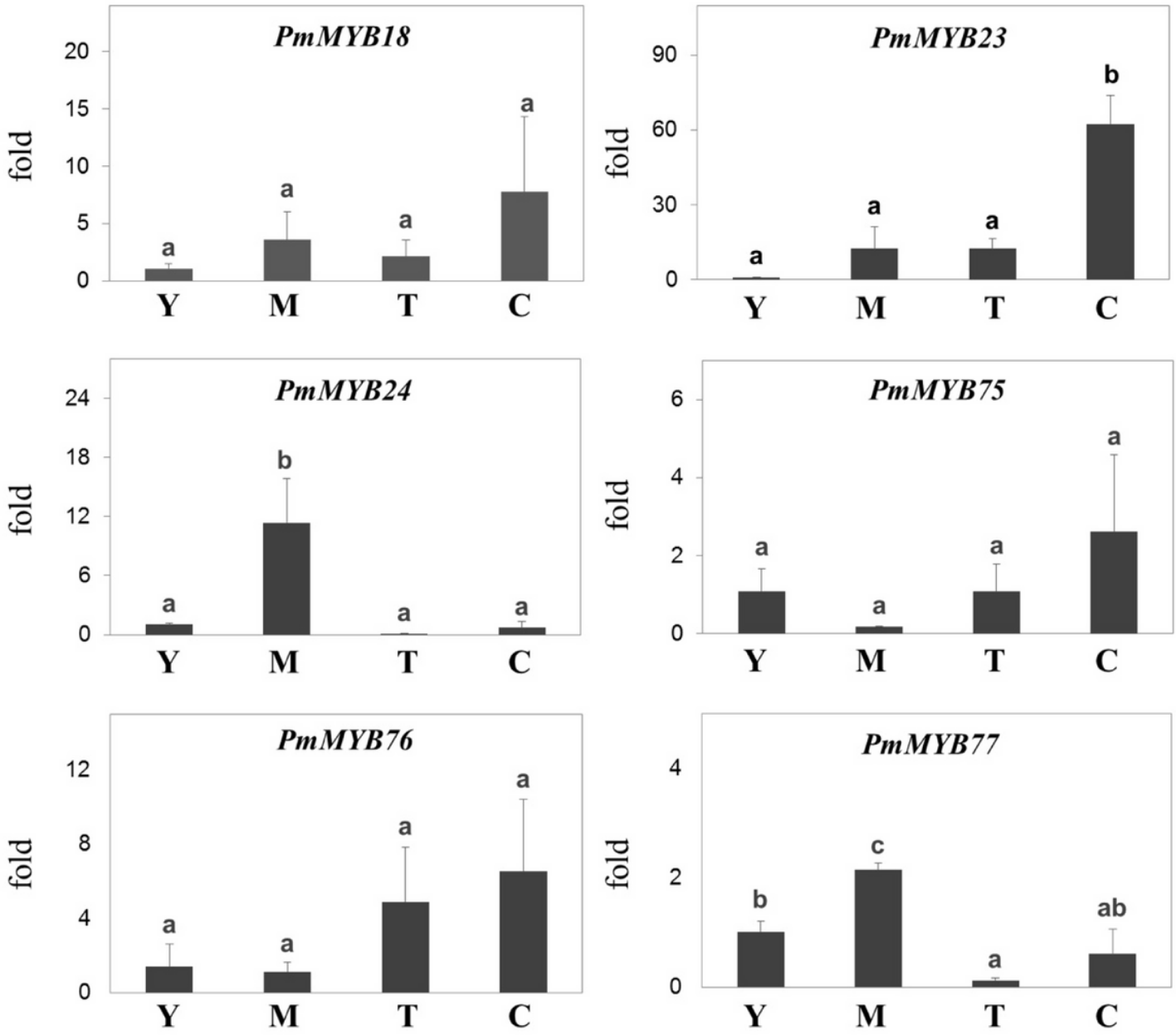

Figure 5

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.

- Fig.S5.tif

- Fig.S2.tif
- Fig.S1.tif
- Fig.S6.tif
- Additionalfile1.docx
- Additionalfile3.docx
- Additionalfile2.docx