BMC Plant Biology

RESEARCH Open Access

Genome-wide identification of the phenylalanine ammonia-lyase gene from *Epimedium Pubescens* Maxim. (Berberidaceae): novel insight into the evolution of the PAL gene family

Chaoqun Xu[1†], Xuelan Fan[1,2†], Guoan Shen[1] and Baolin Guo[1*]

Abstract

Background Phenylalanine ammonia-lyase (PAL) serves as a key gateway enzyme, bridging primary metabolism and the phenylpropanoid pathway, and thus playing an indispensable role in flavonoid, anthocyanin and lignin biosynthesis. PAL gene families have been extensively studied across species using public genomes. However, a comprehensive exploration of PAL genes in *Epimedium* species, especially those involved in prenylated flavonol glycoside, anthocyanin, or lignin biosynthesis, is still lacking. Moreover, an in-depth investigation into PAL gene family evolution is warranted.

Results Seven PAL genes (*EpPAL1*-*EpPAL7*) were identified. *EpPAL2* and *EpPAL3* exhibit low sequence identity to other *EpPALs* (ranging from 61.09 to 64.38%) and contain two unique introns, indicating distinct evolutionary origins. They evolve at a rate ~ 10 to ~ 54 times slower compared to *EpPAL1* and *EpPAL4-7*, suggesting strong purifying selection. *EpPAL1* evolved independently and is another ancestral gene. *EpPAL1* formed *EpPAL4* through segmental duplication, which lead to *EpPAL5* and *EpPAL6* through tandem duplications, and *EpPAL7* through transposed duplication, shaping modern *EpPALs*. Correlation analysis suggests *EpPAL1*, *EpPAL2* and *EpPAL3* play important roles in prenylated flavonol glycosides biosynthesis, with *EpPAL2* and *EpPAL3* strongly correlated with both Epimedin C and total prenylated flavonol glycosides. *EpPAL1*, *EpPAL2* and *EpPAL3* may play a role in anthocyanin biosynthesis in leaves. *EpPAL2*, *EpPAL3*, *EpPAL6*, and *EpPAL7* might be engaged in anthocyanin production in petals, and *EpPAL2* and *EpPAL3* might also contribute to anthocyanin synthesis in sepals. Further experiments are needed to confirm these hypotheses. Novel insights into the evolution of PAL gene family suggest that it might have evolved from a monophyletic group in bryophytes to large-scale sequence differentiation in gymnosperms, basal angiosperms, and Magnoliidae. Ancestral gene duplications and vertical inheritance from gymnosperms to angiosperms likely occurred during PAL evolution.

[†] Chaoqun Xu and Xuelan Fan contributed equally to this work and share the first authorship.

*Correspondence:
Baolin Guo
blguo@implad.ac.cn

Full list of author information is available at the end of the article

Most early-diverging eudicotyledons and monocotyledons have distinct histories, while modern angiosperm PAL gene families share similar patterns and lack distant gene types.

Conclusions *EpPAL2* and *EpPAL3* may play crucial roles in biosynthesis of prenylated flavonol glycosides and anthocyanins in leaves and flowers. This study provides novel insights into PAL gene family evolution. The findings on PAL genes in *E. pubescens* will aid in synthetic biology research on prenylated flavonol glycosides production.

Keywords *Epimedium pubescens*, Phenylalanine ammonia-lyase gene, Evolution, Prenylated flavonol glycoside, Expression profiling

Introduction

Phenylalanine ammonia-lyase (PAL, EC 4.3.1.24) is the first critical enzyme in the phenylpropanoid pathway, catalyzing the biotransformation of L-phenylalanine to trans-cinnamic acid. Acting as a bridge, PAL mediates carbon flux from primary to secondary metabolism, leading to the production of flavonoids, anthocyanins, tannins, lignins, phytoalexins and other benzene-based compounds with pharmaceutical value [1, 2]. The phenylpropanoid derivatives play crucial roles in plant defense against a range of biotic (e.g., insects and pathogens) and abiotic stresses (e.g., UV light, low temperature and nutrient stress). These compounds function as regulatory molecules, participating in signal transduction and communication with other organisms [3]. Furthermore, they contribute to lignin biosynthesis, which is essential for maintaining stem rigidity, vascular integrity, and providing a physical barrier against invading pathogens in plants [4, 5].

PAL enzymes in dicots typically exhibit monofunctionality, specifically catalyzing the PAL reaction. However, in certain monocots, particularly those belonging to the grass family Poaceae, PAL enzymes can display bifunctionality, catalyzing both PAL and TAL reactions with phenylalanine and tyrosine as substrates, respectively. Notably, PAL and TAL enzymes are absent in animals, where they have been replaced by HAL (L-histidine ammonia-lyase) [2, 6, 7]. The PAL gene family typically consists of 2–6 copies, although some species possess significantly more members [8, 9]. Over the course of evolution, the expression of PAL genes in response to biotic and abiotic stresses has become highly regulated in a temporal and spatial manner, often resulting in the diversification of gene copies with redundant functions [10–12]. Given the diverse functions of PAL gene copies, it can be challenging to determine which copy predominantly modulates the biosynthesis of different branch end-products.

Herba Epimedii, a valued traditional Chinese medicine (TCM), is sourced exclusively from the dried leaves of four *Epimedium* species: namely *E. pubescens*, *E. sagittatum*, *E. brevicornu* and *E. koreanum*. Besides its traditional uses as a kidney tonic and antirheumatic agent [13, 14], the aglycone of its primary bioactive components, known as prenylated flavonol glycosides (PFGs), particularly icaritin, has garnered recognition as a novel drug effective in inhibiting hepatocellular carcinoma (HCC) initiation and malignant growth [15, 16]. However, the genes involved in PFGs biosynthesis in *Epimedium*, including PAL, remain fragmented. To date, only three PALs (*EsPAL1*, *EsPAL2* and *EsPAL3*) in *E. sagittatum* [17] and one PAL (*EwPAL*) in *E. wushanense* [18] have been reported. Through qRT-PCR and correlation analysis, only *EsPAL3* has been implicated in both PFGs and anthocyanin pathways. Meanwhile, *EsPAL1* is suspected to play a role in lignin biosynthesis, while *EsPAL2* demonstrates constitutive expression across all tissues, hinting at its potential involvement in lignin, PFGs and anthocyanin pathways. *EwPAL*, on the other hand, has been solely linked to the biosynthesis of naringenin. Further research into the PALs responsible for PFGs, anthocyanin or lignin biosynthesis needs to be clearly explored, which would facilitate more efficient synthesis of PFGs.

Previous studies on the evolution of the PAL gene family are limited, with only a few notable investigations reported: in *Nelumbo nucifera* by Wu et al. (2014) [19] and in three Cucurbitaceae plants by Dong et al. (2016) [8]. While the former study identified a distinct ancient *NnPAL1* gene in the early-diverging dicotyledonous plant *N. nucifera*, it failed to explore similar patterns in other early-diverging dicotyledonous species. On the other hand, the latter study exclusively examined PAL evolution within Cucurbitaceae plants without delving into the evolutionary origins of the discovered PAL genes. *E. pubescens*, as another early-diverging dicotyledonous plant, emerges as a valuable subject for studying the evolution of the PAL gene family. Therefore, conducting in-depth research on this species becomes particularly significant.

In this study, a genome-wide search of *E. pubescens* led to the identification of 7 PAL genes. Among these, *EpPAL2* and *EpPAL3* exhibit significant sequence divergences, yet there is a lack of research exploring their distinct functional traits. This study aims to delve deeper into the gene functions of each *EpPAL*, with a particular focus on *EpPAL2* and *EpPAL3*. Furthermore, by utilizing *E. pubescens* as a representative of an early-diverging angiosperm, we aim to integrate prior research and employ a comprehensive set of 24 representative species to thoroughly examine the evolutionary history of the PAL gene family. The findings of this study have implications for deepening our understanding of the molecular functions of various *EpPALs* and providing valuable insights into the evolution of the PAL gene family.

Results

Identification and chromosomal localization of *EpPAL*s

Putative PAL genes were retrieved and identified from *E. pubescens* by HMM search and BLASTP. A total of 7 PAL genes were identified. The full length of candidate genes was further confirmed to be correct using available transcriptome data of *E. pubescens*. According to the subcellular localization predictions, all *EpPALs* are in the cytoplasm. Detailed information of those genes is presented in Table 1, including gene IDs, gene names, exon numbers, chromosome locations and protein length. The naming of *EpPAL*s was done according to the order of PAL on the chromosome of *E. pubescens*. 7 *EpPALs* were unevenly distributed across the whole genome. Specifically, *EpPAL1* was located on chromosome 1, *EpPAL2* and *EpPAL3* were on chromosome 4, while *EpPAL4-EpPAL7* were allocated on chromosome 6 (Table 1).

Table 1 The general information, structural features and properties of PALs in *E. pubescens*

Gene ID	Name	Exon numbers	Location	Length (aa)
Ebr01G053210.1	*EpPAL1*	2	Chr01: 458320539-458323905	705
Ebr04G040710.1	*EpPAL2*	3	Chr04: 319850975-319854451	739
Ebr04G040750.1	*EpPAL3*	3	Chr04: 320081476-320084690	711
Ebr06G007280.1	*EpPAL4*	2	Chr06: 50165407-50170790	710
Ebr06G007290.1	*EpPAL5*	2	Chr06: 50172988-50175932	710
Ebr06G007300.1	*EpPAL6*	2	Chr06: 50265138-50268288	708
Ebr06G041720.1	*EpPAL7*	2	Chr06: 330631097-330686057	708

Evolutionary analysis of PAL genes

To gain a deeper understanding of the evolution of PAL genes, we conducted a comprehensive analysis utilizing a diverse set of PAL members from 24 species, including *Chara braunii* (a charophyte), *Physcomitrium patens* (a bryophyte), *Ceratopteris richardii* (a fern), *Ginkgo biloba* and *Sequoiadendron giganteum* (gymnosperms), *Amborella trichopoda* and *Nymphaea colorata* (basal angiosperms), *Ceratophyllum demersum* (Ceratophyllales), *Cinnamomum kanehirae*, *Liriodendron chinense* and *Piper nigrum* (Magnoliidae), *Spirodela polyrhiza* (a basal monocot), *Papaver somniferum*, *N. nucifera*, *Tetracentron sinense*, *Macadamia integrifolia*, *Aquilegia coerulea*, *Kingdonia uniflora* and *E. pubescens* (early-diverging eudicotyledons), *Brachypodium distachyon* (an early-diverging monocotyledon), *Oryza sativa* (a representative modern monocotyledon), as well as *A. thaliana*, *Cucumis sativus* and *Vitis vinifera* (representative modern dicotyledons). Detailed gene mining instructions and sequences of the PAL genes from these 24 species are provided in Table S1 and Table S2, respectively.

We constructed a phylogenetic tree using the maximum likelihood method to illustrate the differentiation profile of PAL genes among the aforementioned taxa (Fig. 1). As shown in Fig. 1, the tree clearly divides all PAL genes into six distinct clusters (Cluster 1–6). Cluster 1 represents an evolutionary branch unique to charophytes, with a significantly longer branch length compared to other clusters, indicating that its PAL evolutionary clade is the farthest from the others. Cluster 2 and 4 exclusively contain gymnosperm genes, suggesting they may represent specific evolutionary branches unique to gymnosperms. Cluster 3 and 5 mainly include PAL genes from bryophytes, ferns, gymnosperms, basal angiosperms, Ceratophyllales, Magnoliidae, and early-diverging eudicotyledons, indicating a more primitive group of PALs. Cluster 6 primarily encompasses PAL clades from typical modern dicotyledons and monocotyledons, as well as homologous genes in ferns, gymnosperms, basal angiosperms, Magnoliidae, Ceratophyllales, and other taxa corresponding to modern PAL genes.

Based on these 12 different evolutionary branches (charophyte, bryophyte, fern, gymnosperms, basal angiosperms, Ceratophyllales, Magnoliidae, basic taxa of eudicotyledon, early-diverging eudicotyledons, an early-diverging monocotyledon, a representative modern monocotyledon, and representative modern dicotyledons), we can roughly outline the differentiation profile of PAL genes among various taxonomic groups. The PAL genes of charophytes and bryophytes each form a monophyletic group, respectively. Ferns exhibit two major sequence divergences. Significant differentiation occurred in gymnosperm, resulting in five branches: two unique to gymnosperms, two shared by some primitive taxa and early-diverging angiosperm ancestors, and one clustering with modern taxa. Both basal angiosperms and Magnoliidae also show substantial differentiation. Most early-diverging angiosperms, including *Epimedium*, and ancestral monocotyledon taxa possess ancestral primitive group genes. Modern monocotyledons and dicotyledons, represented by *O. sativa* and *A. thaliana*, mostly form monophyletic groups, and most of them do not have ancestral group genes. Thus, the evolution of PAL genes has progressed from a monophyletic group in bryophyte with small-scale functional differentiation

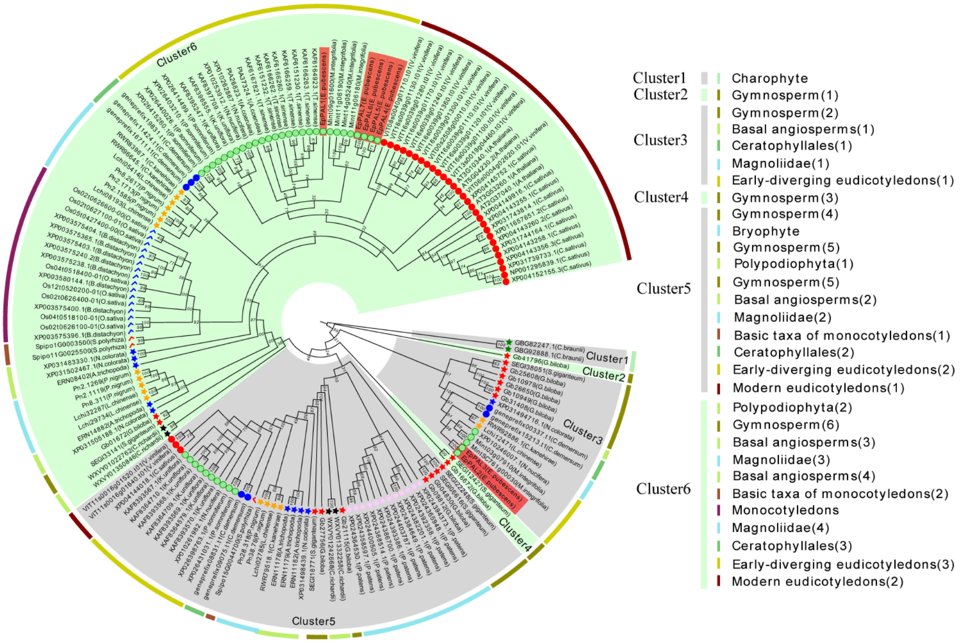


Fig. 1 Phylogenetic tree of PAL sequences of 24 plant species with typical evolutionary relationships. *EpPALs* are marked with red box. The representation of the markings at the branch nodes is as follows: green five-pointed star: charophyte, red five-pointed star: gymnosperm, blue five-pointed star: basal angiosperms, orange five-pointed star: Magnoliidae, pink five-pointed star: bryophyte, black five-pointed star: polypodiophyta, green circle: early differentiated angiosperms, blue circle: Ceratophyllales, red circle: modern eudicotyledons, red checkmark: basic taxa of monocotyledons, and blue checkmark: monocotyledons. The attributes of each gene in the six clusters represented on the right side, and the numbers in parentheses represent branches of corresponding attributes. Taking gymnosperm as an example, it can be subdivided into six branches, represented by Gymnosperm (1) to Gymnosperm (6) respectively. Detailed species source information of PAL, as well as sequences can be referred to Table S1 and Table S2, respectively

to large-scale sequence differentiation in gymnosperms, basal angiosperms, and Magnoliidae. Most early-diverging eudicotyledons and monocotyledons have different evolutionary histories, while modern angiosperm taxa tend to be monophyletic with few ancestral group genes.

Among the seven *EpPALs* identified in our study, two genes (*EpPAL2* and *EpPAL3*) cluster together with the ancestral type PAL genes represented by Cluster 3, supported by high bootstrap values. The remaining five genes (*EpPAL1* and *EpPAL4-7*) form a separate monophyletic group (Cluster 6) with equally strong bootstrap support (Fig. 1). These findings suggest that the PAL genes in *Epimedium* originated from at least two ancestral PAL homologous genes. The presence of *EpPALs* in different branches of the phylogenetic tree indicates their diverse evolutionary histories and potential functional diversification within the species.

Gene structure analysis of PAL genes

Statistical analysis of 167 PAL genes from 24 species primarily revealed three distinct intron insertion patterns. Pattern 1 (59/167) lacked introns, Pattern 2 (65/167) had a single intron with a shorter front-end exon (~400 bp) compared to the back-end (~1750 bp), and Pattern 3 (7/167) contained two introns with exon lengths of ~1140 bp, ~540 bp, and ~540 bp, respectively (Figure S1). Notably, in *E. pubescens*, intron lengths varied widely, from 422 bp in *EpPAL5* to 2862 bp in *EpPAL4*, while exon lengths were highly conserved. *EpPAL2* and *EpPAL3* followed Pattern 3, but differed in intron phase: *EpPAL2* had two phase 0 introns, whereas *EpPAL3* had one phase 2 and one phase 0 intron. Both genes had conserved glutamine codon (CAG) at the second exon/intron boundary. All other *EpPALs* exhibited Pattern 2, with conserved intron insertion sites between the second and third bases of specific codons: arginine (CGA) for *EpPAL1*, *EpPAL6*, and *EpPAL7*, and isoleucine (AUU) for *EpPAL4* and *EpPAL5* (Fig. 2b and Figure S1). These differences suggest independent origins for the intron insertion events in these two gene sets.

Using a fungal PAL gene from *Neurospora tetrasperma* (NCBI accession number: EGZ69514.1) as an outgroup, we determined the root position of the phylogenetic tree composed of *EpPALs* to infer their lineage. The research results support the division of *EpPALs* into two major clades, with Clade1 being further subdivided into two sub-clades (clade1_1 and clade1_2). Detailed homology detection and structural alignment among *EpPALs* were provided (Fig. 2b and Table S3). Pairwise identity between *EpPALs* (excluding *EpPAL2* and *EpPAL3*) ranged from 85.10 to 100%, indicating close relatedness. However, protein identity between Clade2 and Clade1 was

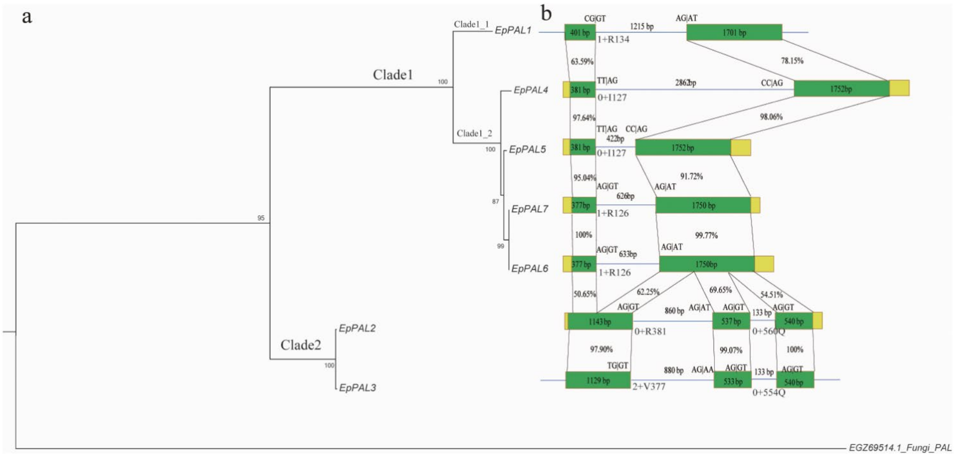


Fig. 2 Phylogenetic tree and genomic structure of *EpPALs*. (**a**) Phylogenetic tree of *EpPALs*. The numbers below the branches represent the bootstrap values; (**b**) Gene structure of *EpPALs*. Green boxes, lines and yellow boxes represent the exon, intron and UTR, respectively. The percentages indicate the similarity of fragments between each pair of *EpPALs*. The exon/intron borders were displayed on top of the structural model, the intron phase (the numbers 0, 1 and 2, which represent introns between codons, introns between the first and second bases of a codon, introns between the second and the third bases of a codon, respectively), the amino acid residues affected by intron insertion events and its position were displayed under the structural model

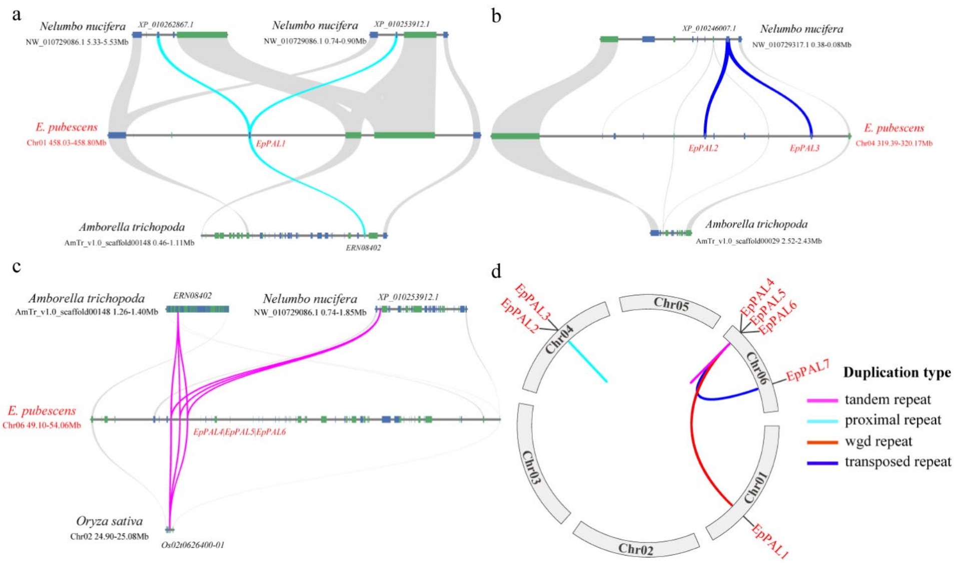


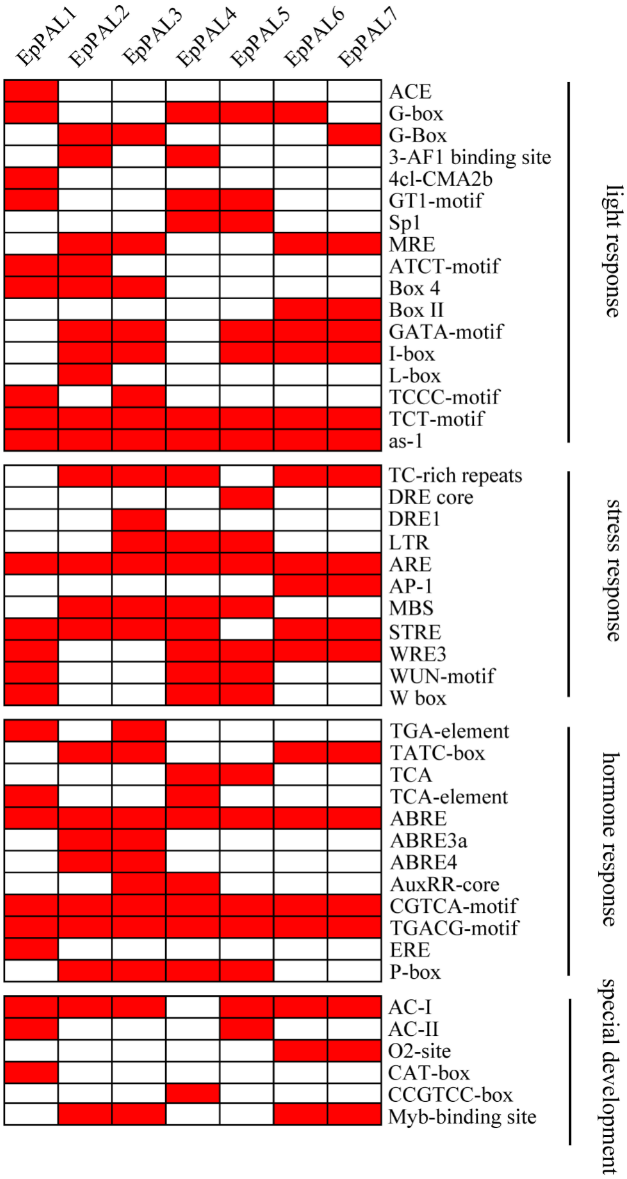
Fig. 3 Inter- and intra-specific collinearity analysis of *EpPALs*. (**a**) Microsynteny analysis between the *EpPAL1* loci and their respective collinear counterparts in *A.trichopoda* and *N. nucifera*. The collinear PAL genes and the syntenic flanking genes are connected by colored and gray lines, respectively; (**b**) Microsynteny analysis of *EpPAL2* and *EpPAL3*; (**c**) Microsynteny analysis of *EpPAL4*, *EpPAL5* and *EpPAL6*; (**d**) Intraspecific collinearity analysis of *EpPALs*. Red line represents a collinear gene pair of *EpPALs* (the two ends were *EpPAL1* and *EpPAL4*, respectively). Purple line represents a tandem repeat between *EpPAL4* and *EpPAL5-6*. Cyan line represents a proximal repeat between *EpPAL2* and *EpPAL3*. Blue line represents a transposed repeat between *EpPAL4* and *EpPAL7*

lower, ranging from 61.09 to 64.38% (Fig. 2a and Table S3), suggesting multiple ancestral origins for the formation of *EpPALs*. Phylogenetic analysis (Fig. 1), sequence alignment (Table S3) and gene structure (Fig. 2b) collectively provide compelling evidence supporting the hypothesis that *EpPAL2* and *EpPAL3* trace their origins to a unique ancestral gene, whereas the other *EpPALs* descended from a different ancestral gene.

Collinearity analysis of PALs

To determine the origin of *EpPALs* through duplication events, both inter- and intra-specific collinearity analyses were conducted (Fig. 3). We selected three species from different evolutionary branches (*A. trichopoda*, *N. nucifera*, and *O. sativa*) along with *E. pubescens* for microsynteny analysis of interspecies local regions related to PAL genes. Collinear blocks were detected for *EpPAL1* to *EpPAL6*, except for *EpPAL7*. The analysis of *EpPAL1*, *EpPAL2* and *EpPAL3* revealed collinear blocks only among *E. pubescens*, *A. trichopoda*, and *N. nucifera*, with no collinear block detected in *O. sativa*. Notably, homologous genes corresponding to *EpPAL2* and *EpPAL3* were not detected in *A. trichopoda*. Considering the detection of collinear blocks for *EpPAL1-3* only in relatively primitive evolutionary branches and not in more modern species, it is speculated that the *EpPAL1-3* may represent primitive types of the PAL family in *E. pubescens*. Further referencing Fig. 1, we infer that the ancestral genes of *EpPAL2* and *EpPAL3* originated differently from that of *EpPAL1*. The ancestral genes of *EpPAL2* and *EpPAL3* belong to Clade 3, a more primitive branch, while the ancestral gene of *EpPAL1* belongs to Cluster 6, a branch present in modern monocots and dicots. For *EpPAL4-6*, we detected relevant collinear blocks in species from three different evolutionary branches. Combining this with the evolutionary positions of these three genes in Fig. 1, we infer that they may have originated from gene duplication of either the *EpPAL1* branch or the *EpPAL2-3* branch.

To gain a deeper understanding of the relationships among these genes, we conducted further intraspecific collinearity analysis in *E. pubescens* and analyzed gene duplication patterns using DupGen-finder (Fig. 3d and Table S4). The results indicated that the gene pair *EpPAL1* and *EpPAL4* underwent segmental/whole-genome duplication (Fig. 3d). *EpPAL5* and *EpPAL6* originated from tandem gene duplication of *EpPAL4*, while *EpPAL7* emerged from a transposed gene duplication event involving *EpPAL4* (Fig. 3d, Figure S2 and Table S4). Therefore, duplication events have played a significant role in the evolution of *EpPALs*. The evolutionary profile of *EpPALs* can be summarized as follows: the ancestral genes of *E. pubescens* are *EpPAL2* and *EpPAL3*, with *EpPAL1* originating independently. *EpPAL4* was then acquired through intraspecific whole-genome duplication from *EpPAL1*. *EpPAL4* underwent tandem duplication to produce *EpPAL5* and *EpPAL6*, and transposition duplication to generate *EpPAL7*.



	EpPAL1	EpPAL2	EpPAL3	EpPAL4	EpPAL5	EpPAL6	EpPAL7	
ACE	🟥							light response
G-box	🟥			🟥	🟥	🟥		
G-Box		🟥	🟥				🟥	
3-AF1 binding site		🟥		🟥				
4cl-CMA2b	🟥							
GT1-motif	🟥			🟥	🟥			
Sp1				🟥	🟥			
MRE		🟥	🟥			🟥	🟥	
ATCT-motif	🟥	🟥						
Box 4	🟥	🟥	🟥					
Box II						🟥	🟥	
GATA-motif		🟥	🟥		🟥	🟥	🟥	
I-box		🟥	🟥		🟥	🟥	🟥	
L-box		🟥						
TCCC-motif	🟥		🟥					
TCT-motif	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
as-1	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
TC-rich repeats		🟥	🟥	🟥		🟥	🟥	stress response
DRE core					🟥			
DRE1			🟥					
LTR			🟥	🟥	🟥			
ARE	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
AP-1						🟥	🟥	
MBS		🟥	🟥	🟥	🟥			
STRE	🟥	🟥	🟥	🟥		🟥	🟥	
WRE3	🟥			🟥	🟥	🟥	🟥	
WUN-motif	🟥			🟥	🟥			
W box	🟥			🟥	🟥			
TGA-element	🟥		🟥					hormone response
TATC-box		🟥	🟥			🟥	🟥	
TCA				🟥	🟥			
TCA-element	🟥			🟥				
ABRE	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
ABRE3a		🟥	🟥					
ABRE4		🟥	🟥					
AuxRR-core			🟥	🟥				
CGTCA-motif	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
TGACG-motif	🟥	🟥	🟥	🟥	🟥	🟥	🟥	
ERE	🟥							
P-box		🟥	🟥	🟥	🟥			
AC-I	🟥	🟥	🟥		🟥	🟥	🟥	special development
AC-II	🟥				🟥			
O2-site						🟥	🟥	
CAT-box	🟥							
CCGTCC-box				🟥				
Myb-binding site		🟥	🟥			🟥	🟥	

Fig. 4 *Cis*-regulatory elements of *EpPAL*s in upstream region of 1500 bp. Four different types of *cis*-regulatory elements and different *cis*-regulatory elements of all *EpPALs* were provided

Conserved motif identification and *cis*-regulatory elements analysis

Eight conserved motifs followed a consistent distribution pattern of 6-4-7-3-8-1-2-5 among all EpPAL proteins (Figure S3). These motifs, ranging from 26 to 100 amino acids in length, were identified across all EpPALs (Figure S3 and Table S5). Notably, the MIO (4-methylidene-imidazolone-5-one) domain, characterized by the highly conserved signature sequence GTITASGDLVPLSYIA, contained the enzymatic active site Ala-Ser-Gly, which was preserved in all EpPAL proteins (Figure S3). With the exception of EpPAL3, which lacked the active site 158 L, the 388 F substrate-selective binding site, and the 350R phosphorylation site, both the active sites and substrate-specific binding sites were conserved across all EpPAL proteins (Figure S4). Additionally, while the 538 S phosphorylation site was serine in EpPAL2, EpPAL3 and EpPAL4, it was replaced by threonine in the remaining EpPAL proteins.

A total of 56 *cis*-regulatory elements (CREs) with known functions were identified, including 17, 19, 12, and 8 CREs related to light, stress, hormone, growth, and development responses, respectively (Fig. 4). With the exception of *EpPAL4*, all *EpPAL* genes possess at least one AC element essential for lignin synthesis. However, no MBSI element related to the regulation of flavonoid biosynthetic genes was found. ARE (anaerobic induction), ABRE (abscisic acid-responsiveness), CGTCA-motif, and

TGACG-motif (MeJA-responsiveness) were identified in all *EpPAL* genes, suggesting that responses to oxygen, abscisic acid, and methyl jasmonate are essential functions shared by all *EpPALs*. Additionally, ERE (ethylene-responsiveness) was exclusively present in *EpPAL1*, while the CCGTCC-box was only found in *EpPAL4*. These findings suggest that *EpPAL1* and *EpPAL4* may be specifically associated with ethylene response and activation of the meristem, respectively. Overall, these results imply that different *EpPAL* genes may have distinct yet overlapping biological functions, playing roles in processes such as growth and development, hormone response, and environmental stress response.

Natural selection analysis

Natural selection tests were conducted using PAML (v.4.1) under different hypotheses. For the branch-specific model, four hypotheses were tested as outlined in Table S6. The three-ratio model, which assigns distinct ω values to each of the three clades (ω[Clade1_1] ≠ ω[Clade1_2] ≠ ω[Clade2]), emerged as a significantly better fit to the dataset compared to the other models tested (df=1, *P*=1.13e-07). Notably, the two-ratio model also outperformed the one-ratio model (df=1, *P*=2.62e-08), as illustrated in Fig. 2a and detailed in Table S6. These findings suggest that each clade experienced unique selection pressures, with ω ratios of 0.20, 0.037, and 0.0037 for Clade1_1, Clade1_2, and Clade2, respectively (Fig. 2a and Table S6). Notably, Clade2 exhibited strong purifying selection, evolving approximately 54 and 10 times slower than Clade1_1 and Clade1_2, respectively. Clade1_2 followed in terms of purifying selection, while *EpPAL1*, originating from Clade1_1, exhibited relatively higher divergence.

To further investigate whether ω varied across all amino acid sites or specific sites within particular branches, both the site-specific model and the branch-site model were employed. Under the site-specific model, three amino acid residues under positive selection when comparing selection M1 versus neutral M1, as well as M7 versus M8. Additionally, when setting Clade1_1, Clade1_2, and Clade2 as the foreground branches, 8, 7, and 76 amino acid sites were found under positive selection, respectively (Table S6). These candidate positively selected sites provide valuable insights into the evolutionary history of *EpPALs*.

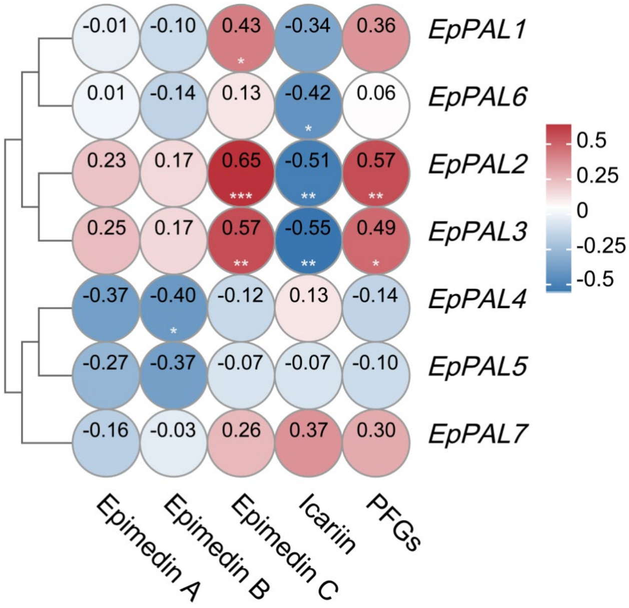


Fig. 5 Correlation heatmap of *EpPALs* with Epimedin **A**, Epimedin **B**, Epimedin **C**, Icariin and total PFGs. The significance levels were set as follows: unmarked stands for *P*>=0.05, * stands for 0.01<*P* <0.05, ** stands for 0.001<*P*<0.01, *** stands for *P*<=0.001

Expression patterns of *EpPALs* and determination of *EpPALs* related to prenylated flavonol glycosides

The *EpPALs* were identified in five distinct tissues, as depicted in Figure S5, with varying expression patterns across these tissues. *EpPAL1* demonstrated high expression levels in all tested tissues, peaking in the leaf and flower, and gradually decreasing in the root, fruit, and stem, suggesting a fundamental role. In contrast, *EpPAL2* and *EpPAL3* showed elevated expression in leaf but were barely detectable in the stem, while *EpPAL5* was predominantly expressed in leaf and root. *EpPAL4*, *EpPAL6*, and *EpPAL7* had minimal or low levels of expression.

To further investigate the relationship between *EpPAL* expression and the contents of key bioactive compounds, a transcriptome analysis using samples from five different leaf development stages. The findings, along with the corresponding *EpPAL* expression levels and bioactive compound contents, are outlined in Table S7. Additionally, a correlation analysis was performed, and the results are visually presented in Fig. 5.

By integrating transcriptome data with targeted metabolite measurements, significant correlations were discovered between the expression of *EpPAL2* and *EpPAL3* with Epimedin C (*EpPAL2*: *r*=0.65, *P*<0.001; *EpPAL3*: *r*=0.57, *P*<0.01) and total PFGs (*EpPAL2*: *r*=0.57, *P*<0.01; *EpPAL3*: *r*=0.49, *P*<0.05). Notably, *EpPAL1* holds the third position in terms of correlation strength. However, a notable negative correlation was observed between these genes and Icariin (*EpPAL2*: *r* = -0.51, *P*<0.01; *EpPAL3*: *r* = -0.55, *P*<0.01). The remaining *EpPALs* demonstrated either weak (*EpPAL6*) or no correlation (*EpPAL4*, *EpPAL5* and *EpPAL7*), implying that *EpPAL1*, *EpPAL2*, and *EpPAL3*, particularly the latter two, might have a pivotal role in the biosynthesis of PFGs.

Determination of *EpPALs* related to anthocyanin biosynthesis in flowers and leaves

To further identity the *EpPAL* isoforms responsible for anthocyanin biosynthesis, various groups of *Epimedium* species with distinct petal, sepal, and leaf colors were studied (Fig. 6a). In leaves, *EpPAL1-3* align with the observed color phenotypes, with *EpPAL2* and *EpPAL3* showing significantly higher expression in magenta leaves compared to green leaves, while *EpPAL1* did not exhibit a notably high expression pattern (Fig. 6b). Co-expression analysis with *EpANSs* and *EpDFRs* revealed a significant positive correlation between *EpPAL1-EpPAL3* and the expression of DFR and ANS genes, suggesting their involvement in anthocyanin biosynthesis in leaves, whereas *EpPAL6* and *EpPAL7* showed a significant negative correlation (Fig. 6e and Table S9).

Similarly, in flowers, *EpPAL2* and *EpPAL3* showed significantly elevated expression in magenta petals compared to yellow petals and green petals (Fig. 6c). Both genes exhibited similar expression patterns across different sepal colors (Fig. 6d). Co-expression analysis in petals and sepals revealed a significant positive correlation between *EpPAL2*, *EpPAL3*, *EpPAL6* and *EpPAL7* with *EpANSs* and *EpDFRs* in petals (Fig. 6f and g and Table S9), and between *EpPAL2* and *EpPAL3* with *EpANSs* and *EpDFRs* in sepals (Fig. 6g and Table S9). Based on these expression profiles of *EpPALs* and co-expression patterns, we speculate that *EpPAL2* and *EpPAL3* may primarily be involved in anthocyanin synthesis in petals and sepals, while *EpPAL6* and *EpPAL7* may also play a role in anthocyanin synthesis in petals.

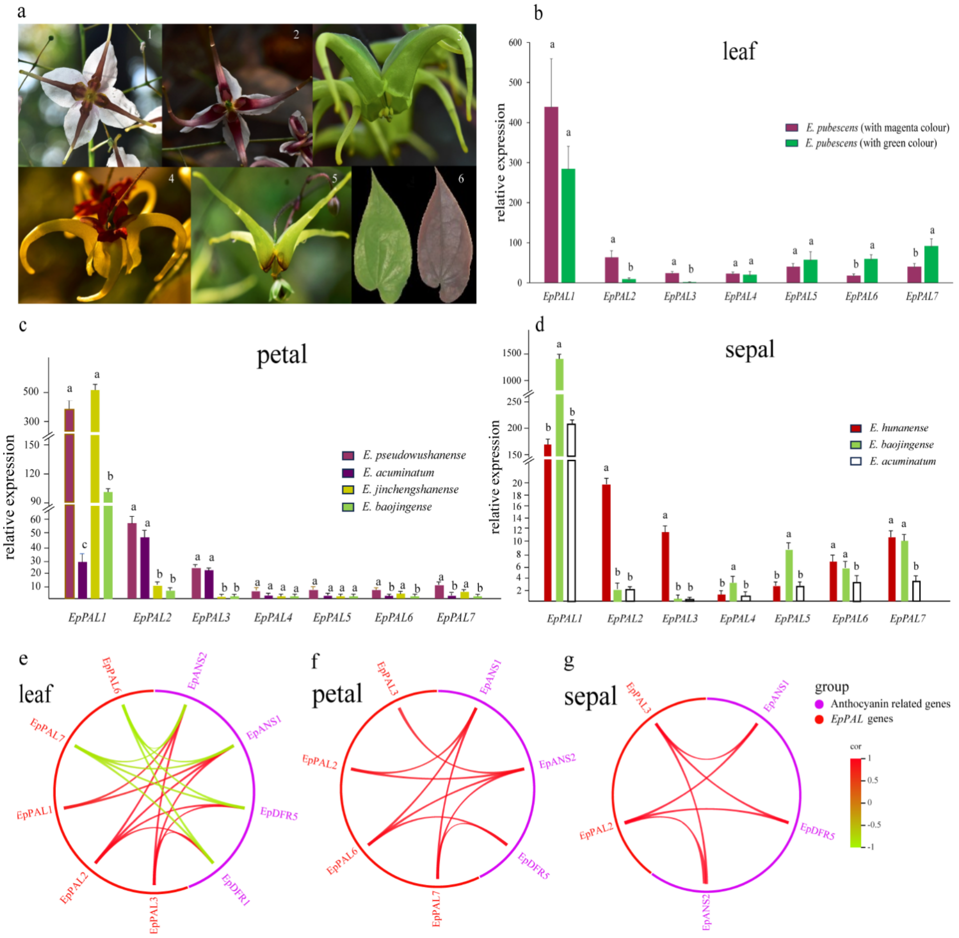


Fig. 6 Phenotype of flower and leaf colors in different species of *Epimedium* plants, their PAL expression profiles, and the co-expression relationship with major anthocyanin-related genes. (**a**) Different colors in leaf, petal and sepal. a1 ~ a5 showed *E. pseudowushanense*, *E. acuminatum*, *E. baojingense*, *E. hunanense* and *E. jinchengshanense*, respectively. a6 showed leaf colors in *E. pubescens* with green and magenta, respectively. (**b**) Expression levels of *EpPALs* with different leaf colors. Significance tests for each *EpPAL* gene in two types of leaves were conducted and were labled with italicized '*a*' and '*b*' accordingly; (**c**) Expression levels of *EpPALs* in *E. pseudowushanense*, *E. acuminatum*, *E. jinchengshanense* and *E. baojingense*, with petal colors of magenta, magenta, yellow, and green, respectively. (**d**) Expression levels of *EpPALs* in *E. hunanense*, *E. baojingense* and *E. acuminatum* with sepal colors of red, green and white, respectively. Significance tests for each *EpPAL* gene in different colours of petals or sepals were conducted and were labeled with regular 'a' and 'b' accordingly. (**e-g**) Co-expression relationship between *EpPALs* and anthocyanin-related genes in leaf, petal and sepal, respectively. Three biological replicates were provided, and each repeat represent a mixing sample originated from three individuals

Repeatability verification of PFGs content dynamics and expression levels

To further validate the casual PAL genes involved in PFGs biosynthesis, a parallel validation experiment was conducted, comprising the determination of Epimedin C and total PFGs using the UPLC method (Fig. 7c and Table S8), and the absolute quantification of seven *EpPALs* gene expressions across five developmental stages (S1 to S5) using the qRT-PCR method (Fig. 7b). Six pairs of specific primers were designed for this purpose, with a single pair used for *EpPAL6* and *EpPAL7* due to their high sequence conservation (Table S10).

The results revealed that *EpPAL2* and *EpPAL3* exhibited peak expression in S1, followed by a continuous decrease from S2 to S4, and a slight increase in S5 (Fig. 7b), consistent with the relative quantification observed in transcriptome results (Fig. 7a). A similar

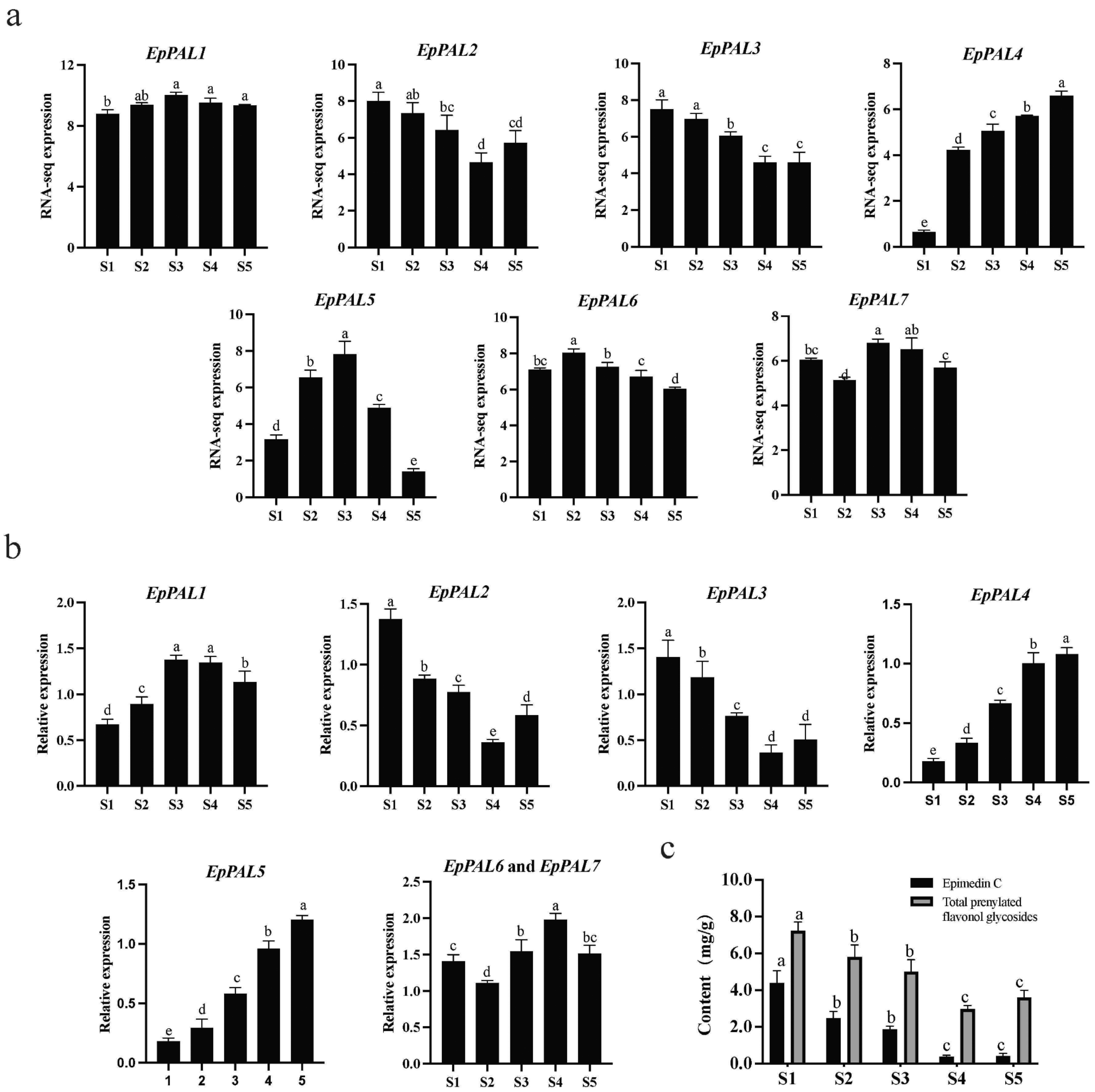


Fig. 7 Expression patterns of *EpPALs*. (**a**) Expression patterns of *EpPALs* in S1-S5 by RNA-seq. The replicates can be referred to Table S7. (**b**) Expression levels of *EpPALs* in S1-S5 by qRT-PCR. Three biological replicates were provided, and each repeat represent a mixing sample originated from three individuals. (**c**) The content of Epimedin C and total PFGs in S1-S5. S1-S5 represent the different stages of leaf development in *E. pubescens*. Error bars indicate standard error. The significance level is 0.05

trend was observed for other *EpPALs*, except for *EpPAL5*, indicating the overall reliability of the transcriptome data.

Importantly, the dynamic expression patterns of *EpPAL2* and *EpPAL3* across the five stages aligned with the dynamic changes in the content of Epimedin C and total PFGs (Fig. 7). However, no apparent correlation was found between the PFGs contents and qRT-PCR results for other *EpPALs*. For instance, *EpPAL1* showed an initial increase (S1 to S3) followed by a decrease (S3 to S5), while *EpPAL4* and *EpPAL5* exhibited a continuous increasing trend. *EpPAL6* and *EpPAL7* fluctuated significantly, with the lowest expression in S2 and the highest in S4, respectively (Fig. 7b), and did not align with the dynamic changes observed in Epimedin C and total PFGs content (Fig. 7c).

Overall, our findings suggest that *EpPAL2* and *EpPAL3* may be the most critical genes responsible for the biosynthesis of PFGs in *E. pubescens.*

Discussion

Characterization of the PAL gene family in *E. pubescens*

In the present study, the high-quality chromosome-level genome of *E. pubescens* served as a valuable resource for PAL research. A comprehensive exploration identified seven PAL genes (*EpPAL1* to *EpPAL7*), offering a more extensive analysis compared to previous studies in *E. sagittatum* [17]. All seven *EpPALs* were located in the cytoplasm, consistent with current PAL studies [8, 9, 12, 19]. Despite the widely recognized conservation among *EpPALs*, two genes, *EpPAL2* and *EpPAL3*, exhibited lower identity (61.09–64.38%) to other *EpPALs*, reminiscent of *ClPAL2* in *Citrullus lanatus* [20], which was proven to be an ancestral PAL, sharing only ~ 60% identity with other *ClPALs* despite having pairwise identities ranging from 71.2 to 99.0% [8]. Significant differences in PAL evolution with distinct origins were observed in *E. pubescens*, contributing to the varying sequence identities.

This study included a substantial proportion (20 out of 24) of primitive taxa which is different from previous studies [8, 19]. This facilitated an investigation into intron insertion events in *E. pubescens*. While similar exon/intron structures tend to cluster together, as observed in *E. sagittatum* [17], pear [21], and common walnut [22], *EpPAL2* and *EpPAL3* gained two introns (Fig. 2b and Figure S1), whereas other *EpPALs* possess only one intron insertion [8, 9, 11, 12, 17, 23, 24], suggesting significant structural divergence within the PAL gene family in *E. pubescens*.

Differences in intron/exon structures between *EpPAL2-3* and other angiosperms were also detected (Figure S1), further indicating their ancient with distinct evolutionary origins. The clustering of the ancient gene *NnPAL1* (*XP_010246007.1* in this study) with *EpPAL2* and *EpPAL3* supports this conclusion [19]. The increased intron number in ancient *EpPAL* genes may be significant, as intron-gain events can enhance mRNA stability or harbor regulatory elements without disrupting the coding frame of genes [25, 26–28]. However, intron insertions in *EpPAL2*, *EpPAL3* and other *EpPALs* may have occurred independently based on their phylogenetic positions (Fig. 1).

Evolution of the plant PAL gene family

Previous studies have primarily focused on research at the species or family level [20, 21, 23, 24, 29], with limited investigations on the evolution of PAL genes, except for those in *N. nucifera* [19] and cucurbit species [8]. To bridge this gap, we conducted a phylogenomic study to elucidate the evolution of plant PAL (Fig. 1). Our findings reveal a large-scale gene duplication event in gymnosperms, with two gymnosperm-specific branches (cluster2 and cluster 4), emphasizing their unique origin. The presence of PALs in gymnosperms across other branches (cluster3, cluster5 and cluster6) suggests ancestral gene duplication and vertical inheritance during evolution, supporting widespread differentiation [19]. We hypothesize that the multigene families of PAL in gymnosperms may confer diverse functions, such as producing additional trans-cinnamic acid for downstream metabolic pathways or enhancing lignin biosynthesis to defend against adverse environments like insect and pathogen attacks [25], contributing to their widespread habitat adaptation.

Furthermore, our results indicate that the origin of PAL in angiosperm plants may not be monophyletic. This is evident from the evolutionary tree, where PAL genes of dicots like cucumber and lotus are clustered within cluster6 and also share clustering with PAL genes of primitive types within cluster3 and cluster5 (Fig. 1). This finding validates previous conclusions in *N. nucifera* [19]. Except for *A. coerulea* (only in cluster6), the PALs of other early-diverging eudicotyledons including *E. pubescens* (in cluster3 and 6), *P. somniferum* (in cluster5 and 6), *M. integrifolia* (in cluster3 and 6), *K. uniflora* (in cluster5 and 6), *N. nucifera* (in cluster3, 5 and 6), and *T. sinense* (cluster5 and 6) are divided into two clusters or three clusters (*N. nucifera*), with one clustering with cluster3 or cluster5, speculated to be relatively ancient genes, and the other with PALs of modern dicots (cluster6). This suggests different evolutionary origins may underlie the PAL gene evolution in early-diverging dicots [19]. Additionally, we speculate that ancient genes like *EpPAL2* and *EpPAL3* may have played a crucial role in the survival of these “pioneer species” in harsh environments during evolution, highlighting their importance for related plants. For instance, in *E. pubescens*, the positive selection test indicates that *EpPAL2* and *EpPAL3* experienced the strictest selection pressure, evolving ~ 10 and ~ 54 times slower than genes of Clade 1_2 (*EpPAL4*~*7*) and Clade 1_1 (*EpPAL1*), respectively (Table S6).

Expression profiles of *EpPALs*

Duplicate genes can lead to pseudogenization, subfunctionalization and neofunctionalization [2]. Gene expression patterns offer valuable insights into the functional differentiation. Previous research has shown that *AtPAL1* and *AtPAL2* in *A. thaliana* are highly expressed in roots and involved in stress-induced flavonoid biosynthesis [30, 31]. In *Pyrus bretschneideri*, *PbPAL1* and *PbPAL2* are predominantly expressed in stems and roots, suggesting involvement in lignin synthesis and stone cell development [21].

In this study, the expression levels of *EpPAL2* and *EpPAL3* gradually decreased in correlation with the PFGs content during leaf development, indicating a potential significant role in PFGs biosynthesis. Our findings support the speculation that both genes positively correlate with Epimedin C and total PFGs content (Figs. 5 and 7). *EpPAL1* may also be involved in PFGs biosynthesis in leaves, as evidenced by the third-highest correlation with Epimedin C and total PFGs (Fig. 5). As an ancestral gene with an independent origin, *EpPAL1* is constitutively highly expressed across tissues, suggesting divergent functionality. Similar expression patterns of PAL can be observed in *Cuminum cyminum* and cucurbit plants [8, 32]. Given the pivotal role of PAL in both primary and secondary metabolism, constitutive multifunctional expression may be crucial for maintaining diverse biological processes and adapting to changing environments.

In *E. sagittatum*, the expression patterns of *EsPAL1*, *EsPAL2*, and *EsPAL3* align with lignification and active components accumulation [17]. In our study, *EpPALs* exhibited distinct yet overlapping expression patterns during different stages of leaf development (Fig. 5 and Table S7), hinting at possible functional diversification and redundancy. *EpPAL2* and *EpPAL3* demonstrated overlapping expression in PFGs and anthocyanin biosynthesis, ensuring functional redundancy, preventing the complete loss of weaker gene copies, a phenomenon often observed in key genes involved in different metabolic pathways [33, 34]. In contrast, *EpPAL1* and other *EpPALs* displayed distinct expression patterns, potentially stemming from functional differentiation among duplicate *EpPAL* genes, ultimately leading to their subfunctionalization and neofunctionalization.

In summary, this study comprehensively investigated PAL genes in *E. pubescens*. We speculate that *EpPAL2* and *EpPAL3* may participate in PFGs and anthocyanin pathways in leaves and flowers, while *EpPAL1*, characterized by its constitutively high expression, may not only be involved in the biosynthesis of PFGs and anthocyanins in leaves, but also play a significant role in defense and protection. *EpPAL6* and *EpPAL7* may participate in anthocyanin synthesis in petals, but there is no evidence of their involvement in PFGs biosynthesis. Given the low expression levels, we hypothesize that *EpPAL4* to *EpPAL7* may function as stress responders or have become nonfunctional following duplication events. Further experimental validation is needed to confirm these speculations.

Materials and methods

Experimental materials

The experimental materials were sourced from the Epimedium Germplasm Resources Nursery located in Xiuwen County, Guizhou Province. The plant species used in our studies are authenticated by Professor Baolin Guo. Specially, the plant selected for genome sequencing, as well as RNA-seq analysis across different tissues and leaf developmental stages, qRT-PCR and UPLC experimentation, was confirmed as *E. pubescens*. Additionally, Professor Baolin Guo identified the plants utilized in RNA-seq investigations of various flower hues, including *E. pseudowushanense*, *E. acuminatum*, *E. jinchengshanense*, *E. baojingense* and *E. hunanense*. Voucher specimens for all plant samples are preserved at the Plant Specimen Museum, part of the Institute of Medicinal Plant Resources Development, Chinese Academy of Medical Sciences (coded as IMD). The deposition numbers assigned to *E. pubescens*, *E. pseudowushanense*, *E. acuminatum*, *E. jinchengshanense*, *E. baojingense* and *E. hunanense* were B. L. Guo 0711-3, B. L. Guo 0312, B. L. Guo 0342, B. L. Guo 0524, B. L. Guo 0332 and B. L. Guo 0402, respectively.

PAL gene identification and sequence analysis

The genome sequences of *E. pubescens* were accessed from the National Center for Biotechnology Information (NCBI) under project PRJNA747870 [35]. Utilizing APG IV [36] as a reference, genome sequences of 23 species were located and retrieved from Phytozome (http://www.phytozome.net) and the NCBI database (https://www.ncbi.nlm.nih.gov/) (Table S1 and Table S2). The analyzed species are as follows: one algae (*Chara braunii*), one bryophyte (*Physcomitrella patens*), one fern (*Ceratopteris richardii*), two gymnosperms (*Ginkgo biloba* and *Sequoiadendron giganteum*), two basal angiosperms (*Amborella trichopoda* and *Nymphaea colorata*), four Magnoliidae (*Ceratophyllum demersum*, *Cinnamomum kanehirae*, *Liriodendron chinense*, *Piper nigrum*), a basal monocotyledon (*Spirodela polyrrhiza*), six early-diverging eudicotyledons (*Nelumbo nucifera*, *Macadamia integrifolia*, *Tetracentron sinense*, *Aquilegia coerulea*, *Kingdonia uniflora*, *Papaver somniferum*, *E. pubescens*), three typical dicotyledons (*Vitis vinifera*, *Cucumis sativus*, *Arabidopsis thaliana*), an early-diverging monocotyledons

(*Brachypodium distachyon*) and a typical monocotyledon (*Oryza sativa*). Predicted proteins from these genomes underwent screening with HMMER v3 [37], employing the Hidden Markov Model (HMM) corresponding to Pfam [38] (PF00221; http://pfam.sanger.ac.uk/). Among the proteins identified using the PAL HMM, a subset of high-quality proteins (E-value < 1e-20 and confirmed integrity of the PAL domain) was selected for alignment. In cases where only genome information was available for certain species, a localized TBLASTN search was conducted against the PAL genes of *A. thaliana* and *O. sativa*, considering records with maximum identity > 95%, length > 400 bp, and E-value < 1e-20. To validate the results of the HMM and BLAST searches, all potential PAL genes were further subjected to analysis in the NCBI-CDD database (https://www.ncbi.nlm.nih.gov/cdd/) to confirm the presence of conserved domains, and candidates lacking the "PAL-HAL" shorthand designation were discarded. Protein sequences were excluded if the PAL domain appeared truncated or if the PAL domain match E-value exceeded 1e-5. Following these stringent criteria, 167 PAL genes were ultimately identified across the nine species studied (Table S2). Sequences of the 7 *EpPALs* have been submitted to China National Center for Bioinformation (CNCB), with the accession numbers for *EpPAL1-EpPAL7* are C_AA071439.1, C_AA071440.1, C_AA071441.1, C_AA071442.1, C_AA071443.1, C_AA071444.1 and C_AA071445.1, respectively. The accession number of *Ep-actin* gene used in qPCR is C_AA071459.1.

Protein sequence properties analysis, conserved domain and motifs analysis

The physiological and biochemical characteristics of the full-length proteins were determined using the ProtParam tool (http://web.expasy.org/protparam/) [39]. SignalP (V.4.1) (http://www.cbs.dtu.dk/services/SignalP/) [40] and Euk-mPLoc (V.2.0) (http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/#) [41] were utilized to analyze the signal peptide and subcellular localization of each protein, respectively. Additionally, MEME (V.5.0.2) (http://meme-suite.org/) [42] was employed to identify conserved motifs, including the PAL domain, using optimized parameters: a maximum of 10 motifs were searched for, with each motif ranging from 6 to 50 residues in width.

Phylogenetic analysis, synteny block identification and gene duplication pattern analysis

The protein sequences were aligned using ClustalW2 [43] with its default settings. Phylogenetic trees were inferred using the maximum likelihood (ML) method with the JTT + R9 model, which was automatic selected by IQ-TREE [44]. The evolutionary tree was then visualized and further refined using iTOL (https://itol.embl.de/) [45]. Synteny blocks between genomes and intra-specific collinearity analysis were identified using the jcvi pipeline (https://github.com/tanghaibao/jcvi). BLASTP was performed to identify paralogous or orthologous gene pairs, with an E-value cutoff of 1e-05. To identify patterns of gene duplication, DupGen-finder (https://github.com/qiao-xin/DupGen_finder) [46] was employed.

Chromosome location, *cis*-acting element and gene structure analysis

The gene location map was constructed using MapChart V.2.0 (http://mg2c.iask.in/mg2c_v2.0/) [47]. *Cis*-acting elements located within the 1.5 kb upstream sequences of the 5′ regulatory region, starting from the transcriptional start site, were identified using PlantCARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/). To assess the divergence between upstream sequences of each paralogous gene pair, the GATA program [48] was employed with a window size of seven and a lower cut-off score of 12 bits. Lastly, the visualization of gene structure was facilitated by TBtools software [49].

Natural selection test

Codeml program in PAML (V.4.8) [50] was conducted to detect changes in evolutionary rates and signatures of positive selection. Four levels of positive selection tests were performed. (1) Detection of positive selection in pairwise genes of all *EpPALs*. For this, the main parameter settings were: runmode = -2 and NSsites = 0; (2) Site-specific model was applied for positive selection detection of sites in genes. This model assumes a constant ω (ω = dN/dS; where dN is the non-synonymous substitution rate and dS is the synonymous substitution rate) across all branches. The main parameters were set as runmode = 0 and NSites = 0 1 2 7 8. To determine the most suitable model for detection, we compared Neutral M1 vs. Selection M1 and M7 vs. M8; (3) Branch model was applied to detect the rapidly evolving genes in the target branch. This model assumes a constant ω for all sites with a gene. Three scenarios were tested: the one-ratio model (assuming a constant ω for all branches with parameters as model = 0 and NSites = 0), the two-ratio model (assuming a foreground ω for designated branches and a background ω for all others with parameters as model = 2 and Nsites = 0), and the free-ratio model (allowing different ω for each branch with parameters as model = 1 and Nsites = 0) [51]. Models were compared using likelihood ratio tests based on the log likelihood (lnL). The chi-square test with a significance threshold of $P < 0.05$ was used to compare 2|ΔlnL| values between models; (4) Branch-site model was applied to detect whether there exist positive selection sites in a specific branch. This model assumes one ω for the target branch and another constant ω for all other branches. We compared the branch-site model A (model = 2, NSites = 2, fix_omega = 0, omega = 2) with its null model (model = 2, NSites = 2, fix_omega = 1, omega = 1). If the chi-square test yielded a significance of $P < 0.05$, we employed the Bayes Empirical Bayes (BEB) method to calculate the posterior probability. Genes in the specific branch were considered under positive selection if this value exceeded 0.95 [52].

RNA-seq and correlation analysis

To identify the gene expression profiles of *EpPALs*, we conducted three independent RNA-seq experiments. Firstly, we sampled different tissues (roots, stems, leaves, flowers, and fruits) from *E. pubescens*. Secondly, we collected leaves from various developmental stages of *E. pubescens*, specifically: Stage 1 (S1) with leaf width of 0.5–1 cm and low leatheriness; Stage 2 (S2) with leaf width of 1.5–2 cm and low leatheriness; Stage 3 (S3) with leaf width of 2.5 ~ 4 cm and low leatheriness; Stage 4 (S4) with leaf width of 5 cm and medium leatheriness; and Stage 5 (S5) with leaf width of 5 cm and high leatheriness. Thirdly, we included six species of *Epimedium* with diverse petal colors (magenta in *E. pseudowushanense* and *E. acuminatum*, yellow in *E. jinchengshanense*, and green in *E. baojingense*), sepal colors (red in *E. hunanense*, green in *E. baojingense*, and white in *E. acuminatum*), and leaf colors (green and magenta in *E. pubescens*). The RNA-seq protocol and classification criteria followed Xu et al. (2023) [53]. Initial protein contamination screening was performed using the NanoDrop ND 1000 (Nanodrop technologies), ensuring a tightly controlled OD260/OD280 ratio within the range of 1.9 to 2.1. Subsequently, the RNA Integrity Number (RIN) was evaluated using the Agilent Technologies 2100 bioanalyzer (Agilent, Santa Clara, CA). Sequencing was only initiated if the RIN exceeded 8 and the 28 S/18S ratio was greater than or equal to 0.7. Software tools including Trimmomatic (version 0.36) [54], HISAT2 [55], and the R package Rsubread [56] were utilized for quality control, sequence alignments, and gene expression quantification, respectively. The reference genome utilized was that of *E. pubescens*, as published by Shen et al. (2022) [35]. All samples were collected between 10:00–11:30 am on a sunny day and immediately treated with liquid nitrogen before being stored in dry ice for transport to Beijing. All samples were conserved at -80 °C under ultralow temperature for subsequent RNA extraction and chemical component identification. We used the R packages Tidyverse and ggcor to compute the pearson correlation between PALs and the relative content of PFGs.

UPLC experiment

For the analysis of PFGs content, approximately 0.1 g of ground sample was soaked in 10 ml of 50% ethanol and ultrasonicated for 30 min before being filtrated through 0.22 μm filter membrane (Millipore, Nylon) for UPLC analysis. UPLC under 270 nm was conducted at a flow rate of 0.3 ml/min using the ACQUITY UPLC system (UPLC I-class; Waters, Milford, MA, USA) equipped with an ACQUITY UPLC BEH C18 column (2.1 × 100 mm. 1.7 μm; Waters, Milford, MA, USA) maintained at 25 °C. The mobile phase comprised of water (eluent A) and 100% acetonitrile (eluent B). The authentic flavonoids were purchased from the Shanghai Yuanye Bio-Technology Co., Ltd., Shanghai, China.

qRT-qPCR

To ensure the reliability of our transcriptome data, we conducted qRT-PCR analysis on leaves from five distinct developmental stages of *E. pubescens*, focusing on six selected *EpPALs*. Total RNA was extracted using a plant total RNA extraction kit from Aidlab (China). We assessed RNA integrity on a 1.2% agarose gel and quantified it using a NanoDrop 2000 C Spectrophotometer from Thermo Scientific (USA). cDNA synthesis was achieved using the TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix Kit from Transgen Biotech (China). qRT-PCR reactions were performed for each tissue sample with gene-specific primers (Table S10). The qRT-PCR program consisted of pre-denaturation at 95 °C for 2 min, followed by 40 cycles of amplification at 95 °C for 15 s, 60 °C for 30 s, and 72 °C for 30 s. We analyzed the relative abundance of transcripts using the comparative Ct method, applying the formula 2-ΔΔCt for relative quantification. Our gene expression results were calculated based on the 2-ΔΔCt method, and the reported data represents the average of three biological and three technical replicates.

Conclusions

7 PALs were firstly and comprehensively identified based on the genome of *E. pubescens*. *EpPAL2*, *EpPAL3* and *EpPAL1* were identified as the ancient isoforms. *EpPAL2* and *EpPAL3* exhibited a homology range of 61.09 to 64.38%, contained two introns and underwent strong purifying selection, evolving at a rate ~ 10 to ~ 54 times slower compared to *EpPAL1* and modern *EpPALs* (*EpPAL4-7*). The evolutionary trajectory of modern *EpPALs* was shaped by multiple duplication events. Initially, *EpPAL4* emerged through intraspecific whole-genome duplication from *EpPAL1*. This was followed by a sequence of tandem duplications resulting in *EpPAL5* and *EpPAL6*, and transposed duplications that gave rise to *EpPAL7*, all originating from *EpPAL4*. Analysis of expression profiles through RNA-seq and UPLC

techniques revealed that *EpPAL2* and *EpPAL3* are key genes involved in the biosynthesis of prenylated flavonol glycosides. This finding was further validated through parallel UPLC and qRT-PCR experiments. Novel insights into the evolution of 24 PAL gene families were provided, revealing the evolutionary characteristics of 12 different evolutionary clade groups. Overall, this study offers a unique perspective on PAL evolution and clarifies the role of PAL genes in *Epimedium* plants.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05480-z.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements

We thank all our colleagues for providing useful discussions and technical assistance. We are very grateful to the editor and reviewers for critically evaluating the manuscript and providing constructive comments for its improvement.

Author contributions

Chaoqun Xu conceived and designed the study, put into effect the main bioinformatics analyses, wrote the manuscript, and prepared the figures and tables. Xuelan Fan prepared the materials, conducted the experiments, data analysis, and revised the manuscript drafts. Guoan Shen participated in the design of this study and revised the manuscript. Baolin Guo conceived and designed the study, involved in data interpretation and finalizing the manuscript draft. The authors read and approved the final manuscript.

Funding

This work was financially supported by the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-031).

Data availability

The experimental materials were stored at the Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences. The datasets containing the E. pubescens genome sequences can be accessed from the National Center for Biotechnology Information (NCBI) repository using the accession number PRJNA747870. Additionally, the RNA-seq datasets generated in this study have been deposited in the China National Center for Bioinformation (CNCB) repository. Specifically, the RNA-seq data related to different tissues of E. pubescens, various developmental stages of E. pubescens leaves, and six species of Epimedium with distinct petal colors can be retrieved using the accession numbers CRA014527, CRA014549, and CRA014550, respectively.

Declarations

Ethics approval and consent to participate

All the experimental materials in this study have obtained the authority. The experimental research and method on all the experimental materials, including the collection of plant material, comply with relevant institutional, national, and international guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Key Laboratory of Bioactive Substances and Resources Utilization of Chinese Herbal Medicines, Ministry of Education, Institute of Medicinal Plant Development, Peking Union Medical College and Chinese Academy of Medical Sciences, No.151 MaLianWa North Road, Haidian District, Beijing 100193, China
[2]College of Pharmacy, Jiangxi University of Chinese Medicine, Nanchang 330004, China

Received: 9 January 2024 / Accepted: 1 August 2024
Published online: 04 September 2024

References

1. Naoumkina M, Zhao Q, Gallego-Giraldo L, Dai X, Zhao PX, Dixon R. Genome-wide analysis of phenylpropanoid defence pathways. Mol Plant Pathol. 2010;11(6):829–46.
2. Barros J, Dixon RA. Plant phenylalanine/tyrosine ammonia-lyases. Trends Plant Sci. 2020;25(1):66–79.
3. Bennici A. Origin and early evolution of land plants: problems and considerations. Commun Integr Biol. 2008;1(2):212–8.
4. Shalaby S, Horwitz BA. Plant phenolic compounds and oxidative stress: integrated signals in fungal-plant interactions. Curr Genet. 2015;61(3):347–57.
5. Liu CW, Murray JD. The role of flavonoids in nodulation host-range specificity: an update. Plants (Basel). 2016;5(3):33.
6. MacDonald MJ, D'Cunha GB. A modern view of phenylalanine ammonia lyase. Biochem Cell Biol. 2007;85(3):273–82.
7. Schwede TF, Rétey J, Schulz GE. Crystal structure of histidine ammonia-lyase revealing a novel polypeptide modification as the catalytic electrophile. Biochemistry. 1999;38(17):5355–61.
8. Dong C, Cao N, Zhang Z, Shang Q. Phenylalanine ammonia-lyase gene families in cucurbit species: structure, evolution, and expression. J Integr Agric. 2016;15(6):1239–55.
9. Chang A, Lim MH, Lee SW, Robb EJ, Nazar RN. Tomato phenylalanine ammonia-lyase gene family, highly redundant but strongly underutilized. J Biol Chem. 2008;283(48):33591–601.
10. Bate NJ, Orr J, Ni W, Meromi A, Nadler-Hassar T, Doerner PW, et al. Quantitative relationship between phenylalanine ammonia-lyase levels and phenylpropanoid accumulation in transgenic tobacco identifies a rate-determining step in natural product synthesis. J Clin Periodontol. 1994;91(16):7608–12.
11. Huang J, Gu M, Lai Z, Fan B, Shi K, Zhou YH, et al. Functional analysis of the arabidopsis PAL gene family in plant growth, development, and response to environmental stress. Plant Physiol. 2010;153(4):1526–38.
12. de Jong F, Hanley SJ, Beale MH, Karp A. Characterisation of the willow phenylalanine ammonia-lyase (PAL) gene family reveals expression differences compared with poplar. Phytochemistry. 2015;117:90–7.
13. Ma H, He X, Yang Y, Li M, Hao D, Jia Z. The genus *Epimedium*: an ethnopharmacological and phytochemical review. J Ethnopharmacol. 2011;134(3):519–41.
14. Jiang J, Zhao Bj, Song J, Jia X. Pharmacology and clinical application of plants in *Epimedium* L. Chin Herb Med. 2016;8(1):12–23.
15. Zhu Jf L, Zj Z, Gs, Meng K, Kuang Wy, Li J, et al. Icaritin shows potent anti-leukemia activity on chronic myeloid leukemia *in vitro* and *in vivo* by regulating MAPK/ERK/JNK and JAK2/STAT3 /AKT signalings. PLoS ONE. 2011;6(8):e23720.
16. Zhao H, Guo Y, Li S, Han R, Ying J, Zhu H, et al. A novel anti-cancer agent Icaritin suppresses hepatocellular carcinoma initiation and malignant growth through the IL-6/Jak2/Stat3 pathway. Oncotarget. 2015;6(31):31927–43.
17. Zeng S, Liu Y, Zou C, Huang W, Wang Y. Cloning and characterization of phenylalanine ammonia-lyase in medicinal *Epimedium* species. Plant Cell Tissue Organ Cult. 2013;113(2):257–67.
18. Liu Y, Wu L, Deng Z, Yu Y. Two putative parallel pathways for naringenin biosynthesis in *Epimedium wushanense*. RSC Adv. 2021;11(23):13919–27.
19. Wu Z, Gui S, Wang S, Ding Y. Molecular evolution and functional characterisation of an ancient phenylalanine ammonia-lyase gene (NnPAL1) from *Nelumbo nucifera*: novel insight into the evolution of the PAL family in angiosperms. BMC Evol Biol. 2014;14:100.

20. Dong CJ, Shang QM. Genome-wide characterization of phenylalanine ammonia-lyase gene family in watermelon (*Citrullus lanatus*). Planta. 2013;238(1):35–49.
21. Li G, Wang H, Cheng X, Su X, Zhao Y, Jiang T, et al. Comparative genomic analysis of the *PAL* genes in five Rosaceae species and functional identification of Chinese white pear. PeerJ. 2019;7:e8064.
22. Yan F, Li H, Zhao P. Genome-wide identification and transcriptional expression of the PAL gene family in common walnut (*Juglans Regia* L). Genes. 2019;10(1):46.
23. Hou X, Shao F, Ma Y, Lu S. The phenylalanine ammonia-lyase gene family in *Salvia miltiorrhiza*: genome-wide characterization, molecular cloning and expression analysis. Mol Biol Rep. 2013;40(7):4301–10.
24. Thiyagarajan K, Vitali F, Tolaini V, Galeffi P, Cantale C, Vikram P, et al. Genomic characterization of phenylalanine ammonia lyase gene in buckwheat. PLoS ONE. 2016;11(3):e0151187.
25. Bagal UR, Leebens-Mack JH, Lorenz WW, Dean JFD. The phenylalanine ammonia lyase (PAL) gene family shows a gymnosperm-specific lineage. BMC Genomics. 2012;13(3):S1.
26. Duret L. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet. 2001;17(4):172–5.
27. Wang HF, Feng L, Niu DK. Relationship between mRNA stability and intron presence. Biochem Biophys Res Commun. 2007;354(1):203–8.
28. Vogt T. Phenylpropanoid biosynthesis. Mol Plant. 2010;3(1):2–20.
29. Hu GS, Jia JM, Hur YJ, Chung YS, Lee JH, Yun DJ, et al. Molecular characterization of phenylalanine ammonia lyase gene from *Cistanche deserticola*. Mol Biol Rep. 2011;38(6):3741–50.
30. Wanner LA, Li G, Ware D, Somssich IE, Davis KR. The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana*. Plant Mol Biol. 1995;27(2):327–38.
31. Olsen KM, Lea US, Slimestad R, Verheul M, Lillo C. Differential expression of four *Arabidopsis PAL* genes; *PAL1* and *PAL2* have functional specialization in abiotic environmental-triggered flavonoid synthesis. J Plant Physiol. 2008;165(14):1491–9.
32. Habibollahi M, Kavousi HR, Lohrasbi-Nejad A, Rahpeyma SA. Cloning, characterization and expression of a phenylalanine ammonia-lyase gene (*CcPAL*) from cumin (*Cuminum cyminum* L). J Appl Res Med Aromat Plants. 2020;18:100253.
33. Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, et al. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Mol Biol Evol. 2006;23(2):469–78.
34. Lei L, Zhou SL, Ma H, Zhang LS. Expansion and diversification of the *SET* domain gene family following whole-genome duplications in *Populus trichocarpa*. BMC Evol Biol. 2012;12:51.
35. Shen G, Luo Y, Yao Y, Meng G, Zhang Y, Wang Y, et al. The discovery of a key prenyltransferase gene assisted by a chromosome-level *Epimedium pubescens* genome. Front Plant Sci. 2022;13:1034943.
36. Group TAP, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 2016;181(1):1–20.
37. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39(Web Server issue):W29–37.
38. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222–230.
39. Wang L, Wang L, Zhang Z, Ma M, Wang R, Qian M, et al. Genome-wide identification and comparative analysis of the superoxide dismutase gene family in pear and their functions during fruit ripening. Postharvest Biol Technol. 2018;143:68–77.
40. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6.
41. Chou KC, Shen HB. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE. 2010;5(4):e9931.
42. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39–49.
43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8.
44. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
45. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256–9.
46. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol. 2019;20(1):38.
47. Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Heredity. 2002;93(1):77–8.
48. Nix D, Eisen M. GATA: a graphic alignment tool for comparative sequence analysis. BMC Bioinformatics. 2005;6:9.
49. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13(8):1194–202.
50. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91.
51. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15(5):568–73.
52. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19(6):908–17.
53. Xu C, Liu X, Shen G, Fan X, Zhang Y, Sun C, et al. Time-series transcriptome provides insights into the gene regulation network involved in the icariin-flavonoid metabolism during the leaf development of *Epimedium pubescens*. Front Plant Sci. 2023;14:1183481.
54. Anthony M, Marc L, Bjoern U. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
55. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
56. Liao Y, Smyth GK, Shi W. The r package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nuc Acids Res. 2019;47(8):e47.

Publisher's Note