Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR

Nancy Choudhary[1] and Boas Pucker[1,*]

[1] Plant Biotechnology and Bioinformatics, Institute of Plant Biology & BRICS, TU Braunschweig, Braunschweig, Germany

* corresponding author: b.pucker@tu-braunschweig.de

Abstract

Flavonoids, an important class of specialized metabolites, are synthesized from phenylalanine and present in almost all plant species. Different branches of flavonoid biosynthesis lead to products like flavones, flavonols, anthocyanins, and proanthocyanidins. Dihydroflavonols form the branching point towards the production of non-coloured flavonols via flavonol synthase (FLS) and coloured anthocyanins via dihydroflavonol 4-reductase (DFR). An angiosperm-wide comparison of FLS and DFR sequences revealed the amino acids at positions associated with the substrate specificity in both enzymes. A global analysis of the phylogenetic distribution of these amino acid residues revealed that monocots generally possess FLS with Y132 (FLS_Y) and DFR with N133 (DFR_N). In contrast, dicots generally possess FLS_H and DFR_N, DFR_D and DFR_A. DFR_A, which restricts substrate preference to dihydrokaempferol, previously believed to be unique to strawberry species, is found to be more widespread in angiosperms and has evolved independently multiple times. Generally, angiosperm FLS appears to prefer dihydrokaempferol, whereas DFR appears to favor dihydroquercetin or dihydromyricetin. Moreover, in the FLS-DFR competition, dominance of one over the other is observed, with typically only one expressed at

any given time. This study illustrates how substrate-specifying residues could mitigate competition between FLS and DFR, delineates the evolution of these enzymes, and provides insights into mechanisms directing the metabolic flux of the flavonoid biosynthesis, with potential implications for ornamental plants and molecular breeding strategies.

Keywords: Flavonoid biosynthesis, substrate competition, flavonol synthase, dihydroflavonol 4-reductase, substrate specificity

Background

Flavonoids comprise a range of different subclasses including flavones, flavonols, anthocyanins and proanthocyanidins, all of which have their own biological functions (1). Flavonoids exhibit a remarkable diversity, encompassing a vast array of compounds that possess distinct properties, conferring specific evolutionary benefits (2). The biosynthesis of these compounds involves a complex network of enzymatic reactions. Mutants with observable phenotypic variations in flavonoid biosynthesis have served as valuable model systems for studying the intricacies of biosynthetic pathways and their regulatory mechanisms (3–6).

Anthocyanins are renowned for their role as red, purple, or blue pigments, contributing to the vibrant colors seen in flowers and fruits (1). These color patterns serve to attract animals for important biological processes such as pollination and seed dispersal (7). Flavonols have been described as co-pigments that contribute to the overall coloration of various plant structures (8,9). On their own, flavonols are colorless or pale yellow and have been described as important factors in the UV stress response (10). Additionally, the flavonoid biosynthesis pathway produces other compounds such as flavones and proanthocyanidins (11,12). Flavones play a crucial role in signalling and pathogen defense (11), whereas proanthocyanidins are key contributors to seed coat pigmentation (13). Beyond their pigmentation functions, flavonoids

have diverse roles in response to environmental factors like drought (14), cold (15–17), and highlight (18,19), playing a vital physiological role in protecting against UV-induced oxidative stress (20,21), detoxifying reactive oxygen species (ROS) (22,23) and toxic metals (24), defending against pathogens (25), regulating symbiotic associations (26,27), and influencing auxin flux (28,29).

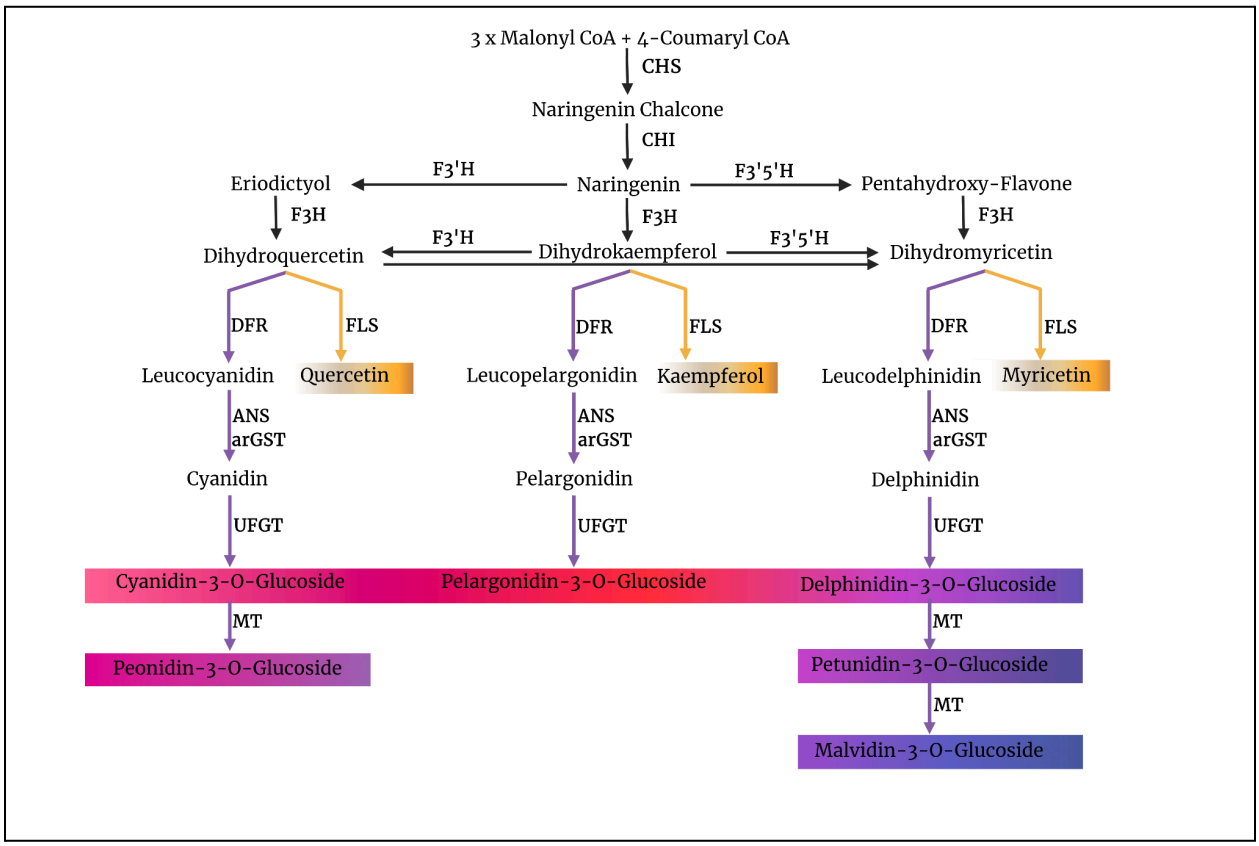

Figure 1: Simplified illustration of the flavonoid biosynthesis pathway in plants showing the competition between FLS and DFR for common substrate, dihydroflavonols. CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3′H, flavonoid 3′-hydroxylase; F3′5′H, flavonoid 3′,5′-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; arGST, anthocyanin-related glutathione S-transferase; UFGT, UDP-glucose flavonoid 3-O-glucosyl transferase; MT, methyltransferase.

A range of different colors is provided by three classes of anthocyanins: pelargonin (orange to bright red), cyanin (red to pink), and delphinin (purple or blue) (1). These three classes are characterized by different numbers of hydroxyl groups at the B phenyl ring: pelargonin (one), cyanin (two), and delphinin (three) (30). The biosynthesis of anthocyanins involves the enzymatic conversion of colorless dihydroflavonols into leucoanthocyanidins by DFR (31). ANS further converts leucoanthocyanidins into anthocyanidins. Recent research suggests the involvement of a glutathione S-transferase in anthocyanin biosynthesis (32,33). Anthocyanidins are transformed into anthocyanins through the addition of a sugar moiety by glycosyl transferases (34). Subsequent modifications such as the addition of other sugar moieties (35) or acyl groups (36) are catalyzed by a large number of often promiscuous enzymes.

Dihydroflavonols are not exclusively channeled into anthocyanin biosynthesis but are also the substrate of the flavonol synthase (FLS) that produces kaempferol, quercetin, or myricetin depending on the available dihydroflavonol (37). This leads to a competition for the dihydroflavonol substrates between DFR and FLS (**Figure 1**). Previous studies suggest that the ratio of FLS to DFR activity is an important factor in color determination (37–41). A study across multiple flowering plants discovered that high FLS activity can result in a lack of anthocyanin pigmentation despite a functional anthocyanin biosynthesis machinery being present (40). The flavonoid 3'-hydroxylase (F3'H) converts dihydrokaempferol into dihydroquercetin by adding one hydroxyl group in the C3' position of the B ring, while the flavonoid 3',5'-hydroxylase (F3'5'H) converts dihydrokaempferol into dihydromyricetin by adding hydroxyl groups at the C3' and C5' positions (42). Differences in the stability of the different dihydroflavonols pose a challenge for quantitative *in vitro* studies (43).

The activity of different branches of the flavonoid biosynthesis is controlled by transcriptional activation of the involved genes. *FLS* gene expression is activated by MYB transcription factors

of the subgroup7 and this mechanism appears to be generally conserved across plant species with minor lineage-specific variations (44,45). Recently, a pollen-specific activation of *FLS* by sugroup19 MYBs was described (46,47). Transcriptional activation of DFR is controlled by an MBW transcription factor complex that is named after the three components: MYB, bHLH, and WD40 (48). While research on *A. thaliana* mutants suggests that positive regulation dominates the gene expression control of flavonoid biosynthesis genes, several negative regulators have been identified through research on many different plant species (49).

Apart from the overall activity of FLS and DFR, the substrate selectivity plays a crucial role in determining the production of flavonols and anthocyanins (50,51). This selectivity is evident in species like *Freesia hybrida*, in which the presence of abundant kaempferol glycosides and delphinidin derivatives (52) indicates the separation of flavonol and anthocyanin biosynthesis based on the contrasting substrate preferences of DFR and FLS (52). Despite testing various DFR copies in *Freesia hybrida*, none of them exhibited a preference for DHK, while all favored DHM as the substrate (52). Some DFRs have been found to accept all three potential dihydroflavonols, indicating that the availability of specific substrates determines the class of anthocyanins produced. Other DFRs are incapable of accepting dihydrokaempferol as a substrate, resulting in the absence of orange anthocyanin-based pigmentation in species like petunia, apple, pear, hawthorn, rose, and cranberry (43,51,53). The substrate specificity determining region was predicted by a multiple alignment of polypeptide sequences (54). This region was further investigated based on the *Gerbera hybrida* DFR which is able to accept DHK and leads to orange flower pigmentation (55). Two important amino acid residues implicated in the substrate specificity of DFR were identified based on chimeric *Gerbera hybrida* and *Petunia hybrida* DFRs followed by site-directed mutation experiments (51). The relevance of all positions in a region of 26 amino acid residues was evaluated by permutation experiments. The mutation E145L (of the *G. hybrida* DFR resulted in a non-functional enzyme while the N134L mutation

restricted the substrate preference to dihydrokaempferol and excluded dihydroquercetin or dihydromyrecitin (51). Many following studies confirmed the particular relevance of this N134 residue (position 133 in *Arabidopsis thaliana* DFR) and identified additional amino acid residues with specific properties (50,56–58). In summary, N allows the acceptance of all dihydroflavonols, L restricts the substrate to DHK, D enables acceptance of DHQ and DHM and A leads to a high DHK affinity (**Figure 2**).

While the substrate preference of DFR received substantial attention, little is known about the substrate preferences of FLS. Five specific residues, namely H132, F134, K202, F293, and E295, have been reported to be involved in the binding of the DHQ substrate (59). Previous research in *A. thaliana* studied the impact of amino acid substitutions in AtFLS through incubation with DHQ. The mutants, H132F and H132Y exhibited 124% and 83% activity, respectively, compared to the wild-type (WT) (59). The K_M values of mutants H132F and H132Y displayed an approximately 2-fold decrease compared to the WT counterpart (59). Conversely, mutations in the other four residues led to a significant decrease in FLS activity. Interestingly, the four highly conserved residues were found to be crucial for substrate binding and catalytic activity. The H132 residue was found to be variable across different flavonol synthases, with some having F residues and others having H residues. These findings suggested that the H132 position may play a role in determining substrate preferences among different flavonol synthases (59). It is postulated that the position of H132 interacts via a hydrogen bond with the B-ring hydroxyl group of the dihydroflavonol substrates (60). Given the previous reports about the relevance of N133 in DFR and H132 in FLS, we will refer to them as substrate preference determining residues throughout this manuscript, while not ruling out the contribution of additional amino acid residues.

While the model organism *A. thaliana* has only a single DFR gene (61), multiple DFR copies have been reported in many other plant species. For example, studies in buckwheat (62) and strawberry (63) have identified DFR copies that show substantial differences in their substrate preference and consequently also in the produced anthocyanin classes. Genetic analysis of *DFR1* and *DFR2* in buckwheat suggested that both genes are genetically linked, but the genetic distance does not suggest a recent tandem duplication (62). Buckwheat *DFR2* differs from *DFR1* by an asparagine to valine substitution at a position that corresponds to 134 in the *Gerbera hybrida* sequence and the presence of an additional glycine between positions 138 and 139. Enzyme assays suggest that *DFR1* prefers DHK and DHM, while *DFR2* prefers DHQ. *DFR1* is expressed in most tissues, while *DFR2* expression is restricted to seeds and roots. Strawberry DFR1 has a high preference for DHK and does not accept DHQ and DHM ('DHK specialist'), while DFR2 does not accept DHK as a substrate, but works on DHQ and DHM ('DHK rejector') (63).

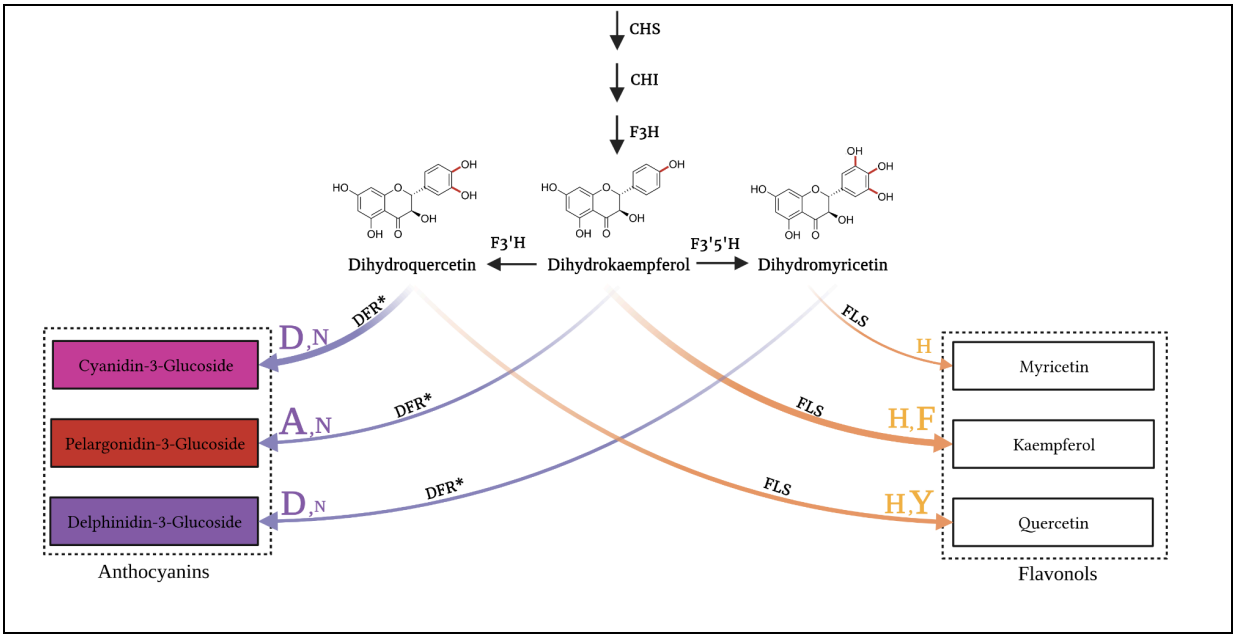

Figure 2: Illustration of the branching point in the flavonoid biosynthesis pathway leading to anthocyanins and flavonols, respectively. Asterisk (*) shows that the reactions from dihydroflavonols to anthocyanins also required the additional enzymes anthocyanidin synthase (ANS), anthocyanin-related glutathione

S-transferase (arGST) and UDP-dependent anthocyanidin 3-O-glucosyltransferase (UGT) downstream of DFR. DFR and FLS can process substrates with different hydroxylation patterns leading to three distinct products each. The characters at the arrows represent the previously reported substrate-specificity determining residues of FLS and DFR. The DFR residues, represented as purple characters depict the 133 position while the FLS residues, represented as yellow characters depict the 132 position in FLS (position according to *AtDFR* and *AtFLS1*).

The objective of this study was to reveal mechanisms that control and mitigate the substrate competition of FLS and DFR. A systematic analysis of substrate specificity determining amino acid residues in the polypeptide sequences of FLS and DFR revealed differences between monocots and dicots. The results were connected to a cross-species gene expression analysis of relative *F3'H* activity which serves as a proxy for the relative availability of DHK and DHQ, respectively.

Methods

Analyzed species and data sets

The coding sequences for 211 plant species were retrieved from Phytozome (64) and NCBI/GenBank (65) (Additional File 1). The selection of plant species encompassed a diverse taxonomic range, including 20 Thallophyta (algae and fungi), 4 Bryophyta (mosses), 1 Pteridophyta (ferns and fern allies), 11 Gymnosperms (naked seed plants), and 175 Angiosperms (flowering plants). A Python script (66) was used to translate the coding sequences into corresponding peptide sequences.

Flavonoid biosynthesis gene identification

The polypeptide sequences were analyzed with KIPEs3 v0.35 (67,68) to detect the presence of flavonoid biosynthesis genes in each plant species. KIPEs3 itself supplied the bait sequences and essential residue information required for the analysis (flavonoid biosynthesis dataset v3); leading to the identification of candidate genes for *FLS* (*flavonol synthase*) and *DFR* (*dihydroflavonol 4-reductase*). In addition, flavanone 3-hydroxylase (*F3H)*, flavonoid 3'-hydroxylase (*F3'H)*, and flavonoid 3'5'-hydroxylase (*F3'5'H)* genes were identified in the investigated plant species. KIPEs3 also enabled the identification of *FLS*-like and *DFR*-like genes within the datasets.

Multiple sequence alignment and phylogenetic tree construction

The DFR and DFR-like polypeptide sequences and the FLS and FLS-like polypeptide sequences were aligned separately using MAFFTv7.310 (69) with default penalties for gaps and the protein weight matrix of BLOSUM62 (70). The alignments were generated using the L-INS-i accuracy-oriented method and incorporated local pairwise alignment. The amino acid alignments were translated back to codon alignments using pxaa2cdn (71). The alignments were cleaned using pxclsq (71) to remove alignment columns with very low occupancy (<0.1). Multiple ANR and CCR sequences retrieved from GenBank were used as an outgroup for DFR tree construction. Previously characterized DFRs were also included. Multiple F3H and ANS sequences obtained from GenBank were used as an outgroup for FLS tree construction to facilitate a comprehensive investigation in the context of the 2-oxoglutarate-dependent dioxygenase (2-ODD) family. Previously described FLS sequences were also included. The best model of evolution was inferred using ModelFinder (72). Maximum likelihood trees were constructed for deciphering *DFR* and *FLS* evolution using IQ-TREE v2.0.3 (73) with 1,000

bootstrap replicates and the GTR+F+ASC+R10 nucleotide model. Additional phylogenetic trees were constructed based on MAFFTv7.310 and Muscle v5.1 (super5 algorithm) with IQ-TREE v2.0.3, FastTree2.1.10 (74) using the GTR+CAT model, and MEGA v11 (75) using the neighbor-joining method, 1,000 bootstrap replicates and the Maximum Composite Likelihood model. The topologies of resulting trees were manually compared to validate the crucial nodes supporting major conclusions. Phylogenetic trees were visualized using iTOLv6.8 (76).

Transcriptomic analyses

For transcriptome analyses, a subset of 43 species was selected based on specific criteria: (1) the presence of *FLS*, *DFR*, *F3H*, and *F3'H* genes in the species, (2) the availability of sufficient transcriptomic datasets in the Sequence Read Archive (SRA), and (3) the presence of clear labels for RNA-seq samples with a minimum of three accession numbers corresponding to at least a single plant organ (leaf, root, stem, flower, seed, or fruit). To generate the count tables, all available paired-end RNA-seq datasets of the selected species were retrieved from the SRA (www.ncbi.nlm.nih.gov/sra) (77) using fastq-dump (https://github.com/ncbi/sra-tools). Kallisto v0.44 (78) was then employed to quantify transcript abundance based on the coding sequences of the respective plant species. The previously developed Python script kallisto_pipeline3.py (79,80) served as a wrapper to run kallisto across all input files. Next, the Python script merge_kallisto_output3.py (79,80) was executed to merge the individual kallisto output files into one count table per species. The count tables were processed to create a unified count table, encompassing the expression values of *FLS* and *DFR* across all 43 species. To explore the expression patterns of *FLS* and *DFR*, a 2D density heatmap with marginal histograms was generated using a customized python script (available via GitHub). Additionally, the accession numbers of different tissues (leaf, stem, flower, root, seed, and fruit) for the 43 species were

retrieved from the SRA. This sample selection was further reduced to avoid the inclusion of any specific stress treatments i.e. the different samples should represent different plant parts under 'normal' conditions. We aimed to assess the transcript abundance of a specific gene in a particular sample. The Python script generate_subset.py (81) was deployed to extract an expression data subset from the count table containing exclusively the values belonging to the accession numbers of the desired samples. The transcripts per million (TPMs) belonging to isoforms of a gene and close paralogs were aggregated per RNA-seq sample as previously described (82). To obtain an overall representation of the gene's transcript abundance, we calculated the median value of the aggregated TPMs across a group of samples(**Figure 3**). The median was chosen due to its tolerance towards outliers. To visualize the gene expression patterns, a heatmap was generated using the seaborn package (83) in Python. The heatmap depicted the relative gene expression levels of *F3H*, *F3'H*, *F3'5'H*, *FLS*, and *DFR*. Additionally, ratios of *F3'H*/*F3H*, *F3'5'H*/*F3H*, and *DFR*/*FLS* were plotted to provide insights into their relative expression levels and potential interdependencies within the flavonoid biosynthesis pathway.

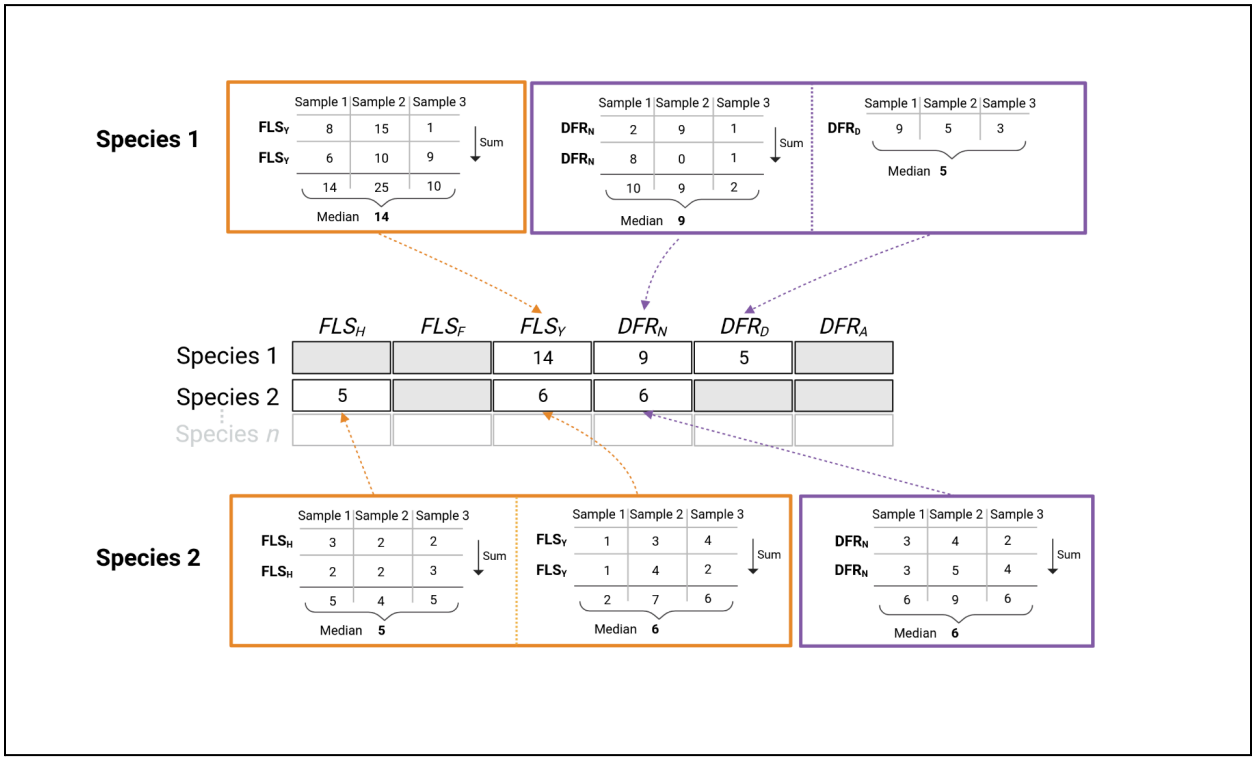

Figure 3: Illustration of the cross-species gene expression comparison. All displayed values are TPMs. The values of close paralogs were replaced by their sum if these paralogs had the same substrate affinity determining residue at the most important position.

Analyses

Substrate preference determining amino acid residues in FLS and DFR and their pattern of occurrence in land plants

From a vast range of diverse taxonomic plants, DFR was identified only in angiosperms. Among the 175 angiosperm datasets analyzed, 129 exhibited at least one most likely functional DFR, characterized by the presence of essential residues associated with DFR activity. Overall, 207 DFR sequences were identified, accounting for cases where multiple DFRs were present in certain species. Multiple sequence alignments of the identified DFR sequences revealed potential NADPH binding and substrate binding sites in N-terminal regions (**Figure 4**). Analysis based on the conservation of the amino acid residue at the position 133 i.e. 3rd position within the 26-amino acid-long substrate binding domain of DFR led to the classification of DFR proteins into three types:

(i) DFR with asparagine (N) at the 3rd position which can recognize all three dihydroflavonols as substrates, (ii) DFR with aspartic acid (D) at the 3rd position which shows higher preference for dihydroquercetin and dihydromyricetin, and (iii) DFR characterized by alanine (A) at the 3rd position which exhibits a preference for dihydrokaempferol and a higher likelihood of rejecting dihydromyricetin. Henceforth, these three DFR types will be referred to as DFR_N, DFR_D, and DFR_A, respectively.

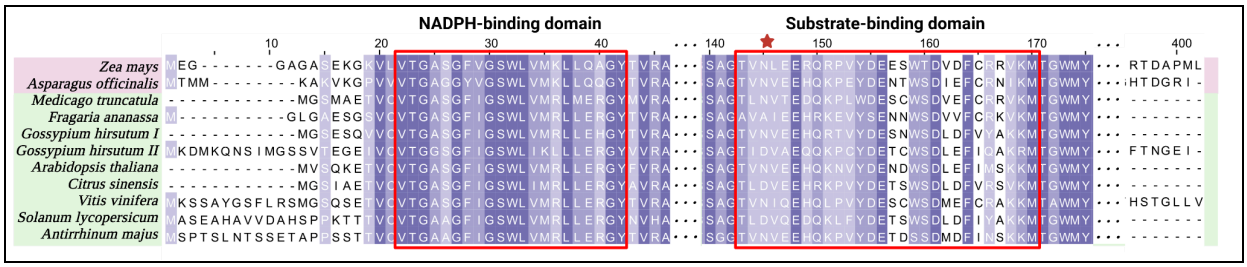

Figure 4: Selected parts of multiple sequence alignment of DFR sequences restricted to 10 species to reduce the complexity. Monocot and dicot species are highlighted by light pink and light green colour, respectively. The dark violet, mid violet, and light violet colours indicate 100%, >75%, and >50% similarity among the sequences, respectively. Important domains associated with DFR functionality are highlighted and some columns are masked with three dots (**...**). Red boxes highlight the NADPH-binding domain and the 26 amino acid long substrate-binding domain. The red-star labeled residue at the 3rd position within this region is crucial for dihydroflavonol recognition (position 133 in the *A. thaliana* DFR). MAFFTv7 was applied to generate the alignment.

Sequences of FLS enzymes were successfully identified in the datasets of 143 out of 175 angiosperm species. The inability to identify FLS sequences in some of the plant species may be attributed to technical reasons i.e. incompleteness of the analyzed data sets. In total, 247 FLS sequences were identified, including instances where multiple copies of FLS were present within a species. FLS enzymes belong to the 2-oxoglutarate-dependent dioxygenase (2-ODD) superfamily. A comprehensive analysis of the multiple sequence alignment revealed the presence of characteristic conserved ferrous iron-binding motif (HX(D/E)XnH) and the 2-oxoglutarate binding residues (RXS) critical for FLS functionality (**Figure 5**). Furthermore, sequences of functional FLSs exhibited additional residues (G68, H75, P207, and G261) responsible for the proper folding of the 2-ODD protein. Specific motifs, such as "PxxxIRxxxEQP" and "SxxTxLVP" which were initially considered to be unique to FLS and distinguished it from other plant 2-ODDs (84) were also identified. However, a previous study identified two *BnaFLS1* homologs which also exhibited F3H activity (85). It was demonstrated

that despite having these reportedly FLS-specific motifs, they could still have additional activities. FLS sequences also possessed five key amino acids (H132, F134, K202, F293, and E295) identified as potential DHQ binding site residues, with the latter four exhibiting notable conservation across diverse species. These residues were also found to be conserved in anthocyanidin synthase (ANS) sequences wherein they had a strongly conserved Y132 residue as well. Analysis based on the conservation of residue 132 led to the classification of FLS into three types: FLS_H (histidine), FLS_F (phenylalanine) and FLS_Y (tyrosine). The positions of amino acids or nucleotides are assigned based on the *Arabidopsis thaliana* sequences (*At*DFR, AT5G42800 and *At*FLS1, AT5G08640) for consistency and reference purposes.

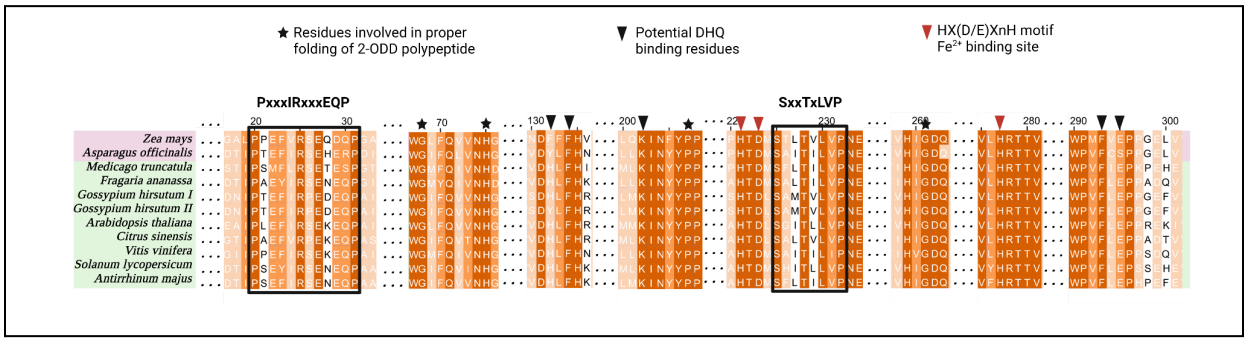

Figure 5: Selected parts of a FLS multiple sequence alignment restricted to 10 species to reduce the complexity. Monocot and dicot species are highlighted by light pink and light green colour, respectively. The dark orange, mid orange and light orange colours indicate 100%, >75%, and >50% similarity among the sequences, respectively. Important domains and residues associated with FLS functionality are highlighted and some columns are omitted as indicated by three dots (**...**). The conserved 2-ODD domain and the Fe^{2+} binding sites are indicated by black asterisks and red arrowheads, respectively. The black boxes highlight the FLS-specific motifs. The black arrowheads indicate potential DHQ-binding sites where the first residue (position 132) is thought to be critical for the substrate preference of FLS. The alignment was generated by MAFFTv7.

The 129 diverse plant species harboring DFRs represented 28 orders out of the 64 orders recognized in the current Angiosperm Phylogeny Group (APG IV;2016) classification of

angiosperms (86). Interestingly, all monocot plants investigated displayed DFR_N, with only a few minor exceptions which displayed neither of the three types. Among dicots, some species exclusively possessed either DFR_N or DFR_D, while others exhibited a combination of either DFR_N and DFR_D or DFR_N and DFR_A. Remarkably, DFR_A, which is known to have a strong preference for DHK as a substrate, previously reported only in *Fragaria*, was also identified in *Spirodela polyrhiza*, *Spirodela intermedium* and *Begonia peltifolia* (Additional File 2). A total of 143 angiosperm plant species encompassing 25 orders were analyzed for functional FLS genes. Among these species, monocots exhibited either FLS_F or FLS_Y, while all dicots possessed FLS_H with a histidine residue at position 132. However, certain orders such as Fabales, Fagales, Malpighiales and Malvales displayed multiple FLS candidates in some species with both FLS_H and FLS_Y, characterized by histidine and tyrosine residues, respectively. (Additional File 3). The overall pattern of amino acid residues at position 133 in DFR sequences and position 132 in FLS sequences at the order level is summarized in **Figure 6**.

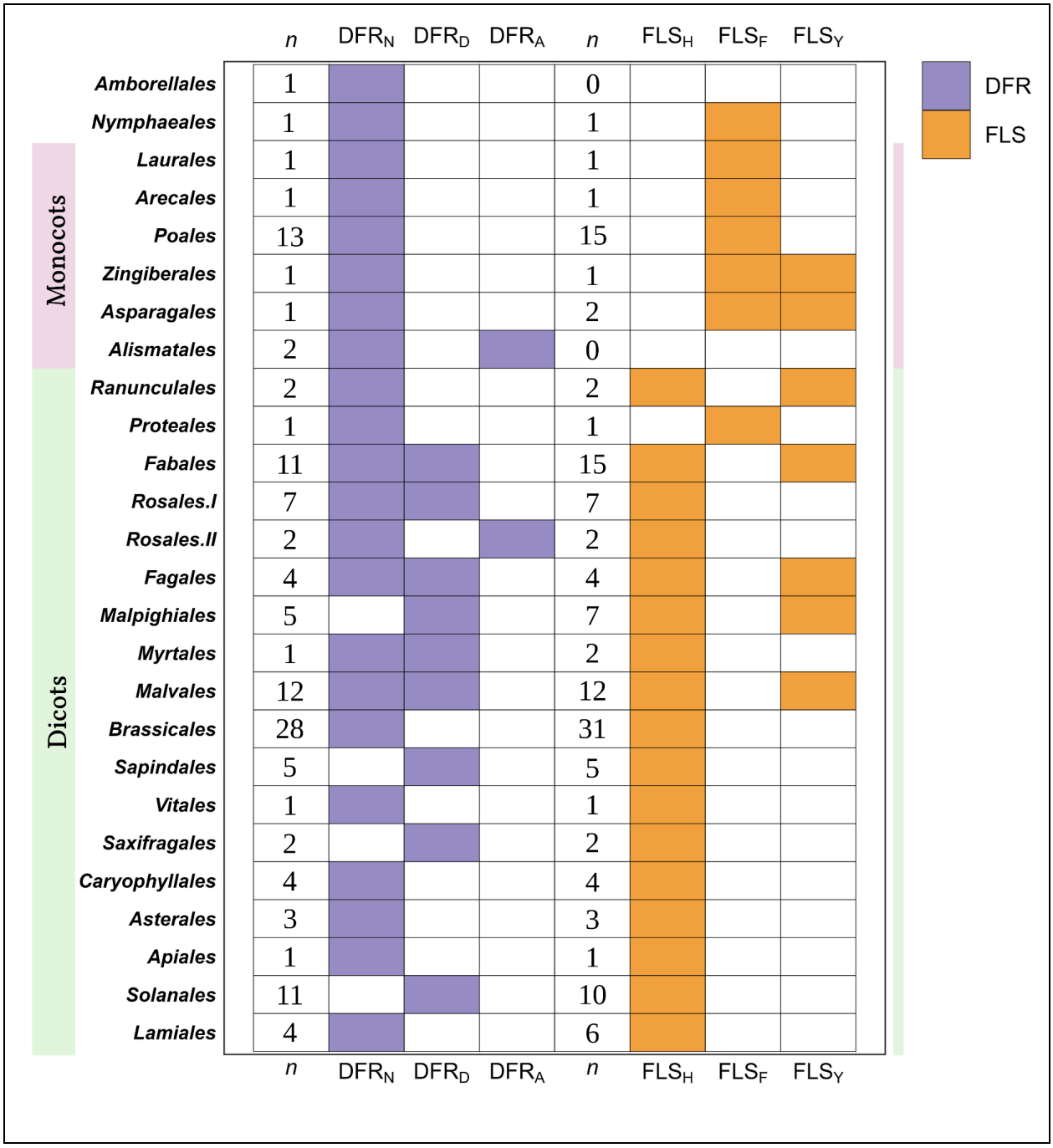

Figure 6: The patterns of commonly occurring amino acid residues at substrate preference determining positions observed in different orders of angiosperms for FLS and DFR. Orders are sorted by branching in the evolution of angiosperms according to the Angiosperm Phylogeny Group (2016). The residues were investigated across 172 angiosperm species and are represented here at the order level. Monocot and dicot species are highlighted by light pink and light green colour, respectively. Rosales II consists of 2

species: *Fragaria x ananassa* and *Fragaria vesca*. *n* represents the number of analyzed species harboring DFR and FLS, respectively, within each order.

Evolution of DFR

To understand the evolutionary relationships among multiple DFR sequences, we constructed a phylogenetic tree with *DFR* protein-encoding sequences. Given that DFR belongs to the short-chain dehydrogenase (SDR) family, the closely related ANR, CCR, and DFR-like sequences from this family were included as outgroups. In the resulting analysis, the ANRs, and DFRs clustered together, distinct from the CCRs, as depicted in Additional File 4. The two clades placed as sisters to the CCR clade had *A. thaliana* Tetraketide α-pyrone reductase (*AtTKPR1* and *AtTKPR2*) in each clade, essential for pollen wall development and male fertility. TKPR2, known to be more active than TKPR1, clustered closer to the CCR clade. The TKPR1-like, TKPR2-like, CCRs and a large clade with an unknown function clustered together and were placed as a sister clade to the DFR and ANR clade. Adjacent to the DFR clade, another clade featured AtBEN1-encoding a DFR-like protein regulating brassinosteroids in *A. thaliana*. This clade did not contain any sequences from the gymnosperms or monocots, only dicots were found. The proximity to the DFR clade and the absence of monocots suggests its evolutionary origin from a DFR predecessor gene.

We did not identify any Gymnosperm DFR in our analysis, possibly due to their early-branching nature and distinct conserved residues compared to angiosperms. Previously characterized DFR sequences (Additional File 5), marked with red asterisks, were generally located within the DFR clade, except for fern *Dryopteris erythrosora*, *DeDFR1* and *DeDFR2*, *Ginkgo biloba*, *GbDFR1*, *GbDFR2*, and *GbDFR3*, and *Camellia sinensis*, *CsDFRc*, which were outside, generally indicating non-DFR character of these sequences or an independent origin of DFR activity. Fern DFRs, found outside the ANR and DFR clade, may exhibit distinct characteristics due to their early-branching nature or possess promiscuous functions. Ginkgo DFRs within

TKPR-like clades likely deviate from true DFRs. The phylogenetic analysis of functional DFRs highlights distinct clades specific to monocots and dicots (**Figure 7**). Monocots exclusively exhibit DFR_N, while dicots generally display various DFR types - DFR_N, DFR_D, and DFR_A. Notably, DFR_D appears to have originated from a duplication event of DFR_N, as observed in the Fabales and Malvales clades. Within the Rosales clade, alanine is present. These findings suggest possible scenarios where dicots experienced a duplication event in DFR_N, followed by a subsequent substitution event resulting in the emergence of DFR_D. Furthermore, in the Rosales clade, the duplication could have been accompanied by a substitution from N to A. Upon examining the codons of the DFR_A coding sequence at position 133 in *F. vesca* and *F. ananassa* and comparison against closely related DFR_N sequences, we found that the codons did not exhibit any predisposition towards the emergence of DFR_A in strawberries (AAT to GCC) (Additional file 6).

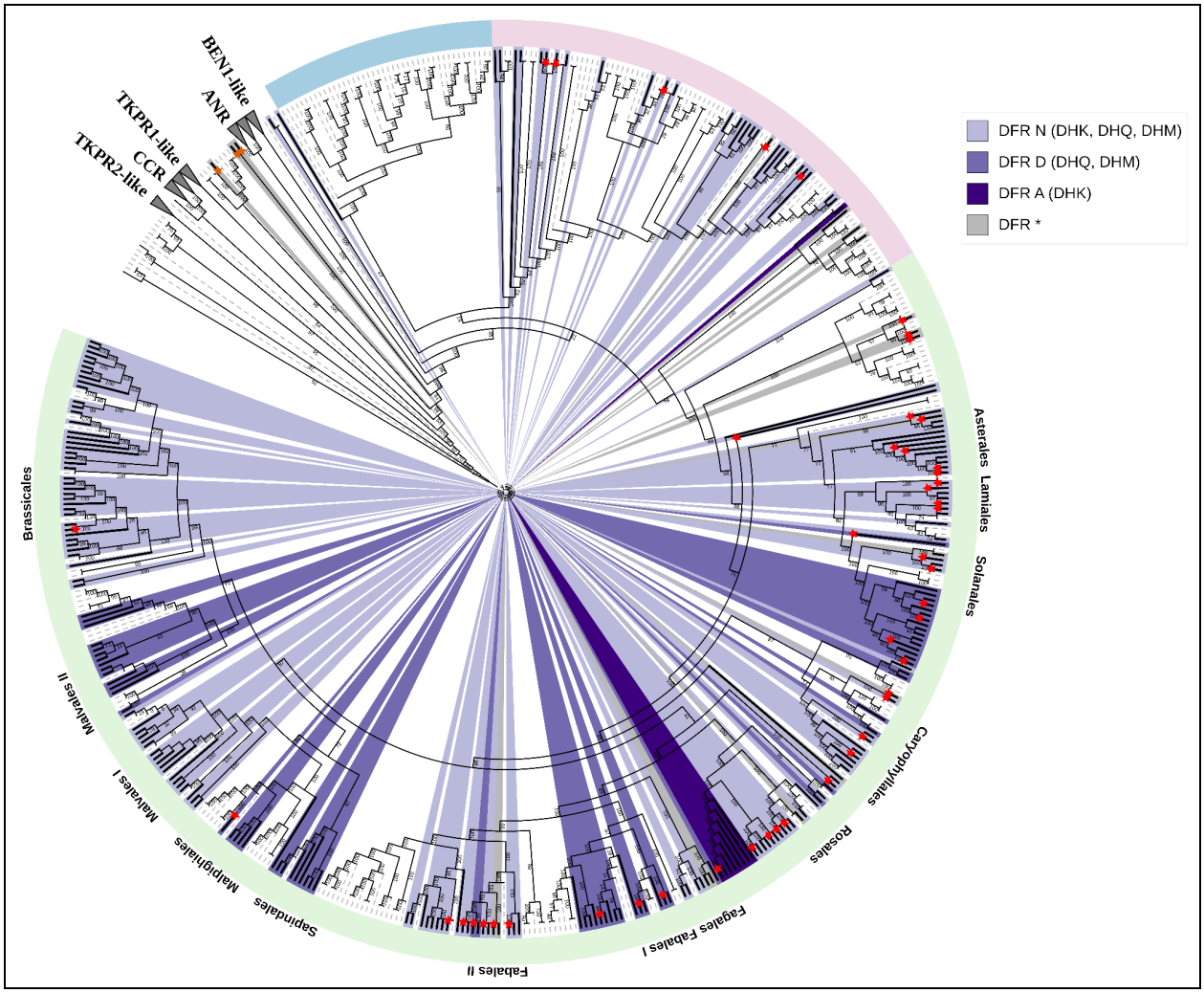

Figure 7: Phylogenetic analysis of DFR sequences in diverse plant species, highlighting amino acid residue diversity at position 133 associated with substrate specificity. Gymnosperm, monocot and dicot species are denoted by light blue, light pink and light green colour stripes, respectively. Non-DFR sequences are represented by dashed gray branches while the functional DFRs are indicated by solid black branches. The color-coded scheme represents different residues: Asparagine (light purple), Aspartic acid (periwinkle blue), Alanine (deep purple), and other amino acids (grey) at position 133. The preferred substrate of the DFR type is written in brackets, DHK, dihydrokaempferol; DHQ, dihydroquercetin and DHM, dihydromyricetin. Distinct clusters of DFRs from major plant orders are labeled for reference. DFR genes characterized in previous studies are highlighted by a red asterisk at the start of the terminal branch (Additional File 5). Leaf labels are hidden to reduce the complexity. The outgroup comprises SDR members like ANRs, CCRs and other DFR-like sequences (Additional File 4).

Deep duplication events during the evolution of FLS

Through a phylogenetic tree generated with *FLS* protein-encoding sequences, we observed deep duplication events during the evolution of FLS enzymes. Considering their classification within the 2-oxoglutarate-dependent dioxygenase (2-ODD) family, closely related ANS, F3H and FNSI sequences from the same family and FLS-like sequences lacking crucial residues for FLS functioning were included as outgroups. The phylogenetic tree revealed a largely distinct clustering of F3H, ANS and FLS sequences. ANS and FLS formed a common cluster, while F3H grouped with FNSI sequences (Additional File 7).

The non-Apiaceae FNSI sequences clustered outside the F3H clade while the liverworts FNSI and Apiaceae FNSI were a part of a single clade with phylogenetically distinct F3H sequences. The liverworts FNSI are present at the root of this clade suggesting that plant F3H evolved from liverworts FNSI. An independent lineage of Apiaceae FNSI is located nested within the Apiaceae F3H sequences further confirming the evolution of Apiaceae *FNSI* from *F3H* (79,87). Gymnosperm, monocot and dicot sequences formed distinct clades within larger clades corresponding to the three 2-ODD members, i.e. F3H, ANS and FLS. The FLS clade was divided into four primary clusters: a Gymnosperm clade branching first, a second clade which we hereafter call aFLS (ancestral FLS), a third monocot clade and lastly, a large dicot clade. The aFLS clade encompasses FLS-like sequences from basal angiosperms and dicots and a few dicot FLS sequences. Only nine FLS sequences, one each from *Linum usitatissimum, Lotus japonicus, Cajanus cajan, Quercus suber, Betula platyphylla, Acacia confusa AcFLS* (JN812062), *Camellia sinensis CsFLSa* (KY615705), *Fagopyrum tataricum FtFLS2* (JX401285) *and Vitis vinifera FLS5* (AB213566) were in this clade, although the FLS-like sequences were from a wide range of angiosperms excluding monocots. The presence of this aFLS clade sister to the distinct monocot and dicot clade implies an FLS duplication event preceding the monocot-dicot split. The large dicot cluster further separates into two distinct subclades: one including sequences exclusively belonging to the Brassicales order, and the other including

sequences from the rest of the dicot plants (**Figure 8**). In the Brassicales subclade, a monophyletic group of Brassicales FLS and FLS-like sequences is evident, with *AtFLS1* and other functional Brassicales FLSs clustered together. The presumed non-functional FLS sequences from the Brassiclaes order group together and are classified based on their phylogenetic relationship to their most likely *A. thaliana* orthologs (*AtFLS2-AtFLS6*). Within the latter subclade which excludes Brassicales, some FLS sequences from the same order appear in distinct clusters as observed in the Fagales, Sapindales, Malpighiales and Malvales order, suggesting gene duplication events in the evolutionary history of FLS in some common ancestor of these orders.

While the first and last residues of the FLS substrate binding residues (H132, F134, K202, F293, and E295) exhibit diverse residues, the central three are strongly conserved in both FLS and ANS sequences. Notably, gymnosperm FLSs and aFLSs feature a Y132 residue, consistent with ANS sequences. Monocotyledonous plants display an F/Y(85%/15%) residue at position 132, whereas the majority of dicots consistently possess an H residue at position 132, except for multiple *Gossypium* species and *Herrania umbratica* from the Malvales order, which have a Tyrosine residue at this position. The E295 residue is conserved in ANS sequences as well as in monocot and dicot FLSs. However, sequences from the gymnosperm, aFLSs, AtFLS4, and AtFLS5 clades exhibit non-conserved amino acid residues at position 295. The ancestral FLSs (aFLSs) exclusively feature aliphatic amino acids at this position (G/A/V/L/I).

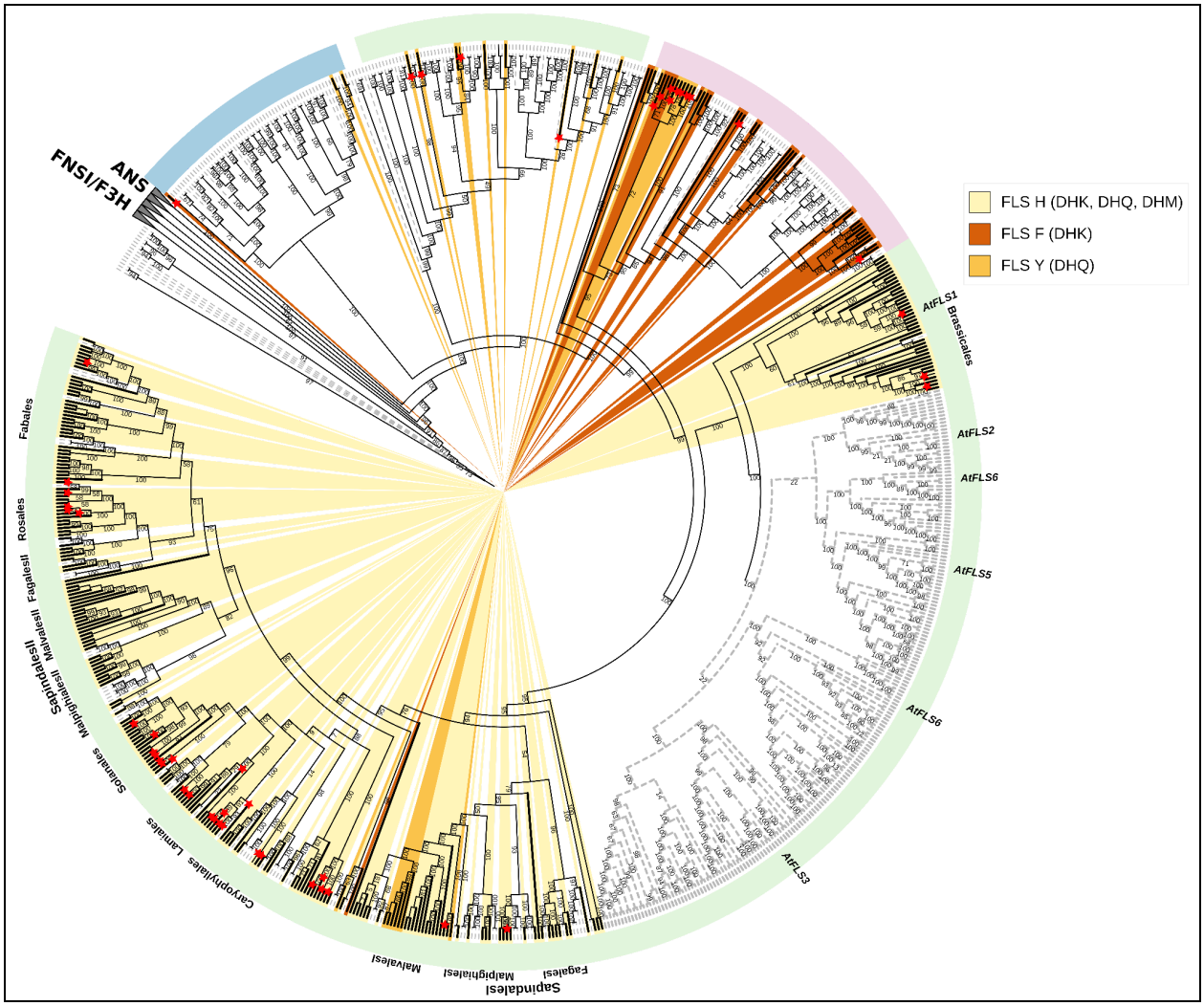

Figure 8: Phylogenetic diversity of FLS sequences in diverse plant species, highlighting amino acid residue diversity at position 132 associated with DHQ-substrate binding. Gymnosperm, monocot and dicot species are denoted by light blue, light pink and light green colour stripes, respectively. Non-FLS sequences are represented by dashed gray branches while the functional FLSs are indicated by solid black branches. The background colour highlights different presumably substrate-determining amino acid residues: histidine (pale yellow), phenylalanine (dark orange), and tyrosine (dark golden) at position 132. The hypothesized preferred substrate of the FLS type is written in brackets, DHK, dihydrokaempferol; DHQ, dihydroquercetin and DHM, dihydromyricetin. Distinct clusters of FLSs from major plant orders are labeled for reference. FLS genes characterized in previous studies (Additional File 5) are highlighted by a red asterisk at the start of the terminal branch. The branches of *Arabidopsis thaliana* AtFLS1-AtFLS6 are

labeled. Leaf labels are hidden to reduce the complexity. The outgroup comprises members of 2-ODD like F3H, ANS and other FLS-like sequences (Additional File 7).

Divergent expression patterns of *FLS* and *DFR*

Generally, either *DFR* or *FLS* is expressed in a sample, but almost never both simultaneously (**Figure 9a**). This observation held true across various *FLS* and *DFR* types, underscoring a general phenomenon. However the pattern intriguingly deviates in some cases.

The mutually exclusive expression patterns were apparent for all combinations of *FLS* and *DFR* types (**Figure9b-g**). These patterns were particularly pronounced in the DFR_A vs FLS_H expression and DFR_N vs FLS_Y expression plots. Contrastingly, in cases such as DFR_N vs FLS_H and DFR_D vs FLS_H, instances of co-expression of *FLS* and *DFR* were observed in some samples.

Theoretically, there could be 9 different *FLS-DFR* combinations, given three *FLS* and three *DFR* types were examined. The absence of certain combinations raises the possibility that not all *FLS* and *DFR* combinations may occur in nature. For instance, to the best of our knowledge, FLS_F is restricted to monocots, exclusively paired with DFR_N. Hence, combinations involving DFR_D and DFR_A alongside FLS_F might not exist in nature.

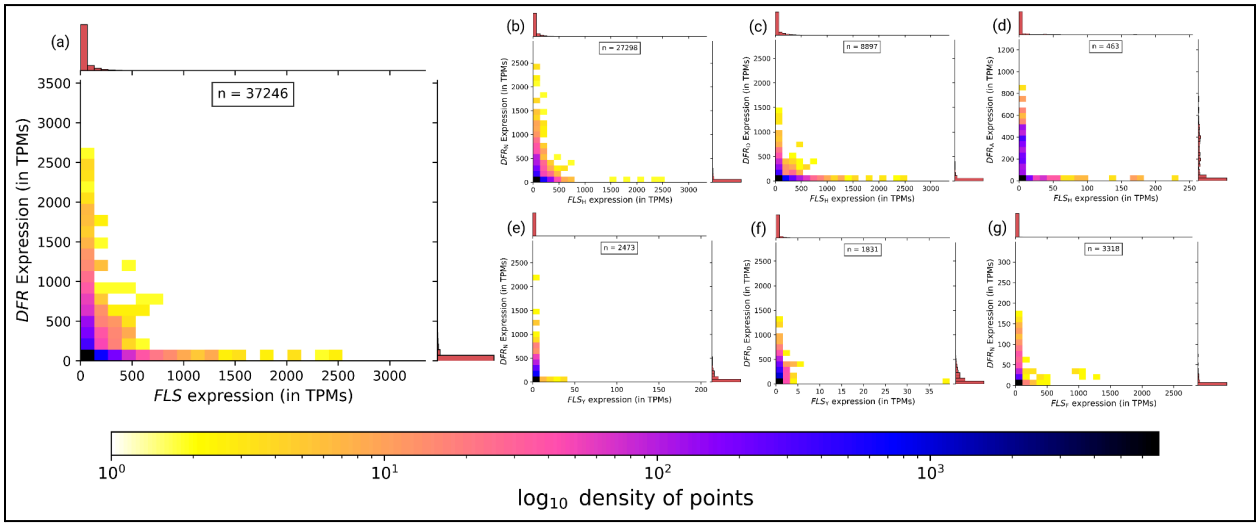

Figure 9: 2D density heatmap with marginal histograms showing divergent expression patterns

of FLS vs DFR across samples from 43 species. Each sample shows only the expression of *FLS* or *DFR*, but not both. The sample size is indicated by n. Colorbar depicts the logarithmic density of points in the plot. (a) Combined expression of all *FLS* and *DFR* types, (b-g) specific combinations of *FLS* and *DFR* types.

Inferring dihydroflavonol availability from relative *F3H*, *F3'H*, and *F3'5'H* gene expression

Besides *FLS* and *DFR*, the major genes influencing the branching into flavonol and anthocyanidin biosynthesis are *F3H*, *F3'H*, and *F3'5'H*. These genes play a crucial role in determining the hydroxylation patterns of dihydroflavonols, which are common substrates for FLS and DFR. F3H catalyzes the formation of DHK, while F3'H adds a hydroxy group converting DHK into DHQ. F3'5'H catalyzes the formation of triple hydroxylated DHM from DHK/DHQ (**Figure 1**). Understanding the substrate preference of FLS and DFR for differently hydroxylated substrates requires information about the relative availability of DHK, DHQ, and DHM. Since metabolic datasets are not available at the necessary scale, we utilized gene expression data as a semi-quantitative proxy for metabolomics. Transcript abundances are considered as well correlated with the final amount of the gene product, because the flavonoid biosynthesis is mostly regulated at the transcriptional level. The relative gene expression of *F3H*, *F3'H*, and *F3'5'H* is analyzed to infer the relative production of the three flavonol/anthocyanin precursors i.e. DHK, DHQ, and DHM. The ratio of *F3'H* to *F3H* and the ratio of *F3'5'H* to *F3H* were calculated across 43 plant species. For example, a high *F3'H* to *F3H* ratio would suggest that DHQ accounts for most of the dihydroflavonols. These gene expression ratios were connected to the expression levels and substrate-determining amino acid residues in the proteins encoded by *FLS* and *DFR*.

To investigate the gene expression patterns, a heatmap showing the transcript abundance of the above-mentioned genes in 43 species representing the taxonomic diversity of angiosperms

was generated (Additional File 8). In **Figure 10**, 10 species are presented to reduce the complexity. The heatmap covers several tissue samples such as leaf, root, stem, flower, fruit, and seed. The samples displayed per individual species depend on data availability. The heatmap analysis revealed distinct expression patterns of *FLS* and *DFR* genes in different plants. In **Figure 10**, the monocot, *Miscanthus sinensis*, demonstrates a higher *F3'H* to *F3H* and *DFR* to *FLS* ratio, suggesting a potential predominance of dihydroquercetin utilization by DFR for anthocyanin or proanthocyanidin production. Cotton exhibits multiple FLS and DFR candidates of FLS_H, FLS_Y, DFR_N, and DFR_D type. Its FLS_Y, limited to the Malvales order in the dicot clade, has low transcript abundance. In cotton, high expression of *F3H*, *F3'H*, *F3'5'H*, FLS_H, DFR_N, and DFR_D signifies flavonoid biosynthesis activity. *Arabidopsis thaliana* lacks *F3'5'H* and exhibits relatively low *DFR* transcript abundance. The *F3H* and *FLS* transcript abundance follow the same patterns, suggesting a preference for DHK as a substrate. Species with exclusive DFR_D (aspartate), like oranges and tomatoes, exhibit low *DFR* transcript abundance despite significant *F3'H* expression. In *Dianthus caryophyllus*, red and orange flowers exhibit significant upregulation in *F3H* and DFR_N expression compared to the white flower, while *FLS* expression remains relatively constant. Interestingly, F3'H expression is the least in orange flowers suggesting that its DFR_N is predominantly acting on DHK as substrate leading to the formation of pelargonidin-based anthocyanins, hence the orange colour. The red flower has high transcript levels of both F3H and F3'H suggesting DFR_N here might be catalyzing both DHK and DHQ,, leading to the production of cyanidin- and pelargonidin-based pigments and contributing to the observed red phenotype.

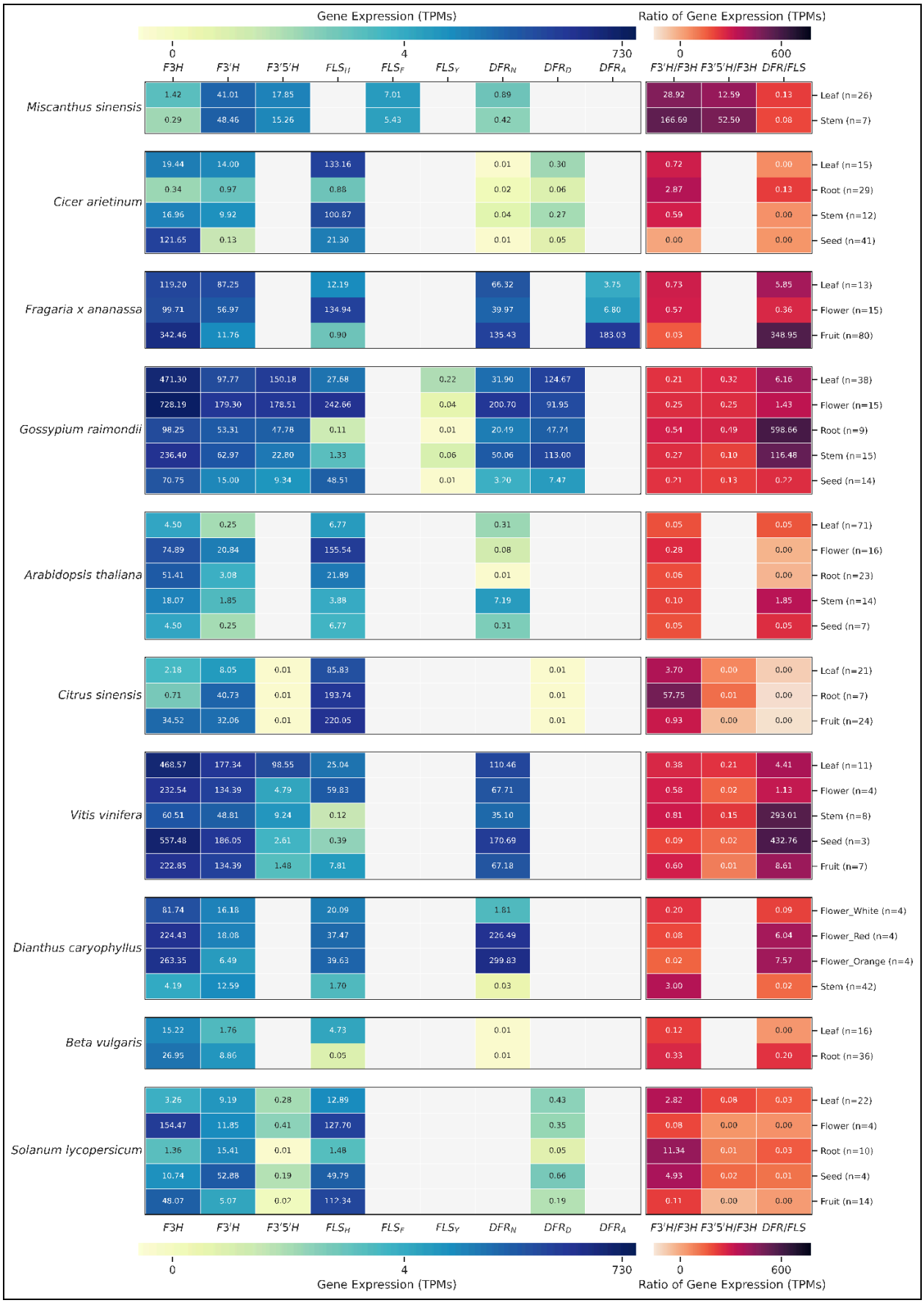
Gene Expression (TPMs)
0
4
730
Ratio of Gene Expression (TPMs)
0
600
F3H
F3'H
F3'5'H
FLS_H
FLS_F
FLS_Y
DFR_N
DFR_D
DFR_A
F3'H/F3H
F3'5'H/F3H
DFR/FLS
Miscanthus sinensis
1.42 41.01 17.85 7.01 0.89
28.92 12.59 0.13 Leaf (n=26)
0.29 48.46 15.26 5.43 0.42
166.69 52.50 0.08 Stem (n=7)
Cicer arietinum
19.44 14.00 133.16 0.01 0.30
0.72 0.00 Leaf (n=15)
0.34 0.97 0.88 0.02 0.06
2.87 0.13 Root (n=29)
16.96 9.92 100.87 0.04 0.27
0.59 0.00 Stem (n=12)
121.65 0.13 21.30 0.01 0.05
0.00 0.00 Seed (n=41)
Fragaria x ananassa
119.20 87.25 12.19 66.32 3.75
0.73 5.85 Leaf (n=13)
99.71 56.97 134.94 39.97 6.80
0.57 0.36 Flower (n=15)
342.46 11.76 0.90 135.43 183.03
0.03 348.95 Fruit (n=80)
Gossypium raimondii
471.30 97.77 150.18 27.68 0.22 31.90 124.67
0.21 0.32 6.16 Leaf (n=38)
728.19 179.30 178.51 242.66 0.04 200.70 91.95
0.25 0.25 1.43 Flower (n=15)
98.25 53.31 47.78 0.11 0.01 20.49 47.71
0.54 0.49 598.66 Root (n=9)
236.40 62.97 22.80 1.33 0.06 50.06 113.00
0.27 0.10 116.48 Stem (n=15)
70.75 15.00 9.34 48.51 0.01 3.70 7.47
0.21 0.13 0.22 Seed (n=14)
Arabidopsis thaliana
4.50 0.25 6.77 0.31
0.05 0.05 Leaf (n=71)
74.89 20.84 155.54 0.08
0.28 0.00 Flower (n=16)
51.41 3.08 21.89 0.01
0.06 0.00 Root (n=23)
18.07 1.85 3.88 7.19
0.10 1.85 Stem (n=14)
4.50 0.25 6.77 0.31
0.05 0.05 Seed (n=7)
Citrus sinensis
2.18 8.05 0.01 85.83 0.01
3.70 0.00 0.00 Leaf (n=21)
0.71 40.73 0.01 193.74 0.01
57.75 0.01 0.00 Root (n=7)
34.52 32.06 0.01 220.05 0.01
0.93 0.00 0.00 Fruit (n=24)
Vitis vinifera
468.57 177.34 98.55 25.04 110.46
0.38 0.21 4.41 Leaf (n=11)
232.54 134.39 4.79 59.83 67.71
0.58 0.02 1.13 Flower (n=4)
60.51 48.81 9.24 0.12 35.10
0.81 0.15 293.01 Stem (n=8)
557.48 186.05 2.61 0.39 170.69
0.09 0.02 432.76 Seed (n=3)
222.85 134.39 1.48 7.81 67.18
0.60 0.01 8.61 Fruit (n=7)
Dianthus caryophyllus
81.74 16.18 20.09 1.81
0.20 0.09 Flower_White (n=4)
224.43 18.08 37.47 226.49
0.08 6.04 Flower_Red (n=4)
263.35 6.49 39.63 299.83
0.02 7.57 Flower_Orange (n=4)
4.19 12.59 1.70 0.03
3.00 0.02 Stem (n=42)
Beta vulgaris
15.22 1.76 4.73 0.01
0.12 0.00 Leaf (n=16)
26.95 8.86 0.05 0.01
0.33 0.20 Root (n=36)
Solanum lycopersicum
3.26 9.19 0.28 12.89 0.43
2.82 0.08 0.03 Leaf (n=22)
154.47 11.85 0.41 127.70 0.35
0.08 0.00 0.00 Flower (n=4)
1.36 15.41 0.01 1.48 0.05
11.34 0.01 0.03 Root (n=10)
10.74 52.88 0.19 49.79 0.66
4.93 0.02 0.01 Seed (n=4)
48.07 5.07 0.02 112.34 0.19
0.11 0.00 0.00 Fruit (n=14)
F3H
F3'H
F3'5'H
FLS_H
FLS_F
FLS_Y
DFR_N
DFR_D
DFR_A
F3'H/F3H
F3'5'H/F3H
DFR/FLS
0
4
730
Gene Expression (TPMs)
0
600
Ratio of Gene Expression (TPMs)

Figure 10: Transcript abundance heatmap and gene expression ratio analysis of 11 selected species across various tissue samples. White tiles indicate that no gene was discovered for a certain function/type. The heatmap tiles represent the actual expression values in TPMs, with the number of samples indicated on the right. The expression heatmap shows genes associated with the branching of *FLS* and *DFR* within the flavonoid biosynthesis pathway. Blue and yellow colours indicate high and low expression levels, respectively (see colour bar). On the right, a heatmap illustrates the ratios of gene expression to investigate the substrate hydroxylation pattern in the FLS-DFR branching. Dark and light colours indicate high and low expression levels, respectively, based on the colour bar. A full heatmap with all 43 investigated species is available in Additional File 8.

Discussion

Factors influencing the competition of the pivotal enzymes FLS and DFR for the common substrate, dihydroflavonols, remain poorly understood. The equilibrium between these enzymes can be modulated by several mechanisms including (i) the unique substrate specificity exhibited by FLS and DFR for different dihydroflavonols, (ii) the existence of multiple FLS/DFR copies with narrow substrate preferences, or (iii) the expression of specific *FLS*/*DFR* copies in specific cell types or under specific conditions, (iv) the availability of substrates i.e. different dihydroflavonols, determined by the activity of flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase, and (v) the channeling of substrate through metabolons.

Mitigating substrate competition between FLS and DFR

A well-studied 26-amino acid region determines substrate specificity in DFR (50,88) and the grape crystal structure further emphasized the significance of the N133 amino acid in binding the 3' and 4' hydroxyl groups of DHQ (58). While the substrate specificity of DFR has been

extensively studied, the substrate specificity determining residues of FLS have received comparatively less attention. The substrate-binding residues of FLS (H132, F134, K202, F293, E295) show high conservation among all plant FLSs, with the exception of the H132 and E295 residues. Through the examination of *A. thaliana fls1* mutant, it was hypothesized that the H132 residue may play a critical role in governing the diverse substrate preferences of FLS enzymes (59). Numerous previous studies have employed enzyme assays to investigate the substrate preferences of plant FLSs and DFRs, while many others identified the type of anthocyanins and flavonols in various plants. Notably, *AcFLS* from the monocot *Allium cepa*, characterized by the Y132 residue, exhibits a preference for DHQ over DHK (89), while monocots such as *OsFLS* (90) and *ZmFLS* (91), with the F132 residue, favor DHK over DHQ. In contrast, DFR_N in maize and *Dendrobium officinale* predominantly produces cyanidin derivatives (92,93). DFR_N of barley displays *in vitro* reduction of DHK and DHQ whereas DHM was not tested (94). Reports of procyanidin and prodelphinidin in testa tissue and a lack of detection of leucopelargonidin and its derivatives (95) do not align with a DHK and DHQ specificity of DFR in barley. DFR_N of grape hyacinth and *Cymbidium* hybrids exhibit high activity towards DHM and DHQ (88,96), and *Anthurium andraeanum* DFR, featuring an uncommon S133 residue, displays a marked specificity for DHK (97).

Previous studies and our analyses identified multiple DFR copies in dicots. For example, in the Fabales order, *M. truncatula* DFR1 and *L. japonicu*s DFR2 and DFR3 with an N133 residue reducesDHK more readily than DHQ, while *M. truncatula* DFR2 and *L. japonicus* DFR5 with a D133 residue show a higher activity towards DHQ (56,57). In these studies, the relative DFR activity was demonstrated by HPLC and spectrophotometric assays, respectively. In the Rosales order, FLS_H from *Fragaria ananassa* (98), *Malus domestica* (99), and *Rosa hybrida* (100) reduces both DHK and DHQ, with strawberry FLS favoring DHK and rose FLS favoring DHQ. Conversely, their DFR_N exhibits a higher affinity for DHQ than DHK (63,101,102). The strawberry species, *Fragaria ananassa* and *Fragaria vesca* have another DFR-type, DFR_A which

has a very high specificity for DHK (63). Other examples are *L. japonicus* and *M. truncatula*, where a DFR_D is present with a high preference for DHQ, while the DFR_N prefers DHK.

This observation seems to contradict previous reports about the substrate preferences of DFR_N. A potential explanation might be sub-/neofunctionalization upon duplication of DFR. We speculate that through a yet unknown mechanism, DFR_N might become more specific for the substrate that is not favored by the other DFR copy. Waki et al. reported that chalcone isomerase-like proteins (CHILs) can bind to chalcone synthase (CHS) and rectify its promiscuous activity making CHS more specific (103). In a conceptually similar mechanism, multiple DFR copies with distinct specificities could be influenced by external factors to adjust their substrate preferences i.e. to make the DFR_N type accept the substrate not favored by the alternative enzyme copy.

In the Brassicales, FLS_H from *A. thaliana* and *Matthiola incana* prefer DHK (104–106), while their DFR_N reduces both DHK and DHQ but favors DHQ (107–109). *Citrus unshiu* FLS_H has higher DHK activity (110), whereas *Zanthoxylum bungeanum* from the same order, Sapindales, has DFR_D and prefers DHM followed by DHQ and DHK (111). In Caryophyllales, FLS_H in *Dianthus caryophyllus*, *Fagopyrum dibotrys*, and *Fagopyrum tataricum* catalyze both DHK and DHQ (112–114), while DFR_N in *Dianthus* reduces DHQ more effectively than DHK (113). In *Fagopyrum esculentum*, DFR_N, active in all tissues, reduces both DHK and DHM at the same rate, while the second DFR enzyme with a V134 residue, expressed only in roots and seeds, exhibits high specificity towards DHQ (62). In Solanales, *Petunia hybrida* and *Nicotiana tabacum* FLS_H catalyze both DHK and DHQ with minimal activity towards DHM (53,115), whereas their DFR_D is specific in reducing 3,4 di- and 3,4,5 tri-hydroxylated substrates but cannot reduce DHK (50,51). *Gentiana triflora* predominantly produces blue flowers and generally have cyanidin, delphinidin, and quercetin derivatives (116).

In our study, we observed that all monocots were mostly FLS_F type, and some were FLS_Y type, while their DFR were predominantly DFR_N type. These enzyme assay studies suggest that in

monocots, FLS_F generally prefers DHK while DFR_N prefers double or triple hydroxylated dihydroflavonol substrates. This difference in substrate preferences could help in avoiding competition between the two enzymes. In dicots, FLS_H is observed in the majority of plants, but in fewer plants with multiple FLS sequences, FLS_Y is also observed. While DFR is more diverse in dicots, some have only DFR_N or only DFR_D, while others have DFR_N along with either DFR_D or DFR_A candidates. The prior studies collectively suggest that FLS_H in dicots can utilize both DHK and DHQ as substrate but may have a preference for one over the other in different plants, implying that the H132 residue remains more neutral between DHK and DHQ. The biosynthesis of DHM requires the presence of flavonoid-3',5'-hydroxylase (F3'5'H) activity, which is rare. Since blue flower color can attract bee pollinators, it is only found in certain angiosperm lineages (117). DHQ and DHM (if *F3'5'H* is present) appear to be the more preferred substrate of DFR in most cases. Both in monocots and dicots, double and triple hydroxylated substrates seem to be more preferred by DFR, aligning with the notion that cyanidin is the most abundant and widely distributed floral anthocyanin (118). While substrate-specifying residues may play a role in identifying the specificity of FLS and DFR enzymes, it is plausible that substrate preference may not be exclusively determined by the H132 and N133 residues, respectively, in all plants. Given the rapid progress in protein structure prediction (119), it might be possible to quantify the contribution of additional residues in the near future through integration of all available data.

Functional divergence during FLS evolution

In the plant kingdom, the 2-oxoglutarate-dependent dioxygenase (2-ODD) superfamily is one of the largest enzyme families only second to cytochrome P450s (CYPs). Within this superfamily, the plant 2-ODD family can be categorized into three classes: DOXA, DOXB, and DOXC, with all enzymes involved in flavonoid biosynthesis falling into DOXC (120). These 2-ODD enzymes

are believed to have originated from a common ancestor before the emergence of land plants, and then underwent species-specific evolution in response to diverse environmental conditions (120). The evolution of FLS and ANS is likely to have occurred after the emergence of F3H during seed plant evolution (120). A noteworthy observation similar to that reported by Wang et al. (2019) (121), is the placement of ginkgo 2-ODD genes (*GbFLS*, *GbANS*, and *GnF3H*) into FLS, ANS, and F3H clades, respectively, indicating the divergence of F3H, FLS, and ANS preceded the separation of gymnosperms and angiosperms.

A comprehensive analysis of the functional FLS clade revealed a deep duplication in the evolutionary history of the FLS gene. We found that the FLS could be split into four different subgroups, gymnosperm sequences, aFLS, monocot sequences, and dicot sequences. The aFLS sequences consisting dicot sequences were more similar to gymnosperm FLSs than to the monocot and dicot FLS sequences in the distinct latter clades suggesting the duplication of ancestral FLS before the divergence of monocots and dicots. The aFLS clade contains FLS-like sequences from basal angiosperms and dicots and some previously characterized dicot FLS sequences. They have a Y132 residue and an aliphatic amino acid residue at position 295, mostly valine. Some of the FLS sequences in the aFLS clade are previously described. For example, the *Fagopyrum tataricum FtFLS2* is mainly expressed in roots, flowers and immature seeds and is upregulated by exogenous application of salicylic acid and NaCl and not affected by abscisic acid (122–124). However, activity analysis of *FtFLS2* was not performed. *Camellia sinensis CsFLSa* has high expression in buds but the buds mostly accumulate catechins. Heterologous expression of *CsFLSa* in tobacco failed, hence, FLS functionality remains unknown (125). In *Vitis vinifera,* FLS5 has very high transcript levels but the quercetin accumulation coincides with the transcription of FLS4 (H132 residue) and not FLS5. The shaded berries accumulate as much *FLS5* mRNA as the control berries while the *FLS4* (H132 residue) mRNA is not accumulated at all (126). *AcFLS* is expressed in almost all plant parts and highest abundance is seen in flowers. The maximum level of *AcFLS* mRNA level was recorded

six hours after wounding (127). We observed in our study that *Lotus japonicus* FLS4 has a low overall transcript level. No previously described aFLS sequence has been tested to show *in vivo* FLS activity. The position of aFLS clade separate from the canonical FLS clade indicates that members of the ancestral FLS clade might not be canonical FLS, but have undergone potential neofunctionalization since separation from the canonical FLS lineage.

The large dicot clade comprises a Brassicales order subclade and another clade comprising the rest of the dicots. Both clades show multiple duplication patterns at a shallower level. We presume that sequences in the Brassicales clade have functions similar to their *A. thaliana* AtFLS1-AtFLS6 orthologs. Preuss et al (2009) demonstrated that only *At*FLS1 and *At*FLS3 exhibit functional FLS activity, with other AtFLSs identified as non-functional and AtFLS3 only showing activity under extended assay conditions (128). However, other studies only found strong evidence for FLS activity in the Brassicales FLS1 clade (85,104). The presence of the FLS2-FLS6 clade across various Brassicales species suggests a potential undiscovered function for these enzymes. Distinct E295 residue patterns are observed among the FLS clades; AtFLS1, AtFLS2, AtFLS3, and AtFLS6 have E/D295 residue, whereas AtFLS4 and AtFLS5 have other non-conserved residues at this position. The AtFLS5 clade features aliphatic amino acids at this position, and *AtFLS5* shows expression primarily in the roots - especially in the roots of seedlings (104). Enzyme assay results by Chua et al. (2018) indicate that the mutations H132Y and H132F show higher specific activity than wild-type *Arabidopsis* FLS1, with E295L mutation resulting in only 7% of WT activity. However, the study only used DHQ as a substrate, and DHK was not tested. Additional studies testing enzyme activity of mutant FLS (H132Y, H132F, and E295(V/L)) with different substrates would help confirm the function of these aFLSs and the substrate preferences of FLSs with different residues.

In the second subclade, an interesting evolutionary pattern is observed, where FLS sequences from some plant orders appear in two different clades, particularly noticeable in the Fagales, Malpighiales, Malvales, and Sapindales orders. This suggests a gene duplication during the

evolution of land plants probably before the split of the aforementioned orders followed by a significant divergence in FLS sequences in these orders due to functional requirements, leading to the formation of distinct FLS clades. Within the Malvales order, including cotton species, an additional duplication event was observed, originating from FLS_H duplication, followed by a subsequent substitution leading to the re-emergence of FLS_Y (H132Y). The exclusive presence of the Y residue at position 132 in the Malvales order indicates a potential evolutionary step to enhance substrate affinity. Other orders like Solanales, Rosales, and Fagales are only represented in a single FLS clade.

Although the *Citrus unshiu* FLS within the Sapindales I clade exhibits a significantly greater affinity for DHK compared to DHQ, it predominantly accumulates quercetin 3-O-rutinoside (rutin) (129). The authors of this study proposed the presence of more than one FLS in *C. unshiu* (110). In other *Citrus* species, like *C. sinensis*, *C. trifoliata* and *C. clementin*a, we observed two FLS candidates, one each in the Sapindales I and II clade. Our analysis and enzyme assay data from previous studies suggest that FLS sequences within the first dicot subclade (Fagales I, Sapindales I, Malpighiales I, and Malvales I) display a pronounced substrate specificity for DHK. In contrast, FLS sequences within the second dicot subclade show a preference for both DHK and DHQ and may have a higher specificity for DHQ (98,100,112,113,125,126,130).

ANS/LDOX, which catalyses the step immediately downstream of DFR, is closely related to FLS and shows 50-60% polypeptide sequence similarity. Besides the conserved 2-ODD family residues, ANS/LDOX and FLS have similar amino acid residues that have been implicated in the DHQ binding of FLS (Y132, F134, K202, F293, and E295 conserved ANS/LDOX residues based on position in AtFLS1) and show the same substrate/ascorbate binding residue, E230 in AtLDOX. Bifunctional ANS enzymes with FLS-like activity have been discovered in *G. biloba* (131), *O. sativa* (132), *A. thaliana* (60), and *M. truncatula* (133). *In vitro* studies have shown that ANS can convert leucocyanidin to cyanidin and also dihydroquercetin to quercetin. *At*LDOX (128) converted DHQ and DHK into quercetin and kaempferol, respectively, in the ratio of 1:0.7

suggesting a higher affinity for DHQ at least in *A. thaliana*. This suggests that if FLS is specific for DHK, ANS might be responsible for different patterns of accumulation of kaempferol and quercetin observed in different plants. While DFR is more likely to accept DHQ, the ANS also might be more adapted to produce leucocyanidins.

Functional divergence during DFR evolution

Short-chain dehydrogenases/reductases (SDR) is an NADP-dependent enzyme superfamily ubiquitous across all life forms. DFR, belonging to the SDR108E enzyme family, converts dihydroflavonols (DHK, DHQ, and DHM) into colorless leucoanthocyanidins. These leucoanthocyanidins subsequently undergo further modifications to produce a wide array of pigments such as anthocyanins and proanthocyanidins (also known as tannins).

Cinnamoyl-CoA reductases (CCRs), involved in lignin biosynthesis, likely originated early in land plants and diverged distinctly from other SDR108E family (134), as seen by the distinct clustering of CCRs with DFR and ANR in our phylogenetic analysis. *DeDFR1*, and *DeDFR2* in *Dryopteris erythrosora*, *GbDFR1*, *GbDFR2*, and *GbDFR3* in *Ginkgo biloba*, and *CsDFRc* in *Camellia sinensis* were positioned outside the clade comprising functionally characterized DFRs. Fern DFRs, showing both *in vitro* and *in vivo* DFR activity, displayed a unique R residue at position 133, exhibiting substrate specificity for DHK and DHQ but an inability to catalyze DHM (135). The positioning of these fern DFRs outside the DFR and ANR clade suggests independent evolution of DFRs in ancient land plants. Notably, ginkgo DFRs (136), questioned by Halbwirth et al (2019) (43) for their authenticity, were found within TKPR-like clades in our analysis, suggesting their classification as DFR-like sequences rather than *bona fide* DFRs. *CsDFRc*, though reported to have DFR activity (125), also raised doubts about its categorization as a genuine DFR. Further investigation revealed that the clade containing *CsDFRc* also

harbored an *Arabidopsis* sequence, AT4G27250, named as ABA HYPERSENSITIVE 2 (ABH2) in a study by Weng et al (2016) (137). This *ABH2* is the long-sought after phaseic acid reductase (PAR), which are DFR-like enzymes involved in abscisic acid catabolism. This implies that *CsDFRc* might lean more towards a PAR functionality than a canonical DFR.

A thorough examination of the clade comprising functionally characterized DFRs revealed primarily three distinct DFR types determined by the amino acid residue at position 133 - DFR_N, DFR_D, and DFR_A. Our phylogenetic analysis indicates that DFR underwent duplication events, likely associated with N133D and N133A substitutions. These duplications and substitutions led to the acquisition of novel or enhanced functions, characterized by narrower substrate specificity, compared to the non-specific DFR_N. Until now, DFR_A was considered exclusive to the *Fragaria* genus; however, our detailed exploration of various plant species, including plants of *Spirodela* genus (*S. intermedium* and *S. polyrhiza*) and *Begonia peltafolia*, has revealed the presence of DFR_A in these species. This discovery suggests a more widespread distribution of DFR_A and multiple independent origins, although the study's power is limited by available plant sequences in public repositories.

While we could not identify DFR in gymnosperms, monocot and dicot DFRs in angiosperms formed distinct clades. Notably, two monocot clades were observed—one dominated by DFR_N sequences and another smaller one situated at the root of the dicot clade, harboring non-DFR_N sequences of monocots. These sequences exhibited amino acid diversity at position 133, such as cysteine (C) in *Musa acuminata*, alanine (A) in *Spirodela polyrrhiza*, and t*hreonine* (T) in *Dioscorea alata* and *D. rotundata*. While cysteine and threonine share similar properties with alanine, predictions about their performance, such as a preference for DHK, remain speculative. Given the lack of enzyme assay studies for these proteins, the functional status of this clade remains an open question.

Our analysis identified DFRs that do not conform to DFR_N, DFR_D, or DFR_A types, albeit in limited numbers. Few such DFRs have already been described, like *Lotus japonicus DFR1*, and

Ipomoea nil DFR-A, featuring residues S and H at position 133, respectively, were found to lack functional DFR activity. Whether these have undergone neo-functionalization or are pseudogenes remains unknown. Others, like monocot *Anthurium andraeanum* DFR, with a S133 residue, exhibit high activity towards DHK, while *Vaccinium macrocarpon DFR1-1*, *DFR2-2*, and *Fagopyrum esculentum FtDFR2*, featuring a V133 residue, prefer DHQ over other dihydroflavonols. Additionally, we identified DFR sequences with residues not studied by Johnson et al, including *Musa acuminata* (C133), *Malus domestica* (I133), *Carica papaya* (S133), *Fagus sylvatica*, and *Quercus suber* (R133). Investigating residues beyond N, D, and A at position 133 could unveil further insights into DFR functionality and substrate specificity.

DFRs with known functional activity and specificity towards DHK, as observed in strawberries where DFR_A is active in a 'false fruit,' and in *A. andraeanum* where DFR_S is active in spathe leaf—a modified leaf—suggest that unconventional or modified structures could prompt the evolution of different DFRs. Additional DFR types may be discovered in the future within diverse structures of plants that assume the function of fruit, flower, or leaves.

FLS and *DFR* expression pattern contributes to competition mitigation across angiosperms

The exploration of gene expression across 43 angiosperm species revealed global patterns and offered insights into the competition mitigation between *FLS* and *DFR*. There is a predominant concentration of data points along the x and y axes, indicating that co-expression of *FLS* and *DFR* almost never occurred. This observation suggests that in the competition between FLS and DFR, only one of them is active, directing the substrate either towards flavonol or anthocyanins. Transcription factors might be responsible for causing an almost mutual exclusion of *FLS* and *DFR* expression For example, the presence of DFR_A, known for preferring DHK, alongside FLS_H,

also inclined towards DHK, was associated with the exclusive expression of only one the corresponding genes within each sample. Similar dynamics were observed for DFR_N and FLS_Y, where both might favor DHQ, providing empirical evidence for substrate competition dynamics. In other cases, like FLS_H and DFR_N or FLS_H and DFR_D, there are some samples where both genes are co-expressed. The preference of above-mentioned DFRs for DHQ and DHM, juxtaposed with FLS_H's affinity for DHK emphasizes the intricate substrate-specific regulation between the genes, thereby mitigating the competition.

From the gene expression analysis (Additional File8), we distilled five significant observations. Firstly, we observed diverse substrate preferences among multiple DFR/FLS copies, as seen in *Lotus japonicus* and various *Gossypium* species harboring both DFR_N and DFR_D, as well as FLS_H and FLS_Y, with DFR_N favoring DHK and DFR_D preferring DHQ. The FLS_Y in *L. japonicus* is an ancestral FLS (aFLS) while ones in *Gossypium* are the re-emerged FLS_Y. Interestingly, FLS_Y in both species exhibited almost negligible expression. Secondly, tissue-specific expression of *FLS* and *DFR* was evident, as seen in *Fragaria ananassa*, where *FLS* displayed high expression in white flowers while *DFR* was more expressed in red fruits. This tissue specific expression of *FLS* and *DFR* would mitigate competition between both in the same tissue. The separation might even take place at the single cell level, which could explain samples that show some activity of both genes. The analyzed RNA-seq samples are mixtures of different cell types thus FLS and DFR expression could be divergent, but would appear as co-expression, because RNA from multiple different cell types is mixed during extraction. Thirdly, a preference for flavonols or anthocyanins might also affect the metabolic flux. Legume species like *Vigna unguiculata*, *Phaseolus acutifolius*, *Glycine max*, and *Cicer arietinum* from Fabales orders are known to produce flavonoids and isoflavonoids since they have a physiological role in nodulation (138). They were found to have a greater gene expression ratio of *FLS* to *DFR*, while others like *Vitis vinifera* with a high anthocyanin content (139) have higher *DFR* to *FLS* expression ratio.

Fourthly, the presence of an alternate pigmentation pathway may also influence the *DFR* expression levels. The betalain-pigmented *Beta vulgaris* does not accumulate anthocyanins (36) and hence, shows no *DFR* expression. Similar patterns were observed in some high-carotenoid species like *Solanum lycopersicum* and *Carica papaya*, which exhibited minimal to no *DFR* transcript abundance. Lastly, we observed that the expression of the same *FLS*/*DFR* type in the same plant tissue with different phenotypes differed. For example, *Vigna unguiculata* green fruit showed lower expression of DFR_N, DFR_D, *F3'H* and *F3H* compared to red fruit, indicating increased substrate availability in red fruit. Despite this, FLS expression remained constant in both red and green fruit. This pattern was consistent in other plants, in white and red fruit of *Fragaria vesca*, green and red leaf of *Lactuca sativa* and white, red and orange flowers of *Dianthus caryophyllus*, suggesting a complex regulation mechanism. Previous studies suggested feedback mechanisms in flavonoid biosynthesis regulation (140–142). Another possible explanation could be the influence of anthocyanin biosynthesis transcription factors like the MBW complex (143). Surprisingly, in *Theobroma cacao*, green fruit exhibited higher expression of *F3H*, *F3'H*, and DFR_N compared to mauve fruit. This apparent disparity in the transcript abundance and observed phenotype may be explained by a required expression of the anthocyanin biosynthesis genes before the anthocyanin pigmentation can become visible. Previous studies have reported that the accumulation of flavonols/anthocyanins may not always be directly correlated with the expression levels of *FLS*/*DFR* (125,144)This can be explained because the accumulation of anthocyanins is preceded by the expression of the required biosynthesis genes (145,146) ,i.e., enzymes need to be available for anthocyanins to be produced.

Power and limitations of big data upcycling

The wealth of available transcriptomic data offers unprecedented opportunities for efficient data upcycling and comprehensive cross-species comparison. However, it is crucial to acknowledge that many datasets in the Sequence Read Archive (SRA) lack comprehensive metadata, and in some instances, samples may suffer from cryptic or mislabeling issues, posing challenges for data upcycling (147). Therefore, careful consideration, filtering, and validation of sample tissue information are imperative to ensure the reliability and robustness of data interpretation and subsequent analyses. Nonetheless, the benefits of data reuse in biology outweigh the challenges, especially with the surge in publicly available data and the affordability of sequencing (148). In this study, we harnessed transcriptomic and genomic data from 43 angiosperm-wide species to perform expression analysis. We also utilized vast amounts of genomic data available for a large number of species and employed tools like KIPEs3 to identify functional flavonoid pathway genes with high specificity and accuracy (68). Achieving this through wet lab methods would not only be laborious, time-consuming, and costly, but also impossible due to unavailability of the biological material. This underscores the importance of open data to enable efficient exploration of complex biological processes and to facilitate the discovery of novel insights.

Conclusion

The branching point of the flavonoid biosynthesis into flavonols, catalyzed by FLS, and anthocyanins/proanthocyanidins, catalyzed by DFR, is a pivotal step for metabolic flux control. The availability of massive plant sequence datasets and gene expression resources enabled the investigation of FLS and DFR across a wide spectrum of plant species. This transformative approach revealed global patterns that cannot be identified by analyzing individual gene

functions in one species through classical wet lab methods. Amino acid residues associated with substrate preferences, evolutionary patterns, and potential competition mitigation mechanisms were explored. These insights into competing branches responsible for producing coloured anthocyanins and non-coloured flavonols can guide molecular breeding or engineering of ornamental plants.

Declarations

Data Availability

Scripts developed for this project and additional datasets are available on GitHub (https://github.com/bpucker/DFR_vs_FLS). The count tables of all 43 species analyzed via RNA-seq are freely available through LeoPARD (https://doi.org/10.24355/dbbs.084-202306231402-0) (149).

Funding

NC received a DAAD-Kospie scholarship for her research stay at TU Braunschweig which enabled her to conduct most of the presented analyses. NC's current position is funded by the DFG - Project number 516566273.

Competing interest

The authors declare that they do not have any competing interests.

Authors' contribution

NC and BP designed the experiments, collected data, wrote scripts for the data analysis, generated figures, and wrote the manuscript. All authors read and approved the final version of this manuscript.

Acknowledgements

We thank the research group Plant Biotechnology and Bioinformatics at TU Braunschweig for excellent discussions. This work was supported by the BMBF-funded de.NBI Cloud within the

German Network for Bioinformatics Infrastructure (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We would like to thank the entire de.NBI team for their excellent technical support. We acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig.

References

1. Winkel-Shirley B. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. Plant Physiol. 2001 Jun 1;126(2):485–93.
2. Koes R, Verweij W, Quattrocchio F. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci. 2005 May 1;10(5):236–42.
3. Paz-Ares J, Ghosal D, Saedler H. Molecular analysis of the C1-I allele from Zea mays: a dominant mutant of the regulatory C1 locus. EMBO J. 1990 Feb;9(2):315–21.
4. Martin C, Prescott A, Mackay S, Bartlett J, Vrijlandt E. Control of anthocyanin biosynthesis in flowers of Antirrhinum majus. Plant J. 1991 Jul;1(1):37–49.
5. Buer CS, Muday GK. The transparent testa4 Mutation Prevents Flavonoid Synthesis and Alters Auxin Transport and the Response of Arabidopsis Roots to Gravity and Light[W]. Plant Cell. 2004 May 12;16(5):1191–205.
6. Mo Y, Nagel C, Taylor LP. Biochemical complementation of chalcone synthase mutants defines a role for flavonols in functional pollen. Proc Natl Acad Sci. 1992 Aug;89(15):7213–7.
7. Winkel-Shirley B. Biosynthesis of flavonoids and effects of stress. Curr Opin Plant Biol. 2002 Jun;5(3):218–23.
8. Asen S, Stewart RN, Norris KH. Co-pigmentation of anthocyanins in plant tissues and its effect on color. Phytochemistry. 1972 Mar;11(3):1139–44.
9. Asen S, Stewart RN, Norris KH. Anthocyanin, flavonol copigments, and pH responsible for larkspur flower color. Phytochemistry. 1975 Dec;14(12):2677–82.
10. Emiliani J, Grotewold E, Falcone Ferreyra ML, Casati P. Flavonols protect Arabidopsis plants against UV-B deleterious effects. Mol Plant. 2013 Jul;6(4):1376–9.
11. Jiang N, Doseff AI, Grotewold E. Flavones: From Biosynthesis to Health Benefits. Plants. 2016 Jun;5(2):27.
12. Dixon RA, Sarnala S. Proanthocyanidin Biosynthesis—a Matter of Protection. Plant Physiol. 2020 Oct 1;184(2):579–91.
13. Todd JJ, Vodkin LO. Pigmented Soybean (Glycine max) Seed Coats Accumulate Proanthocyanidins during Development. Plant Physiol. 1993 Jun 1;102(2):663–70.
14. Nakabayashi R, Yonekura‐Sakakibara K, Urano K, Suzuki M, Yamada Y, Nishizawa T, et al. Enhancement of oxidative and drought tolerance in Arabidopsis by overaccumulation of antioxidant flavonoids. Plant J. 2014 Feb;77(3):367–79.
15. Schulz E, Tohge T, Zuther E, Fernie AR, Hincha DK. Natural variation in flavonol and anthocyanin metabolism during cold acclimation in *A rabidopsis thaliana* accessions: Flavonoids and cold acclimation. Plant Cell Environ. 2015 Aug;38(8):1658–72.
16. Becker C, Klaering HP, Kroh LW, Krumbein A. Cool-cultivated red leaf lettuce accumulates cyanidin-3-O-(6″-O-malonyl)-glucoside and caffeoylmalic acid. Food Chem. 2014 Mar;146:404–11.
17. Schulz E, Tohge T, Zuther E, Fernie AR, Hincha DK. Flavonoids are determinants of freezing tolerance and cold acclimation in Arabidopsis thaliana. Sci Rep. 2016 Sep 23;6(1):34027.
18. SU W, Zhang G, LI X, GU F, Shi B. Effect of light intensity and light quality on growth and total flavonoid accumulation of Erigeron breviscapus. Chin Tradit Herb Drugs. 1994;
19. Idris A, Linatoc AC, Bakar MFA, Ibrahim ZT, Audu Y. Effect of light quality and quantity

on the accumulation of flavonoid in plant species. J Sci Technol. 2018;10(3).
20. Feild TS, Lee DW, Holbrook NM. Why Leaves Turn Red in Autumn. The Role of Anthocyanins in Senescing Leaves of Red-Osier Dogwood. Plant Physiol. 2001 Oct 1;127(2):566–74.
21. falcone ferreyra M, Serra P, Casati P. Recent advances on the roles of flavonoids as plant protective molecules after UV and high light exposure. Physiol Plant. 2021 Aug 1;173.
22. Hanasaki Y, Ogawa S, Fukui S. The correlation between active oxygens scavenging and antioxidative effects of flavonoids. Free Radic Biol Med. 1994 Jun 1;16(6):845–50.
23. Agati G, Azzarello E, Pollastri S, Tattini M. Flavonoids as antioxidants in plants: Location and functional significance. Plant Sci. 2012 Nov 1;196:67–76.
24. Michalak A. Phenolic Compounds and Their Antioxidant Activity in Plants Growing under Heavy Metal Stress. Pol J Environ Stud. 2006 Jan 1;15:523–30.
25. Samanta A, Das G, Das S. Roles of flavonoids in Plants. Int J Pharm Sci Technol. 2011 Jan 1;6:12–35.
26. Subramanian S, Stacey G, Yu O. Distinct, crucial roles of flavonoids during legume nodulation. Trends Plant Sci. 2007 Jul;12(7):282–5.
27. Liu CW, Murray JD. The Role of Flavonoids in Nodulation Host-Range Specificity: An Update. Plants Basel Switz. 2016 Aug 11;5(3):33.
28. Brown DE, Rashotte AM, Murphy AS, Normanly J, Tague BW, Peer WA, et al. Flavonoids act as negative regulators of auxin transport in vivo in arabidopsis. Plant Physiol. 2001 Jun;126(2):524–35.
29. Besseau S, Hoffmann L, Geoffroy P, Catherine L, Pollet B, Legrand M. Flavonoid Accumulation in Arabidopsis Repressed in Lignin Synthesis Affects Auxin Transport and Plant Growth. Plant Cell. 2007 Feb 1;19:148–62.
30. Halbwirth H. The Creation and Physiological Relevance of Divergent Hydroxylation Patterns in the Flavonoid Pathway. Int J Mol Sci. 2010 Feb 4;11(2):595–621.
31. Yoder JI, Belzile F, Tong Y, Goldsbrough A. Visual markers for tomato derived from the anthocyanin biosynthetic pathway. Euphytica. 1994 Jan;79(3):163–7.
32. Shao D, Li Y, Zhu Q, Zhang X, Liu F, Xue F, et al. GhGSTF12, a glutathione S-transferase gene, is essential for anthocyanin accumulation in cotton (Gossypium hirsutum L.). Plant Sci. 2021 Apr 1;305:110827.
33. Eichenberger M, Schwander T, Hüppi S, Kreuzer J, Mittl PRE, Peccati F, et al. The catalytic role of glutathione transferases in heterologous anthocyanin biosynthesis. Nat Catal. 2023 Oct;6(10):927–38.
34. Zhao ZC, Hu GB, Hu FC, Wang HC, Yang ZY, Lai B. The UDP glucose: flavonoid-3-O-glucosyltransferase (UFGT) gene regulates anthocyanin biosynthesis in litchi (Litchi chinesis Sonn.) during fruit coloration. Mol Biol Rep. 2012 Jun;39(6):6409–15.
35. Francis FJ. Food colorants: anthocyanins. Crit Rev Food Sci Nutr. 1989;28(4):273–314.
36. Tanaka Y, Sasaki N, Ohmiya A. Biosynthesis of plant pigments: Anthocyanins, betalains and carotenoids. Plant J Cell Mol Biol. 2008 Jun 1;54:733–49.
37. Davies KM, Schwinn KE, Deroles SC, Manson DG, Lewis DH, Bloor SJ, et al. Enhancing anthocyanin production by altering competition for substrate between flavonol synthase and dihydroflavonol 4-reductase. Euphytica. 2003 Jun 1;131(3):259–68.
38. Holton TA, Brugliera F, Tanaka Y. Cloning and expression of flavonol synthase from Petunia hybrida. Plant J. 1993;4(6):1003–10.
39. Davies K, Winefield C, Lewis D, Nielsen K, Bradley M, Schwinn K, et al. Research into control of flower colour and flowering time in Eustoma grandiflorum (Lisanthus). Flower Newsl. 1997 Jan 1;23:24–32.
40. Luo P, Ning G, Wang Z, Shen Y, Jin H, Li P, et al. Disequilibrium of Flavonol Synthase and Dihydroflavonol-4-Reductase Expression Associated Tightly to White vs. Red Color Flower Formation in Plants. Front Plant Sci [Internet]. 2016 [cited 2023 Jul 7];6. Available

from: https://www.frontiersin.org/articles/10.3389/fpls.2015.01257
41. Nielsen K, Deroles SC, Markham KR, Bradley MJ, Podivinsky E, Manson D. Antisense flavonol synthase alters copigmentation and flower color in lisianthus. Mol Breed. 2002 Dec 1;9(4):217–29.
42. Tanaka Y, Brugliera F. Flower colour and cytochromes P450. Philos Trans R Soc B Biol Sci. 2013 Feb 19;368(1612):20120432.
43. Halbwirth H, Miosic S, Milosevic M, Nitarska D, Thill J, Stich K, et al. Re-investigating substrate specificity of dihydroflavonol 4-reductase with respect to the B-ring hydroxylation pattern of substrates. Acta Hortic. 2019 Jul;(1242):889–98.
44. Stracke R, Ishihara H, Huep G, Barsch A, Mehrtens F, Niehaus K, et al. Differential regulation of closely related R2R3-MYB transcription factors controls flavonol accumulation in different parts of the Arabidopsis thaliana seedling: TFs of A. thaliana R2R3-MYB subgroup 7. Plant J. 2007 Apr 25;50(4):660–77.
45. Schilbert HM, Glover BJ. Analysis of flavonol regulator evolution in the Brassicaceae reveals MYB12, MYB111 and MYB21 duplications and MYB11 and MYB24 gene loss. BMC Genomics. 2022 Aug 19;23(1):604.
46. Battat M, Eitan A, Rogachev I, Hanhineva K, Fernie A, Tohge T, et al. A MYB Triad Controls Primary and Phenylpropanoid Metabolites for Pollen Coat Patterning. Plant Physiol. 2019 May;180(1):87–108.
47. Zhang X, He Y, Li L, Liu H, Hong G. Involvement of the R2R3-MYB transcription factor MYB21 and its homologs in regulating flavonol accumulation in Arabidopsis stamen. Hancock R, editor. J Exp Bot. 2021 May 28;72(12):4319–32.
48. Xu W, Dubos C, Lepiniec L. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. Trends Plant Sci. 2015 Mar;20(3):176–85.
49. LaFountain AM, Yuan Y. Repressors of anthocyanin biosynthesis. New Phytol. 2021 Aug;231(3):933–49.
50. Johnson ET, Ryu S, Yi H, Shin B, Cheong H, Choi G. Alteration of a single amino acid changes the substrate specificity of dihydroflavonol 4-reductase: Substrate specificity of DFR. Plant J. 2001 Dec 23;25(3):325–33.
51. Forkmann G, Ruhnau B. Distinct Substrate Specificity of Dihydroflavonol 4-Reductase from Flowers of Petunia hybrida. Z Für Naturforschung C. 1987 Oct 1;42(9–10):1146–8.
52. Li Y, Liu X, Cai X, Shan X, Gao R, Yang S, et al. Dihydroflavonol 4-Reductase Genes from Freesia hybrida Play Important and Partially Overlapping Roles in the Biosynthesis of Flavonoids. Front Plant Sci [Internet]. 2017 Mar 28 [cited 2023 Jul 7];8. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2017.00428/full
53. Gerats AGM, De Vlaming P, Doodeman M, Al B, Schram AW. Genetic control of the conversion of dihydroflavonols into flavonols and anthocyanins in flowers of Petunia hybrida. Planta. 1982 Aug;155(4):364–8.
54. Beld M, Martin C, Huits H, Stuitje AR, Gerats AGM. Flavonoid synthesis in Petunia hybrida: partial characterization of dihydroflavonol-4-reductase genes. Plant Mol Biol. 1989 Nov;13(5):491–502.
55. Helariutta Y, Elomaa P, Kotilainen M, Griesbach RJ, Schr◆der J, Teeri TH. Chalcone synthase-like genes active during corolla development are differentially expressed and encode enzymes with different catalytic properties in Gerbera hybrida (Asteraceae). Plant Mol Biol. 1995 Apr;28(1):47–60.
56. Xie DY, Jackson LA, Cooper JD, Ferreira D, Paiva NL. Molecular and Biochemical Analysis of Two cDNA Clones Encoding Dihydroflavonol-4-Reductase from *Medicago truncatula*. Plant Physiol. 2004 Mar 1;134(3):979–94.
57. Shimada N, Sasaki R, Sato S, Kaneko T, Tabata S, Aoki T, et al. A comprehensive analysis of six dihydroflavonol 4-reductases encoded by a gene cluster of the Lotus japonicus genome. J Exp Bot. 2005 Sep 1;56(419):2573–85.

58. Petit P, Granier T, d'Estaintot BL, Manigand C, Bathany K, Schmitter JM, et al. Crystal Structure of Grape Dihydroflavonol 4-Reductase, a Key Enzyme in Flavonoid Biosynthesis. J Mol Biol. 2007 May 18;368(5):1345–57.
59. Chua CS, Biermann D, Goo KS, Sim TS. Elucidation of active site residues of Arabidopsis thaliana flavonol synthase provides a molecular platform for engineering flavonols. Phytochemistry. 2008 Jan;69(1):66–75.
60. Wilmouth RC, Turnbull JJ, Welford RWD, Clifton IJ, Prescott AG, Schofield CJ. Structure and Mechanism of Anthocyanidin Synthase from Arabidopsis thaliana. Structure. 2002 Jan;10(1):93–103.
61. Østergaard L, Lauvergeat V, Næsted H, Mattsson O, Mundy J. Two differentially regulated Arabidopsis genes define a new branch of the DFR superfamily. Plant Sci. 2001 Feb;160(3):463–72.
62. Katsu K, Suzuki R, Tsuchiya W, Inagaki N, Yamazaki T, Hisano T, et al. A new buckwheat dihydroflavonol 4-reductase (DFR), with a unique substrate binding structure, has altered substrate specificity. BMC Plant Biol. 2017 Dec;17(1):239.
63. Miosic S, Thill J, Milosevic M, Gosch C, Pober S, Molitor C, et al. Dihydroflavonol 4-Reductase Genes Encode Enzymes with Contrasting Substrate Specificity and Show Divergent Gene Expression Profiles in Fragaria Species. PLOS ONE. 2014 Nov 13;9(11):e112707.
64. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40(D1):D1178–86.
65. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D67–72.
66. Pucker B. script_collection [Internet]. 2023 [cited 2023 Jul 21]. Available from: https://github.com/bpucker/script_collection
67. Pucker B, Reiher F, Schilbert HM. Automatic Identification of Players in the Flavonoid Biosynthesis with Application on the Biomedicinal Plant Croton tiglium. Plants. 2020 Sep;9(9):1103.
68. Rempel A, Choudhary N, Pucker B. KIPEs3: Automatic annotation of biosynthesis pathways. PLOS ONE. 2023 Nov 16;18(11):e0294342.
69. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr;30(4):772–80.
70. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992 Nov 15;89(22):10915–9.
71. Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. Kelso J, editor. Bioinformatics. 2017 Jun 15;33(12):1886–8.
72. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.
73. 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. Teeling E, editor. Mol Biol Evol. 2020 May 1;37(5):1530–4.
74. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. Poon AFY, editor. PLoS ONE. 2010 Mar 10;5(3):e9490.
75. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Battistuzzi FU, editor. Mol Biol Evol. 2021 Jun 25;38(7):3022–7.
76. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.
77. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. 2011 Jan;39(Database issue):D19–21.

78. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525–7.
79. Pucker B, Iorizzo M. Apiaceae FNS I originated from F3H through tandem gene duplication. Remington DL, editor. PLOS ONE. 2023 Jan 19;18(1):e0280155.
80. Pucker B. CoExp [Internet]. 2022 [cited 2023 Nov 8]. Available from: https://github.com/bpucker/CoExp
81. PBBtools/collection/generate_subset.py at main · bpucker/PBBtools · GitHub [Internet]. [cited 2023 Oct 6]. Available from: https://github.com/bpucker/PBBtools/blob/main/collection/generate_subset.py
82. Pucker B, Walker-Hale N, Yim WC, Cushman J, Crum A, Yang Y, et al. Evolutionary blocks to anthocyanin accumulation and the loss of an anthocyanin carrier protein in betalain-pigmented Caryophyllales [Internet]. Evolutionary Biology; 2022 Oct [cited 2023 Jul 21]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.10.19.512958
83. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 2021 Apr 6;6(60):3021.
84. Stracke R, De Vos RCH, Bartelniewoehner L, Ishihara H, Sagasser M, Martens S, et al. Metabolomic and genetic analyses of flavonol synthesis in Arabidopsis thaliana support the in vivo involvement of leucoanthocyanidin dioxygenase. Planta. 2009 Jan 1;229(2):427–45.
85. Schilbert HM, Schöne M, Baier T, Busche M, Viehöver P, Weisshaar B, et al. Characterization of the Brassica napus Flavonol Synthase Gene Family Reveals Bifunctional Flavonol Synthases. Front Plant Sci [Internet]. 2021 [cited 2023 Oct 20];12. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2021.733762
86. The Angiosperm Phylogeny Group, 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 May 1;181(1):1–20.
87. Gebhardt Y, Witte S, Forkmann G, Lukačin R, Matern U, Martens S. Molecular evolution of flavonoid dioxygenases in the family Apiaceae. Phytochemistry. 2005 Jun 1;66(11):1273–84.
88. Johnson ET, Yi H, Shin B, Oh BJ, Cheong H, Choi G. Cymbidium hybrida dihydroflavonol 4-reductase does not efficiently reduce dihydrokaempferol to produce orange pelargonidin-type anthocyanins. Plant J. 1999;19(1):81–5.
89. Park S, Kim DH, Lee JY, Ha SH, Lim SH. Comparative Analysis of Two Flavonol Synthases from Different-Colored Onions Provides Insight into Flavonoid Biosynthesis. J Agric Food Chem. 2017 Jul 5;65(26):5287–98.
90. Park S, Kim DH, Park BR, Lee JY, Lim SH. Molecular and Functional Characterization of *Oryza sativa* Flavonol Synthase (OsFLS), a Bifunctional Dioxygenase. J Agric Food Chem. 2019 Jul 3;67(26):7399–409.
91. Falcone Ferreyra ML, Rius S, Emiliani J, Pourcel L, Feller A, Morohashi K, et al. Cloning and characterization of a UV-B-inducible maize flavonol synthase. Plant J. 2010;62(1):77–91.
92. Žilić S, Serpen A, Akıllıoğlu G, Gökmen V, Vančetović J. Phenolic Compounds, Carotenoids, Anthocyanins, and Antioxidant Capacity of Colored Maize (Zea mays L.) Kernels. J Agric Food Chem. 2012 Feb 8;60(5):1224–31.
93. Aoki H, Kuze N, Kato Y, Ei S, Inc F. Anthocyanin isolated from purple corn (Zea mays L). Foods Food Ingred J Jpn. 2001 Jan 1;199.
94. Jende-Strid B. Gene-enzyme relations in the pathway of flavonoid biosynthesis in barley. Theor Appl Genet. 1991 May 1;81(5):668–74.
95. Nyegaard Kristiansen K, Rohde W. Structure of the Hordeum vulgare gene encoding dihydroflavonol-4-reductase and molecular analysis of ant18 mutants blocked in flavonoid synthesis. Mol Gen Genet MGG. 1991 Nov 1;230(1):49–59.
96. Liu H, Lou Q, Ma J, Su B, Gao Z, Liu Y. Cloning and Functional Characterization of Dihydroflavonol 4-Reductase Gene Involved in Anthocyanidin Biosynthesis of Grape

Hyacinth. Int J Mol Sci. 2019 Jan;20(19):4743.
97. Leonard E, Yan Y, Chemler J, Matern U, Martens S, Koffas MAG. Characterization of dihydroflavonol 4-reductases for recombinant plant pigment biosynthesis applications. Biocatal Biotransformation. 2008 Jan 1;26(3):243–51.
98. Almeida JRM, D'Amico E, Preuss A, Carbone F, de Vos CHR, Deiml B, et al. Characterization of major enzymes and genes involved in flavonoid and proanthocyanidin biosynthesis during fruit development in strawberry (Fragaria ×ananassa). Arch Biochem Biophys. 2007 Sep 1;465(1):61–71.
99. Li P, Lei K, Liu L, Zhang G, Ge H, Zheng C, et al. Identification and functional characterization of a new flavonoid synthase gene MdFLS1 from apple. Planta. 2021 Apr 15;253(5):105.
100. Suzuki K ichi, Tsuda Sh, Fukui Y, Fukuchi-Mizutani M, Yonekura-Sakakibara K, Tanaka Y, et al. Molecular Characterization of Rose Flavonoid Biosynthesis Genes and Their Application in Petunia. Biotechnol Biotechnol Equip. 2000 Jan 1;14(2):56–62.
101. Fischer TC, Halbwirth H, Meisel B, Stich K, Forkmann G. Molecular cloning, substrate specificity of the functionally expressed dihydroflavonol 4-reductases from Malus domestica and Pyrus communis cultivars and the consequences for flavonoid metabolism. Arch Biochem Biophys. 2003 Apr 15;412(2):223–30.
102. Tanaka Y, Fukui Y, Fukuchi-Mizutani M, Holton TA, Higgins E, Kusumi T. Molecular Cloning and Characterization of Rosa hybrida Dihydroflavonol 4-reductase Gene. Plant Cell Physiol. 1995 Sep 1;36(6):1023–31.
103. Waki T, Mameda R, Nakano T, Yamada S, Terashita M, Ito K, et al. A conserved strategy of chalcone isomerase-like protein to rectify promiscuous chalcone synthase specificity. Nat Commun. 2020 Feb 13;11(1):870.
104. Owens DK, Alerding AB, Crosby KC, Bandara AB, Westwood JH, Winkel BSJ. Functional Analysis of a Predicted Flavonol Synthase Gene Family in Arabidopsis. Plant Physiol. 2008 Jul 8;147(3):1046–61.
105. Spribille R, Forkmann G. Conversion of Dihydroflavonols to Flavonols with Enzyme Extracts from Flower Buds of Matthiola incana R. Br. Z Für Naturforschung C. 1984 Aug 1;39(7–8):714–9.
106. Turnbull JJ, Nakajima J ichiro, Welford RWD, Yamazaki M, Saito K, Schofield CJ. Mechanistic Studies on Three 2-Oxoglutarate-dependent Oxygenases of Flavonoid Biosynthesis: ANTHOCYANIDIN SYNTHASE, FLAVONOL SYNTHASE, AND FLAVANONE 3β-HYDROXYLASE*. J Biol Chem. 2004 Jan 9;279(2):1206–16.
107. Bloor SJ, Abrahams S. The structure of the major anthocyanin in Arabidopsis thaliana. Phytochemistry. 2002 Feb 1;59(3):343–6.
108. Martens S, Teeri T, Forkmann G. Heterologous expression of dihydroflavonol 4-reductases from various plants. FEBS Lett. 2002 Nov 20;531(3):453–8.
109. Saito N, Tatsuzawa F, Nishiyama A, Yokoi M, Shigihara A, Honda T. Acylated cyanidin 3-sambubioside-5-glucosides in Matthiola incana. Phytochemistry. 1995 Mar;38(4):1027–32.
110. Wellmann F, Lukačin R, Moriguchi T, Britsch L, Schiltz E, Matern U. Functional expression and mutational analysis of flavonol synthase from *Citrus unshiu*: Flavonol synthase. Eur J Biochem. 2002 Aug;269(16):4134–42.
111. Aiguo Z, Ruiwen D, Cheng W, Cheng C, Dongmei W. Insights into the catalytic and regulatory mechanisms of dihydroflavonol 4-reductase, a key enzyme of anthocyanin synthesis in Zanthoxylum bungeanum. Tree Physiol. 2023 Jan 1;43(1):169–84.
112. Li C, Bai Y, Li S, Chen H, Han X, Zhao H, et al. Cloning, Characterization, and Activity Analysis of a Flavonol Synthase Gene *FtFLS1* and Its Association with Flavonoid Content in Tartary Buckwheat. J Agric Food Chem. 2012 May 23;60(20):5161–8.
113. Stich K, Eidenberger T, Wurst F, Forkmann G. Flavonol Synthase Activity and the Regulation of Flavonol and Anthocyanin Biosynthesis during Flower Development in

Dianthus caryophyllus L. (Carnation). Z Für Naturforschung C. 1992 Aug 1;47(7–8):553–60.
114. Jiang J. Cloning of flavonol synthase gene from Fagopyum dibotrys and its expression in Escherichia coli. Chin Tradit Herb Drugs. 2013;1974–8.
115. McCarthy EW, Landis JB, McCoy AG, Lawhorn AJ, Kurti A, Xu Y, et al. Homeologue differential expression in the flavonoid biosynthetic pathway underlies flower colour variation in natural and synthetic polyploids of Nicotiana tabacum (Solanaceae). Bot J Linn Soc. 2023 Oct 24;boad052.
116. Tanaka Y, Yonekura K, Fukuchi-Mizutani M, Fukui Y, Fujiwara H, Ashikari T, et al. Molecular and Biochemical Characterization of Three Anthocyanin Synthetic Enzymes from Gentiana triflora. Plant Cell Physiol. 1996 Jul 1;37(5):711–6.
117. Gottlieb OR. Micromolecular Evolution, Systematics and Ecology: An Essay into a Novel Botanical Discipline. Springer Science & Business Media; 2012. 183 p.
118. Harborne JB, Williams CA. Advances in flavonoid research since 1992. Phytochemistry. 2000 Nov 1;55(6):481–504.
119. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol. 2019 Nov;20(11):681–97.
120. Kawai Y, Ono E, Mizutani M. Evolution and diversity of the 2–oxoglutarate-dependent dioxygenase superfamily in plants. Plant J. 2014;78(2):328–43.
121. Wang Z, Wang S, Wu M, Li Z, Liu P, Li F, et al. Evolutionary and functional analyses of the 2-oxoglutarate-dependent dioxygenase genes involved in the flavonoid biosynthesis pathway in tobacco. Planta. 2019 Feb 1;249(2):543–61.
122. Yao Y, Sun L, Wu W, Wang S, Xiao X, Hu M, et al. Genome-Wide Investigation of Major Enzyme-Encoding Genes in the Flavonoid Metabolic Pathway in Tartary Buckwheat (Fagopyrum tataricum). J Mol Evol. 2021 Jun 1;89(4):269–86.
123. Li X, Kim YB, Kim Y, Zhao S, Kim HH, Chung E, et al. Differential stress-response expression of two flavonol synthase genes and accumulation of flavonols in tartary buckwheat. J Plant Physiol. 2013 Dec 15;170(18):1630–6.
124. Li X, Park NI, Kim YB, Kim HH, Park CH, Wu Q, et al. Accumulation of flavonoids and expression of flavonoid biosynthetic genes in tartary and rice-tartary buckwheat. Process Biochem. 2012 Dec 1;47(12):2306–10.
125. Jiang X, Shi Y, Fu Z, Li WW, Lai S, Wu Y, et al. Functional characterization of three flavonol synthase genes from Camellia sinensis: Roles in flavonol accumulation. Plant Sci. 2020 Nov;300:110632.
126. FUJITA A, GOTO-YAMAMOTO N, ARAMAKI I, HASHIZUME K. Organ-Specific Transcription of Putative Flavonol Synthase Genes of Grapevine and Effects of Plant Hormones and Shading on Flavonol Biosynthesis in Grape Berry Skins. Biosci Biotechnol Biochem. 2006 Jan 1;70(3):632–8.
127. Toh HC, Wang SY, Chang ST, Chu FH. Molecular cloning and characterization of flavonol synthase in Acacia confusa. Tree Genet Genomes. 2013 Feb 1;9(1):85–92.
128. Preuss A, Stracke R, Weisshaar B, Hillebrecht A, Matern U, Martens S. Arabidopsis thaliana expresses a second functional flavonol synthase. FEBS Lett. 2009 Jun 18;583(12):1981–6.
129. Moriguchi T, Kita M, Ogawa K, Tomono Y, Endo T, Omura M. Flavonol synthase gene expression during citrus fruit development. Physiol Plant. 2002;114(2):251–8.
130. Kou M, Li C, Song W, Shen Y, Tang W, Zhang Y, et al. Identification and functional characterization of a flavonol synthase gene from sweet potato [Ipomoea batatas (L.) Lam.]. Front Plant Sci. 2023 May 10;14:1181173.
131. Xu F, Cheng H, Cai R, Li LL, Chang J, Zhu J, et al. Molecular Cloning and Function Analysis of an Anthocyanidin Synthase Gene from Ginkgo biloba, and Its Expression in Abiotic Stress Responses. Mol Cells. 2008 Dec 31;26(6):536–47.
132. Reddy AM, Reddy VS, Scheffler BE, Wienand U, Reddy AR. Novel transgenic rice

overexpressing anthocyanidin synthase accumulates a mixture of flavonoids leading to an increased antioxidant potential. Metab Eng. 2007 Jan 1;9(1):95–111.
133. Pang Y, Peel GJ, Wright E, Wang Z, Dixon RA. Early Steps in Proanthocyanidin Biosynthesis in the Model Legume Medicago truncatula. Plant Physiol. 2007 Nov;145(3):601–15.
134. Moummou H, Kallberg Y, Tonfack LB, Persson B, van der Rest B. The Plant Short-Chain Dehydrogenase (SDR) superfamily: genome-wide inventory and diversification patterns. BMC Plant Biol. 2012 Nov 20;12(1):219.
135. Chen X, Liu W, Huang X, Fu H, Wang Q, Wang Y, et al. Arg-type dihydroflavonol 4-reductase genes from the fern Dryopteris erythrosora play important roles in the biosynthesis of anthocyanins. PLOS ONE. 2020 May 1;15(5):e0232090.
136. Hua C, Linling L, Shuiyuan C, Fuliang C, Feng X, Honghui Y, et al. Molecular Cloning and Characterization of Three Genes Encoding Dihydroflavonol-4-Reductase from Ginkgo biloba in Anthocyanin Biosynthetic Pathway. PLOS ONE. 2013 Aug 26;8(8):e72017.
137. Weng JK, Ye M, Li B, Noel JP. Co-evolution of Hormone Metabolism and Signaling Networks Expands Plant Adaptive Plasticity. Cell. 2016 Aug;166(4):881–93.
138. Aoki T, Akashi T, Ayabe S ichi. Flavonoids of Leguminous Plants: Structure, Biological Activity, and Biosynthesis. J Plant Res. 2000 Dec;113(4):475–88.
139. Boss PK, Davies C, Robinson SP. Analysis of the Expression of Anthocyanin Pathway Genes in Developing Vitis vinifera L. cv Shiraz Grape Berries and the Implications for Pathway Regulation. Plant Physiol. 1996 Aug 1;111(4):1059–66.
140. Liao L, Vimolmangkang S, Wei G, Zhou H, Korban SS, Han Y. Molecular characterization of genes encoding leucoanthocyanidin reductase involved in proanthocyanidin biosynthesis in apple. Front Plant Sci [Internet]. 2015 [cited 2023 Dec 17];6. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2015.00243
141. Liu Y, Shi Z, Maximova S, Payne MJ, Guiltinan MJ. Proanthocyanidin synthesis in Theobroma cacao: genes encoding anthocyanidin synthase, anthocyanidin reductase, and leucoanthocyanidin reductase. BMC Plant Biol. 2013 Dec 5;13(1):202.
142. Lei T, Huang J, Ruan H, Qian W, Fang Z, Gu C, et al. Competition between FLS and DFR regulates the distribution of flavonols and proanthocyanidins in Rubus chingii Hu. Front Plant Sci [Internet]. 2023 [cited 2023 Dec 17];14. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2023.1134993
143. Li C, Yu W, Xu J, Lu X, Liu Y. Anthocyanin Biosynthesis Induced by MYB Transcription Factors in Plants. Int J Mol Sci. 2022 Jan;23(19):11701.
144. Takos AM, Jaffé FW, Jacob SR, Bogs J, Robinson SP, Walker AR. Light-Induced Expression of a MYB Gene Regulates Anthocyanin Biosynthesis in Red Apples. Plant Physiol. 2006 Nov 1;142(3):1216–32.
145. Zhang Y, Li Y, Li W, Hu Z, Yu X, Tu Y, et al. Metabolic and molecular analysis of nonuniform anthocyanin pigmentation in tomato fruit under high light. Hortic Res. 2019 May 1;6(1):1–21.
146. Jaakola L, Määttä K, Pirttilä AM, Törrönen R, Kärenlampi S, Hohtola A. Expression of Genes Involved in Anthocyanin Biosynthesis in Relation to Anthocyanin, Proanthocyanidin, and Flavonol Levels during Bilberry Fruit Development. Plant Physiol. 2002 Oct;130(2):729–39.
147. Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ. 2020 Sep 22;8:e9954.
148. Pucker B, Irisarri I, Vries J de, Xu B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quant Plant Biol. 2022 Jan;3:e5.
149. Pucker B, Choudhary N. Gene expression data sets of selected plant species [Internet]. Universitätsbibliothek Braunschweig; 2023 [cited 2023 Nov 9]. Available from: https://doi.org/10.24355/dbbs.084-202306231402-0

Additional Files

Additional File 1: References of the coding sequence data sets that were used in this study.
Additional File 2: Collection of ANR, CCR, DFR, and DFR-like sequences that were used for the phylogenetic analysis of DFR.
Additional File 3: Collection of F3H, FNSI, ANS, FLS, and FLS-like sequences that were used for the phylogenetic analyses of FLS.
Additional File 4: Phylogenetic trees of DFR. (1) DFR tree showing outgroups constructed by IQ-TREE based on a MAFFT alignment, (2a-f) DFR-specific trees with collapsed outgroups. Highlighting amino acid residue diversity at position 133 associated with substrate specificity. Gymnosperm, monocot and dicot species are denoted by light blue, light pink and light green colour stripes, respectively. Non-DFR sequences are represented by dashed gray branches while the functional DFRs are indicated by solid black branches. The color-coded scheme represents different residues: Asparagine (light purple), Aspartic acid (periwinkle blue), Alanine (deep purple), and other amino acids (grey) at position 133. The preferred substrate of the DFR type is written in brackets, DHK, dihydrokaempferol; DHQ, dihydroquercetin and DHM, dihydromyricetin. Distinct clusters of DFRs from major plant orders are labeled for reference. DFR genes characterized in previous studies are highlighted by a red asterisk at the start of the terminal branch.(2a) Constructed by IQ-TREE based on a MAFFT alignment, (2b) constructed by IQ-TREE based on a Muscle5 alignment, (2c) constructed by FastTree2 based on a MAFFT alignment, (2d) constructed by FastTree2 based on a Muscle5 alignment, (2e) constructed by MEGA based on a MAFFT alignment, and (2f) constructed by MEGA based on a Muscle5 alignment.
Additional File 5: References of previously characterized FLS and DFR included in phylogenetic tree analyses.
Additional File 6: Examination of codons at position 133 within the DFR_A coding sequence in *Fragaria vesca* and *Fragaria ananassa*, along with a comparison against closely related DFR_N sequences. The codons and corresponding amino acid at position 133 are highlighted in red. The color-coded scheme represents different residues: Asparagine (light purple), Alanine (deep purple), and other amino acids (grey) at position 133.
Additional File 7: Phylogenetic trees of FLS. (1) FLS tree showing outgroups constructed by IQ-TREE based on a MAFFT alignment, (2a-f) FLS-specific trees with collapsed outgroups. Highlighting amino acid residue diversity at position 132 associated with DHQ-substrate binding. Gymnosperm, monocot and dicot species are denoted by light blue, light pink and light green colour stripes, respectively. Non-FLS sequences are represented by dashed gray branches while the functional FLSs are indicated by solid black branches. The background colour highlights different presumably substrate-determining amino acid residues: histidine (pale yellow), phenylalanine (dark orange), and tyrosine (dark golden) at position 132. The hypothesized preferred substrate of the FLS type is written in brackets, DHK, dihydrokaempferol; DHQ, dihydroquercetin and DHM, dihydromyricetin. Distinct clusters of FLSs from major plant orders are labeled for reference. FLS genes characterized in previous studies are highlighted by a red asterisk at the start of the terminal branch. (2a) Constructed by IQ-TREE based on a MAFFT alignment, (2b) constructed by IQ-TREE based on a Muscle5 alignment, (2c) constructed by FastTree2 based on a MAFFT alignment, (2d) constructed by FastTree2 based on a Muscle5 alignment, (2e) constructed by MEGA based on a MAFFT alignment, and (2f) constructed by MEGA based on a Muscle5 alignment.
Additional File 8: Extended heatmap of 43 species depicting the transcript abundance of analyzed genes in different tissues. DFR* refers to DFR sequences where the residue 133 is not N, D or A. In *Musa acuminata* C133 residue, *Lotus japonicus* S133, and *Carica papaya* S133.