CoExpPhylo – A Novel Pipeline for Biosynthesis Gene Discovery

Nele Grünig[1] & Boas Pucker[1,2,*]

[1]Plant Biotechnology and Bioinformatics, Institute of Plant Biology & Braunschweig Integrated, Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany.

[2]Institute for Cellular and Molecular Botany (IZMB), University of Bonn, Kirschallee 1, 53115, Bonn, Germany

[*]Correspondence to: Boas Pucker, pucker@uni-bonn.de

ORCIDs:

NG: 0009-0004-1768-9489

BP: 0000-0002-3321-7471

Abstract

Background

The rapid advancement of sequencing technologies has drastically increased the availability of plant genomic and transcriptomic data, shifting the challenge from data generation to functional interpretation. Identifying genes involved in specialized metabolism remains difficult. While coexpression analysis is a widely used approach to identify genes acting in the same pathway or process, it has limitations, particularly in distinguishing genes coexpressed due to shared regulatory triggers from those directly involved in the same pathway. To enhance functional predictions, integrating phylogenetic analysis provides an additional layer of confidence by considering evolutionary conservation. Here, we introduce CoExpPhylo, a computational pipeline that systematically combines coexpression analysis and phylogenetics to identify candidate genes involved in specialized biosynthetic pathways across multiple species based on one to multiple bait gene candidates.

Results

CoExpPhylo systematically integrates coexpression information and phylogenetic signals to identify candidate genes involved in specialized biosynthetic pathways. The pipeline consists of multiple computational steps: (1) species-specific coexpression analysis, (2) local sequence alignment to identify orthologs, (3) clustering of candidate genes into Orthologous Coexpressed Groups (OCGs), (4) functional annotation, (5) global sequence alignment, (6) phylogenetic tree generation, and optionally (7) visualization. The workflow is highly customizable, allowing users to adjust correlation thresholds, filtering parameters, and annotation sources. Benchmarking CoExpPhylo on multiple pathways, including the biosynthesis of anthocyanins, proanthocyanidins, and flavonols, as well as lutein and zeaxanthin, confirmed its ability to recover known genes while also suggesting novel candidates.

Conclusion

CoExpPhylo provides a systematic framework for identifying candidate genes involved in the specialized metabolism. By integrating coexpression data with phylogenetic clustering, it facilitates the discovery of both conserved and lineage-specific genes. The resulting OCGs offer a strong foundation for further experimental validation, bridging the gap between computational predictions and functional characterization. Future improvements, such as incorporating multi-species reference databases and refining clustering for large gene families, could further enhance its resolution. Overall, CoExpPhylo represents a valuable tool for accelerating pathway elucidation and advancing our understanding of specialized metabolism in plants.

Keywords

Phylogeny, Orthology, Functional annotation, Pathways, Coexpression, Flavonoids, Carotenoids, Biosynthesis route, Specialized metabolism

Background

The continuous advancements in sequencing technologies have led to a drastic reduction in the cost of genome sequencing and the availability of thousands of plant genome sequences [1]. As a result, the primary challenge in genomics has shifted from generating sequencing data to efficiently analyzing and interpreting the vast amount of publicly available data [2]. Databases such as the Sequence Read Archive (SRA) and the National Center for Biotechnology Information (NCBI) provide an extensive collection of sequencing datasets, offering unprecedented opportunities for comparative and functional genomics [3]. While identifying genes structurally becomes more feasible through the integration of transcriptomic data and the facilitated transfer of gene annotations between species through programs such as GeMoMa [4], functionally annotating genes remains a major challenge [5]. In plants, understanding the roles of genes in the complex specialized metabolism is a particular challenge and heavily dependent on information about the biochemical functions of gene products.

One challenge in finding genes belonging to a certain metabolic pathway in plants lies in their scatteredness across the genome. Genes involved in pathways that lead to specialized metabolites are rarely physically clustered [6]. Instead, transcriptional control by a shared transcription factor or transcription factor complex activates all genes involved in a given pathway [7–10]. Notable exceptions are biosynthesis gene clusters, including those reported for withanolide biosynthesis in *Solanaceae*, noscapine biosynthesis in *Papaver somniferum,* or terpenoid biosynthesis pathways such as thalianol, marneral, tirucalladienol, and arabidol in *Arabidopsis thaliana*, among a limited number of other known cases [11–14]. Hence, finding all genes participating in a given pathway based on already known genes merely based on their genomic position can be challenging. Here, coexpression analyses can come to the rescue: Naturally, genes that are required for the same pathway show a similar activity as the encoded enzymes are required simultaneously. This is realized by transcription factors that often regulate the expression of several genes participating in a pathway. Genes whose expression patterns have a correlation to the expression of already known genes of the pathway of interest, might also participate in that pathway due to guilt-by-association [15]. By using Spearman's rank correlation method, which also detects non-linear correlations, even transcription factors involved in the regulation of a target pathway can be identified. However, the results should be taken with caution, as guilt-by-association has its limits and especially genes that are expressed constantly, such as genes involved in the central metabolism or genes coexpressed due to shared regulatory triggers rather than direct functional relationships can produce misleading results [16]. Over the years, a range of computational strategies has been developed to overcome these challenges. Early network-based approaches such as AraNet v2 leverage cofunctional gene networks to predict pathway membership, but often lack resolution among closely related paralogs or across diverse species, as it contains information only about 28 species [17]. More recently, supervised machine-learning models have been applied to plant pathway gene discovery – for example, a support-vector-machine model trained on over 10 000 genomic and network-based features (conservation, protein domains, duplication

events, epigenetic marks, expression and network metrics) to distinguish specialized- from general-metabolism genes in *Arabidopsis thaliana* [18], and RafSee, a Random Forest classifier that integrates sequence-derived, evolutionary, and epigenetic features with coexpression metrics to prioritize candidate genes for specific pathways [19]. These approaches, however, depend on large, carefully curated training sets and do not explicitly incorporate evolutionary context, limiting their applicability to new species or poorly annotated pathways.

To address this limitation, integrating phylogenetic analysis with coexpression data provides a more reliable strategy for identifying functionally relevant genes belonging to one biosynthetic pathway. Phylogenetic approaches help discern evolutionary conservation and functional divergence, enhancing the accuracy of gene function predictions.

A well studied model system is required as proof-of-concept for a new computational approach for the identification of genes in a biosynthesis pathway based on integrated coexpression and phylogenetic integration. The flavonoid biosynthesis with a set of different compounds produced by individual branches is a prime example [20, 21]. Anthocyanins confer colors to fruits and flowers [22], proanthocyanidins (PAs) are responsible for seed coat pigmentation [23, 24] and flavonols are known as UV-response compounds [25, 26] and have been associated with regulatory functions [27]. Due to a plethora of modification reactions catalyzed by numerous (promiscuous) enzymes, there are thousands of different flavonoid derivatives in plants [28, 29]. Despite decades of intense research on the flavonoid biosynthesis [20, 21, 30], there is still potential for new discoveries as demonstrated by recent publications [31, 32].

A second benchmark pathway is the carotenoid biosynthesis, which represents another well-studied yet complex system of plant specialized metabolism. Carotenoids are C40 isoprenoid pigments with essential roles in photosynthesis and photoprotection, as well as in the coloration of flowers and fruits [33]. Lutein and zeaxanthin, two major xanthophylls, are prominent products of this pathway. The extensive research and the identification of many core enzymes [34–36] makes carotenoid biosynthesis, alongside flavonoids, an excellent model for benchmarking computational approaches for gene identification.

Here, we present CoExpPhylo as a novel implementation for automatic integration of coexpression and phylogenetic analyses for the identification of players in a biosynthesis pathway. Starting from minimal prior knowledge about some genes involved in the biosynthesis of a metabolite of interest, CoExpPhylo enables the identification of candidate genes associated with the biosynthesis of that metabolite. The potential is demonstrated by analyses on the biosynthesis of the flavonoids anthocyanin, proanthocyanidin, and flavonol as well as the carotenoids lutein and zeaxanthin.

Implementation

CoExpPhylo was written in Python3 using SciPy v1.13.0 [37], NumPy v1.26.4 [38], Networkx v3.3 [39] and Plotly v5.22.0 [40] as well as the utility GNU parallel 20210822 [41] through bash

scripting. CoExpPhylo utilizes different external tools: Several local alignments are conducted via DIAMOND v2.0.14.152 [42], global alignments are executed using MAFFT v7.490 [43, 44] or optionally by MUSCLE v3.8.1551 [45] and phylogenetic trees are inferred by FastTree v2.1.11 [46, 47], RAxML-NG v1.2.2 [48] or IQ-TREE v2.0.7 [49]. Optionally, the final tree files can be uploaded to iTOL [50].

Input data collection

The analysis is conducted based on transcriptomic datasets of 240 species from various orders (Figure 1, Additional File 1) [51, 52]. Briefly, the coding sequences were collected from Phytozome (https://phytozome-next.jgi.doe.gov), NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank), and species-specific websites. The RNA-seq data was collected from the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) [3] using fastq-dump (https://github.com/ncbi/sra-tools) and processed as described in the methodology of the dataset publications: Count tables and gzip-compressed FASTQ files were processed using kallisto v0.44 [53] and subsequently merged into a single count table per species with custom Python scripts [54]. All values in the count tables are Transcripts Per Million (TPM). Samples were excluded, if the total number of reads was below one million or the read distribution did not match RNA-seq expectation i.e. at least 20% of reads assigned to the most abundant 100 transcripts. This was handled using the Python script filter_RNAseq_samples.py v0.4 (https://github.com/bpucker/CoExp). Each annotation file was generated with the Python script construct_anno.py v0.1 (https://github.com/bpucker/PBBtools). For each dataset, the gene IDs in each file were cleaned as any special characters were replaced. Only one representative transcript was retained per gene by selecting the one with the longest coding sequence (CDS). This established approach was chosen to reduce redundancy in the dataset to limit computational complexity. The isoform with the longest CDS often corresponds to the most functionally relevant isoform [55]. In cases where transcript identifiers were missing from the sequence headers (e.g. no transcript number or suffix indicating alternative isoforms), transcripts belonging to the same gene were determined and reduced via isoform_purger.py v0.21 (https://github.com/bpucker/PBBtools).

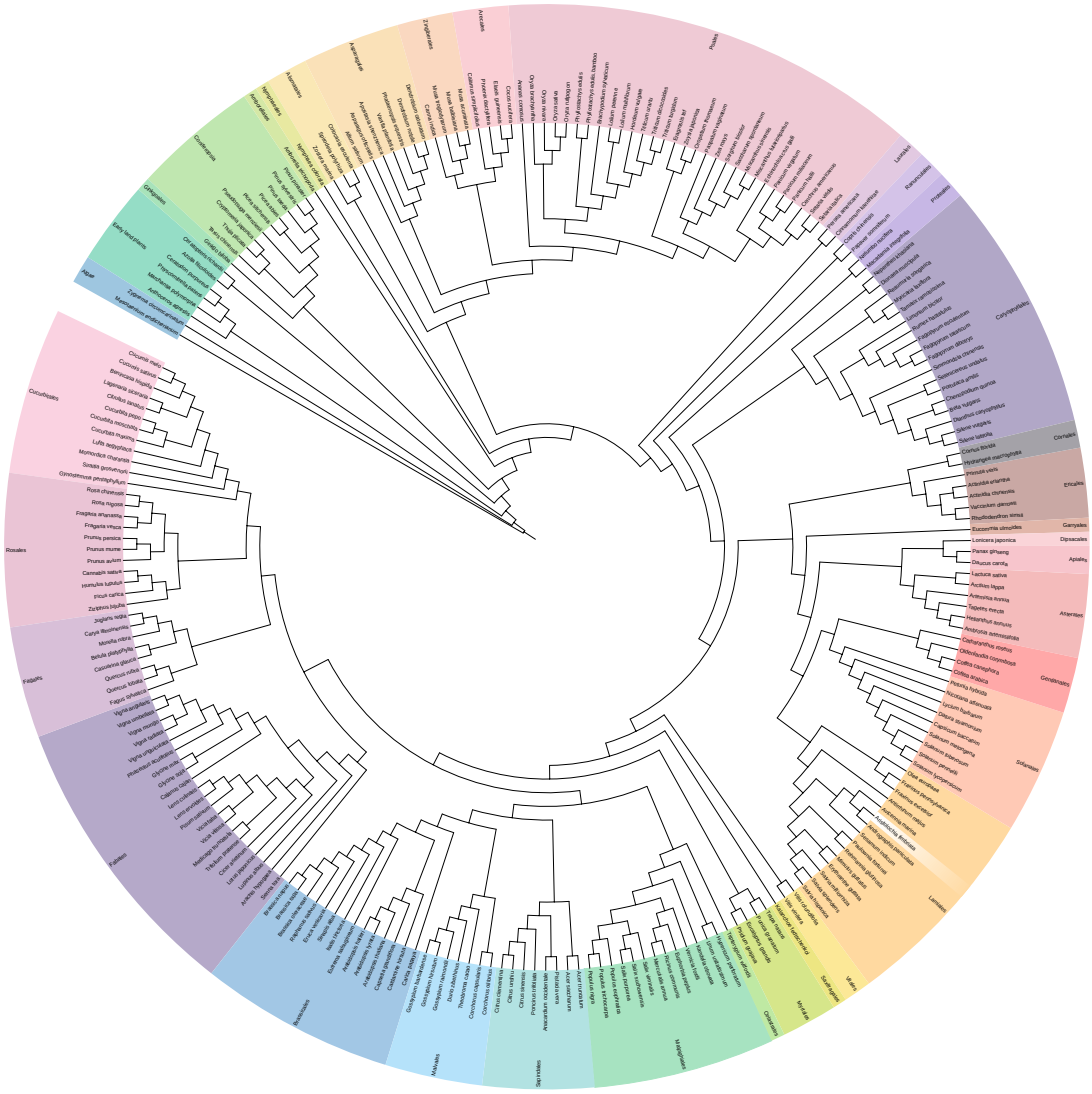

Figure 1: Taxonomic grouping of species included in this study.
Colored ranges represent the taxonomic orders of the species. For better readability, the classification of early land plants and *Coniferopsida* has been simplified. *Aristolochia fimbriata* appears to be misplaced, as it belongs to the *Piperales*, which would be placed close to the *Laurales*. The figure was generated using iTOL [50].

The CoExpPhylo workflow is designed to start from a known metabolite or biosynthetic pathway and a small set of bait genes that are already associated with this pathway. Using these starting points, CoExpPhylo identifies additional candidate genes that might be involved in the same biosynthesis process. To provide the necessary information for this analysis, a configuration (config) file is required. This file defines all datasets (i.e., species) to be included and specifies the paths to the relevant input data. The columns in this file need to be comma-separated without a header. Each row describes one dataset (i.e. one species). The columns contain (species-)IDs, path to the species-specific count table, path to a species-specific multiple FASTA file containing coding sequences, and the path to the species-specific bait file (Table 1). An example config file as well as a synthetic sample dataset demonstrating the required input formats are available in the GitHub repository (https://github.com/bpucker/CoExpPhylo).

Table 1: Explanation and examples of the columns in the config file needed for CoExpPhylo.

ID	Path to TPM file	Path to CDS file	Path to bait file
Unique; usually the species name, is later used as suffix	Path to a count table, first column of count table includes gene IDs, first row contains sample names, other fields are expression values	Path to a multiple FASTA file with coding sequences. Sequence IDs must match sequence IDs in count table	Path to text file with initial bait sequence IDs. IDs must match sequence IDs in TPM and CDS file; one sequence per line
Arabidopsis-thaliana	/path/to/Athaliana.tpm.tsv	/path/to/Athaliana.cds.fasta	/path/to/Athaliana.baits_anthos.txt

A key element of the CoExpPhylo workflow are the baits, which define the set of genes already known to be involved in the biosynthetic pathway of interest. Depending on the available knowledge, this set can consist of multiple genes or even a single gene that serves as the starting point for the analysis. If possible, we recommend selecting late players of the pathway – genes that act close to the final product. Ideally, these bait genes should be specific to the pathway rather than shared with other metabolic branches, and orthologous genes should be provided for each species included in the study. The bait file must be a simple text file, with one sequence ID per line.

To annotate each OCG properly, Araport11 *Arabidopsis thaliana* Col-0 reference polypeptide sequences and annotation terms derived from TAIR10 and Araport11 [56, 57] were used.

Results

The CoExpPhylo workflow systematically analyzes coexpressed genes within one biosynthetic pathway across multiple species, integrating coexpression results, sequence alignment, clustering, and phylogenetic relationships. The following steps outline the approach, with a schematic representation provided in Figure 2.

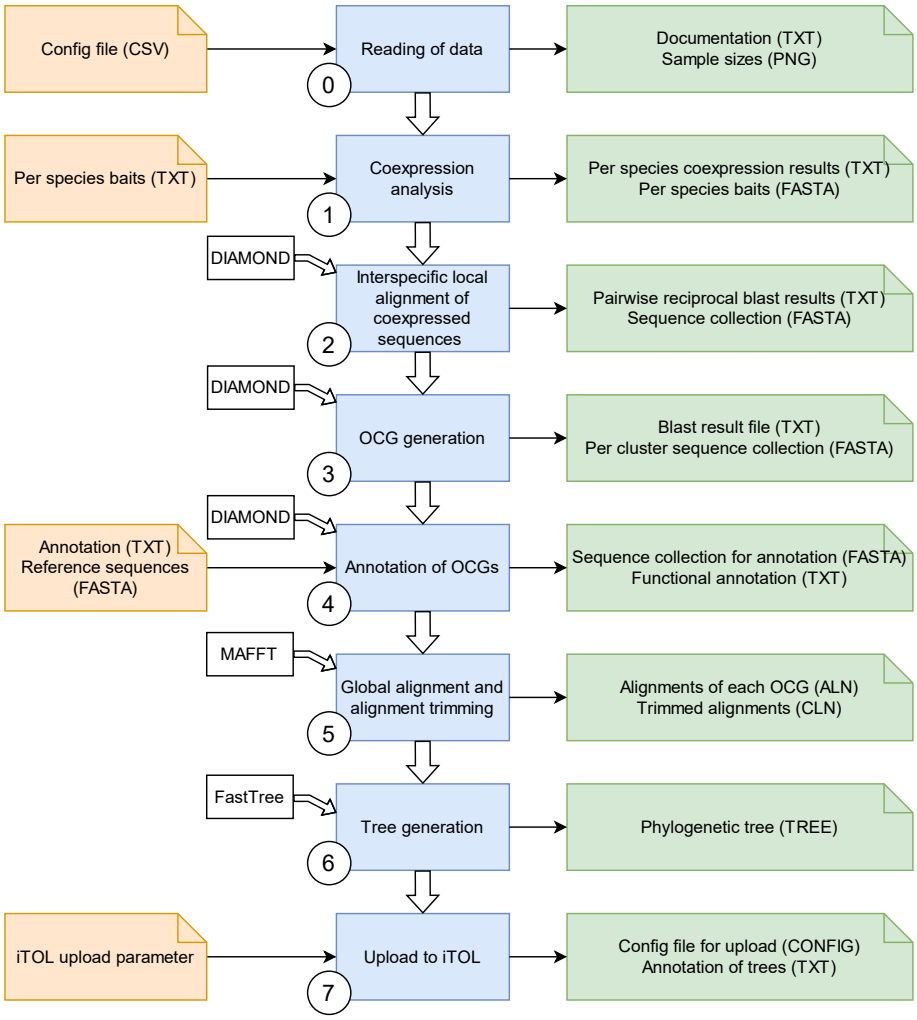

Figure 2: Workflow of CoExpPhylo.
Shown is a schematic overview of the pipeline of coexp_phylo.py. The workflow steps are represented in blue, input files and their file extension are denoted in orange, whereas the output files and their file extensions are colored in green. White boxes indicate the utilization of external tools. OCG means Orthologous Coexpressed Group. This flow chart was generated with drawio.com.

Step 1 – Coexpression analysis

After loading of the input data (Step 0), the coexpression is conducted for every species separately. The IDs of the genes for which the coexpression analysis shall be conducted are loaded from the respective baits file. To evaluate the correlation between the gene expression of the baits and the remaining genes, Spearman's rank correlation was applied *via* the spearmanr-function from the module SciPy [37] in Python to calculate the Spearman's rank correlation coefficient r_s and the corresponding adjusted p-value. The default correlation coefficient cut-off is set to 0.7 but can be adjusted using the argument `--r`. The p-value threshold can be adjusted by specifying the `-p` argument. If not set, the widely accepted default value of 0.05 is used. To further customize the output of the coexpression analyses according to the structure of the input dataset, two additional parameters can be adjusted: Through specification of the argument `--numcut`, the maximal number of genes that are retrieved from each co-expression analysis can be lowered or increased. The default value is set to 100, therefore,

only the top 100 sequences found *via* coexpression analysis are considered for computation in further steps. Lowering the value can reduce the number of relevant genes fetched from the analyses, whereas increasing the value will substantially extend the run time. Secondly, `--min_exp_cutoff` specifies the minimal cumulated expression of any gene that is detected *via* coexpression analysis among all samples. The default is set to 30, which can help to exclude technical artifacts.

Step 2 – Local alignment of coexpressed sequences

Candidate sequences are identified *via* DIAMOND [42]. All sequences collected in the coexpression analyses employed as query, with every species serving as a database. This requires several local alignment runs to be conducted. DIAMOND blastp runs with default parameters, the results are stored in a tabular output format 6. Promising candidate sequences can then be filtered *via* different parameters and are written into one FASTA file. Candidate sequences must exceed the following requirements: Firstly, the e-value of the alignment must not exceed 10^{-5}. Secondly, the bit score must be greater than 100, thirdly, the alignment length must surpass 100, and fourthly, the similarity must be greater than 80%. Those values can be adjusted by setting `--evalue`, `--scorecut`, `--lencut` as well as `--simcut`, respectively. The influence of lowering the length or similarity score cutoff is shown in Additional File 2.

Step 3 – Generation of Orthologous Coexpressed Groups (OCGs)

To divide the sequence collection into clusters, another DIAMOND blastp run is required. At this time, the sequence collection was employed as query and simultaneously as database. The results were filtered according to the previously described parameters. The remaining pairs are incorporated into graphs using the Graph function from the Python module NetworkX [39]. Every sequence that occurs in the filtered BLAST result file is added to the graph as a node, whereas edges are drawn between sequences that have a high blastp result and thus, appear in the filtered blast-result file. Thereby, all sequences with a high sequence similarity are connected within one graph. Finally, all sequences of each OCG are written into one FASTA file per OCG. The OCGs are ranked according to the occurrence of sequences derived from the coexpression analysis. Hence, OCGs with a high ratio between sequences identified *via* coexpression analysis and orthologs identified *via* DIAMOND blastp are labeled firstly, as they presumably contribute more profoundly to the pathway as OCGs with a low number of coexpressed sequences. OCGs with less than ten sequences in general are excluded as well as OCGs that contain coexpressed sequences from less than three species. Thus, the results have a higher significance, as artifacts are excluded.

Step 4 – Functional annotation of OCGs

Optionally, OCGs can be annotated to gauge functionality of the genes within one OCG. To enable this functional annotation, a reference peptide FASTA file must be provided by using `--reference`. If the annotation of the OCGs should not only contain the sequence ID but also a functional description, an annotation table can be specified *via* `--anno`. This

annotation file must be a tab-separated table with the first column containing matching IDs to the provided FASTA file and the second column comprises the respective function. After the sequences collected in step 4 are clustered, a defined percentage of sequences is collected per OCG and written into a separate FASTA file. The percentage can be defined with `--seqs_cluster_anno` and is assigned to 50 by default. Thus, 50% of the sequences of each OCG are selected randomly and are chosen for the annotation in order to save computational resources. However, to ensure the assignment of an annotation valid for the majority of sequences in the OCG, the value cannot be set to less than 10%. If the value is lower, it will automatically be defined as 10, representing 10% of the sequences. Additionally, at least five sequences are employed for the annotation. Howsoever, to annotate an OCG properly, the sequences of each OCG are locally aligned *via* DIAMOND [42] against a database built from the reference FASTA file. Subsequently, the best hit of each alignment is retrieved according to the bit score. Among all annotations of each OCG, the most common one is selected as the functional annotation for the respective OCG. To elucidate the accuracy of the annotation, a reliability score is displayed in the functional annotation table. This score indicates the proportion of sequences that were annotated with the conclusively selected annotation.

Step 5 – Global alignment

The global alignment of each OCG is executed either by MAFFT [43, 44] or optionally with MUSCLE [45]. With the argument `--alnmethod`, the tool can be selected. By default, MAFFT is used. If the executable for the intended tool is not included in the user's system PATH environment variable, the path to the tool must be provided with `--mafft` or `--muscle`, respectively. The global alignment is executed using the default parameters of the tools. Afterwards, the aligned sequences are trimmed, so that only positions with sufficient occupancy are kept. The occupancy cutoff is set to 10% but can be adjusted by the user with the argument `--occupancy`. The output file is written in FASTA format.

Step 6 – Phylogenetic tree generation

The cleaned and trimmed alignment files are subsequently used as input files for the phylogenetic tree construction. With the argument `--treemethod`, the tool for tree generation can be specified. By default, FastTree [46, 47] is utilized with the options `-wag` and `-nosupport`. Alternatively, RAxML-NG [48] using the model "LG+8G+F" or IQTREE [49] can be applied for this step. To enable the upload to common phylogenetic tree visualization programs, especially the batch upload to iTOL [50], the output files have Newick format.

Step 7 – Batch upload to iTOL

After the CoExpPhylo analysis finished, the tree files can be uploaded to any phylogenetic tree visualization software. As the number of generated tree files can be complex to handle manually, it is possible to upload the tree files automatically to iTOL [50]. This upload requires a Perl script that can be downloaded from the iTOL help page (https://itol.embl.de/help.cgi#batch). By using the argument `--upload_script`, the path

to this upload script must be provided if it is not stored in the same folder as the main Python script coexp_phylo.py. Furthermore, an annotation file is uploaded with every tree that enables coloration of the sequences tagged with the suffix "_coexp", i.e. sequences that were collected *via* coexpression analyses, in iTOL. The automatic batch upload is only possible with an active standard subscription and requires the specification of the API using the argument `--API`. All trees are uploaded into the project that is defined by `--pro_name`. This project must already exist in the user's account and should be unique among all workspaces. To avoid failure of the script due to misspelling in the API or project name, a test-tree is uploaded during the initial step if a batch upload to iTOL is desired. An error during the test upload leads to the termination of the whole coexp_phylo.py script with a corresponding error message.

Output files

CoExpPhylo generates several final output files summarizing the analysis results (Table 2). The documentation file records the script version, parameter settings, external tool versions, and input file paths, including an MD5 hash for each file. To assess species representation, an HTML-based species count histogram visualizes the distribution of coexpressed sequences per species. This allows users to detect potential biases, such as underrepresented species with low coexpression signal due to limited sample availability.

A tab-separated ranking table provides an overview of OCGs based on the proportion of coexpressed sequences. This table includes the number of sequences per OCG, annotation details (if applied), and a confidence score reflecting the relationship to the assigned annotation sequence. The confidence score is determined by comparing a subset of sequences from each OCG to a reference database, reflecting the proportion of sequences with high similarity to the final annotation sequence.

For phylogenetic analysis, a Newick-formatted tree file is generated for each OCG, representing the evolutionary relationships among the identified sequences (Figure 3). Additionally, two annotation files are produced per OCG to facilitate visualization in iTOL [50] if automatic upload is not enabled. The first annotation file (*iTOL_coexp_annotation.txt*) highlights sequences identified through the initial intra-specific coexpression analysis. The second annotation file (*iTOL_gradient_labels.txt*) assigns a color bar to each sequence, reflecting its highest correlation coefficient. This visualization aids in assessing the overall correlation strength of an OCG to the biosynthetic pathway under investigation.

Table 2: Overview over the core output files generated by CoExpPhylo.

File name	Content
docu.txt	Documentation about all parameters, versions of applied tools, and input files including MD5 hashes of all input files
species_count_histogram.html	Bar chart showing the distribution of the sample sizes for each species, i.e. number of transcriptomic datasets

functional_annotation.txt	Tab-separated table including the number of sequences within each OCG, the functional annotation for each OCG and the respective confidence score
number.tree	One tree file for each OCG in Newick format
Annotation of tree files: iTOL_coexp_annotation.txt iTOL_gradient_labels.txt	Two annotation files for each tree file, including a color range for the correlation coefficient of each sequence and a marking of the sequences with a "_coexp"-tag, can be uploaded into iTOL [50]

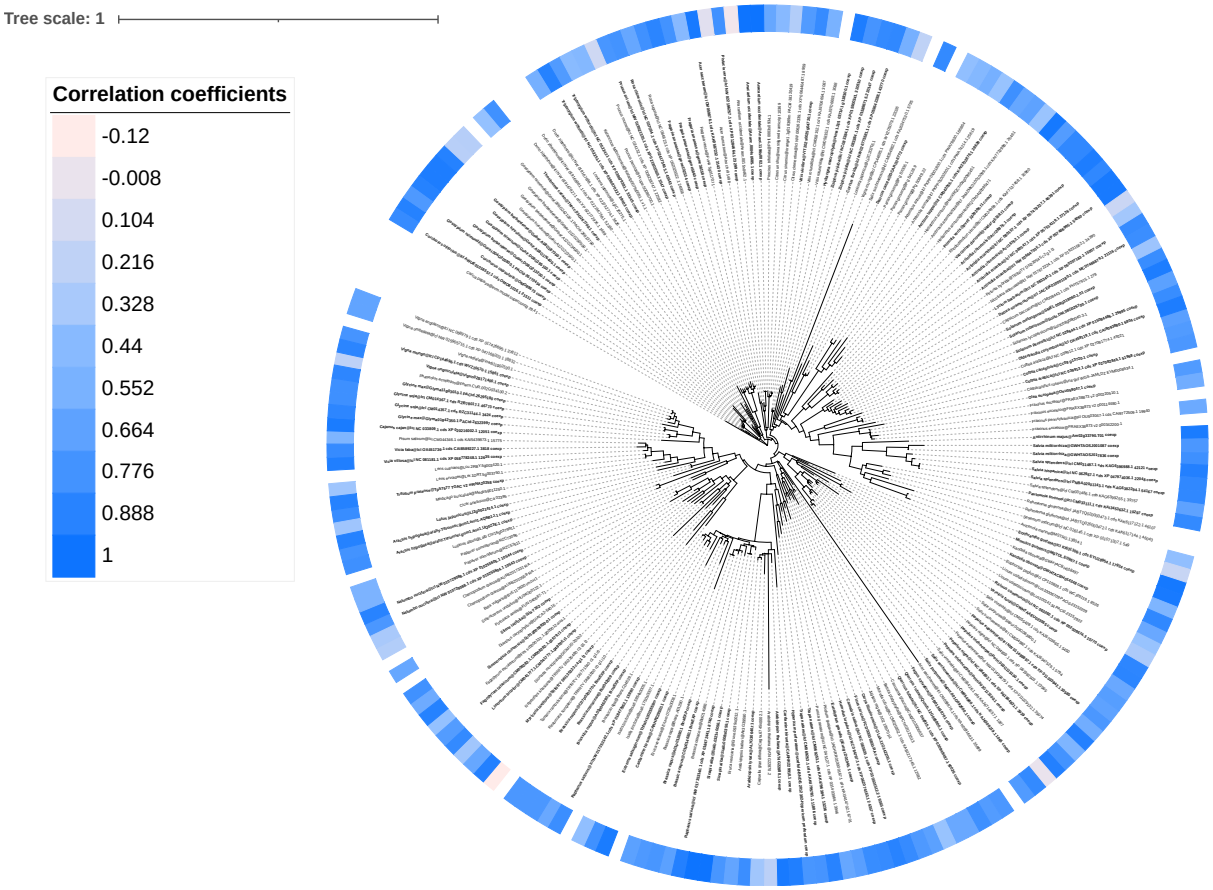

Figure 3: Exemplary representation of the visualization of one Step via iTOL [50]. Shown is the result of Orthologous Coexpressed Group (OCG) 0009 retrieved through CoExpPhylo with respect to the anthocyanin biosynthesis synthesis. CoExpPhylo was conducted based on the transcriptomic data of 239 species [51, 52], dihydroflavonol 4-reductase (DFR), anthocyanidin synthase (ANS), and the anthocyanin-related glutathione S-transferases (arGSTs) Bronze2 (Bz2) as well as transparent testa 19 (TT19*)* were used as baits. Candidate baits were identified via KIPEs3 [58, 59]. Any species without identifiable baits as well as samples with a high expression (> 5 tpm) of leucoanthocyanidin reductase (LAR) or anthocyanidin reductase (ANR) were excluded from the analysis. The global alignment was conducted with MAFFT [43, 44]. The trees were constructed with FastTree [46, 47].

Proof of concept: application to flavonoid biosynthesis

To evaluate the performance, we applied CoExpPhylo to three distinct branches within the flavonoid biosynthesis: anthocyanins, PAs, and flavonols. We used respective genes as input

baits and were able to detect various OCGs annotated with already known players of the biosynthetic pathway as well as possible candidate genes (Figure 4A).

Briefly, flavonoid biosynthesis originates from the general phenylpropanoid pathway, which involves the key enzymes phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), and 4-coumarate:CoA ligase (4CL). The core flavonoid biosynthesis pathway includes chalcone synthase (CHS), chalcone isomerase (CHI), and flavanone 3-hydroxylase (F3H). The emerged dihydroflavonols can be further diversified through the flavonoid 3'-hydroxylase (F3'H) and the flavonoid 3',5'-hydroxylase (F3'5'H) [20, 60]. Flavonol biosynthesis proceeds via flavonol synthase (FLS), while the anthocyanin biosynthesis requires dihydroflavonol 4-reductase (DFR), anthocyanidin synthase/leucoanthocyanidin dioxygenase (ANS/LDOX), anthocyanin-related glutathione S-transferase (arGST), and UDP-dependent anthocyanidin-3-O-glycosyltransferase (3GT) [20, 31]. PAs can be synthesized through leucoanthocyanidin reductase (LAR) or anthocyanidin reductase (ANR).

The flavonoid biosynthesis is transcriptionally regulated by a complex consisting of a MYB, a bHLH, and a WD40 transcription factor, collectively forming an MBW complex [9, 61]. The MYB component determines the target gene, forming a complex with bHLH42 and TTG1 [62]. MYB11, MYB12 and MYB111 regulate the flavonol biosynthesis bHLH-independent [7], while proanthocyanidin biosynthesis is regulated through MYB123 in a complex with a bHLH and a WD40 component [63]. MYB75, MYB90, MYB113, and MYB114 are responsible for anthocyanin biosynthesis regulation [9, 64, 65].

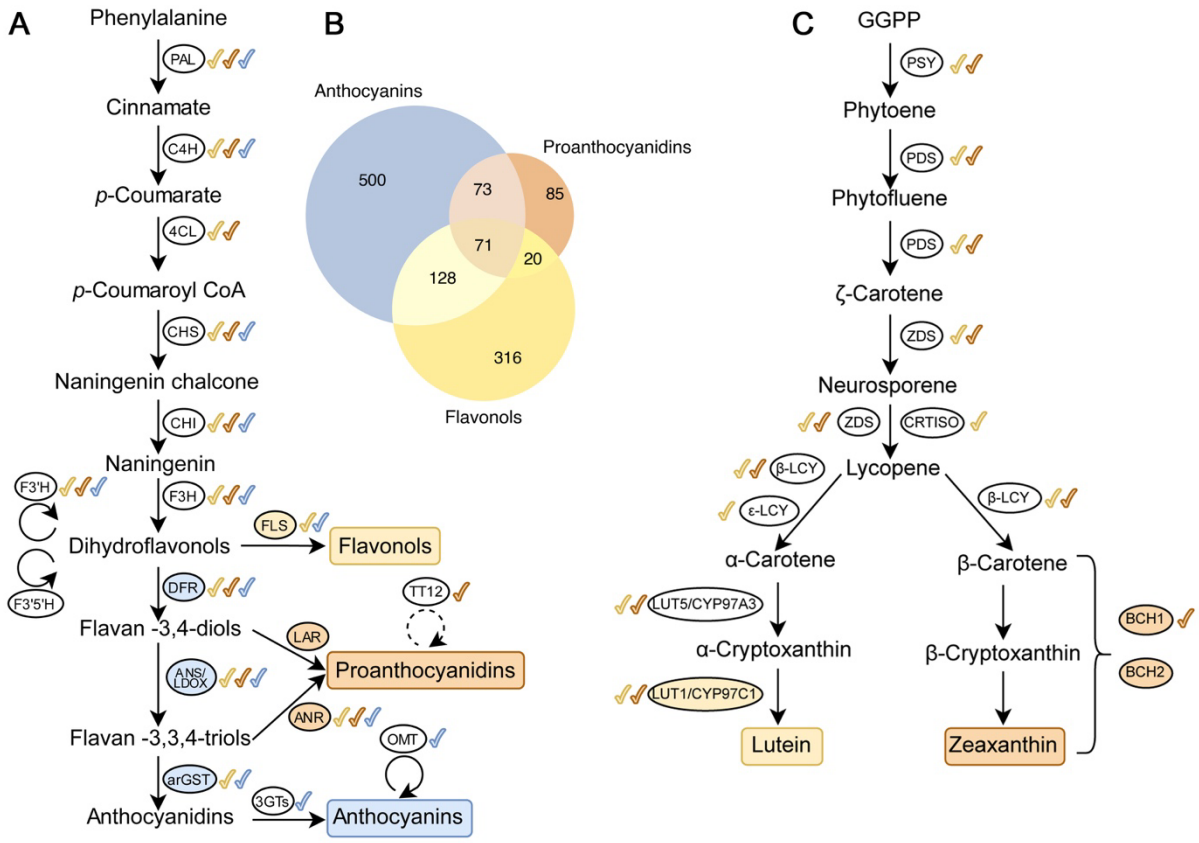

Figure 4: Schematic overview of different branches of the flavonoid and carotenoid biosynthesis and the detection via CoExpPhylo.
Enzymes highlighted with a colored background were used as input baits for the analysis of the corresponding biosynthetic branch. Colored check marks indicate the Orthologous Coexpressed Groups (OCGs) retrieved for that branch.
(A) Displayed enzymes are PHE ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H), 4-coumarate:CoA ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H), flavonoid 3',5'-hydroxylase (F3'5'H), flavonol synthase (FLS), dihydroflavonol 4-reductase (DFR), leucoanthocyanidin reductase (LAR), transparent testa 12 (TT12), anthocyanidin synthase/leucoanthocyanidin dioxygenase (ANS/LDOX), anthocyanidin reductase (ANR), anthocyanin-related glutathione S-transferase (arGST), UDP-dependent anthocyanidin-3-O-glycosyltransferase (3GT), O-methyltransferase (OMT). The dotted arrow indicates a transporter.
(B) Venn diagram showing the overlap of genes identified in the three biosynthetic pathways: anthocyanins, proanthocyanidins and flavonols.
(C) Displayed enzymes are phytoene synthase (PSY), phytoene desaturase (PDS), ζ-carotene desaturase (ZDS), carotenoid isomerase (CRTISO), lycopene β -cyclase (β-LCY), lycopene ε -cyclase (ε-LCY), lutein deficient5/cytochrome P450-family 97-A3 (LUT5/CYP97A3), lutein deficient1/cytochrome P450-family 97-C1 (LUT1/CYP97C1), beta carotenoid hydroxylase 1 (BCH1), and beta carotenoid hydroxylase 2 (BCH2).

Anthocyanins

Focusing on the anthocyanin branch, which is responsible for the vibrant pigmentation in flowers and fruits, late-stage enzymes were selected as bait candidates to evaluate CoExpPhylo's capability in identifying genes involved in specialized metabolic pathways. This approach aimed to determine whether the tool could accurately recover established biosynthetic genes and differentiate them from those associated with related flavonoid branches.

The final enzymatic step in anthocyanin biosynthesis is catalyzed by UDP-glucose:flavonoid glucosyltransferases (UFGTs). However, UFGTs cannot yet be reliably annotated due to the limited knowledge about their conserved amino acid residues. To avoid introducing uncertainty into the analysis, UFGTs were excluded from the bait selection. As the lack of a comprehensive set of amino acid residues required for the activity of arGSTs complicates their identification across species, DFR and ANS/LDOX were also included as baits to ensure a robust test case for CoExpPhylo.

A major challenge in using *DFR* and *ANS* as baits is their dual involvement in anthocyanin and PA biosynthesis. To mitigate potential misassignments, RNA-seq samples with considerable expression (>5 tpm) of either LAR or ANR – late-stage enzymes of PA biosynthesis – were excluded. Since low expression of LAR and ANR suggests minimal PA biosynthesis, the remaining DFR and ANS expression data are more likely linked to anthocyanin biosynthesis. By applying these selection criteria, we aimed to create a controlled test scenario that allows for an initial evaluation of CoExpPhylo's ability to distinguish biosynthetic genes from related pathways and to identify potential new candidates.

The output (Additional File 3) shows OCGs associated with multiple key enzymes of the anthocyanin biosynthesis. One OCG was annotated with PAL and another with C4H, both enzymes of the phenylpropanoid pathway. Additionally, CHS, CHI, F3H and F3'H were assigned to distinct sequence clusters. Interestingly, no OCG corresponding to 4CL, which acts between C4H and CHS in the pathway, was detected. This absence could be due to the gene's diverse

402 functional roles beyond flavonoid biosynthesis its broad substrate specificity (e.g., 4-coumaric
403 acid, ferulic acid, and caffeic acid), or potential limitations in coexpression-based detection
404 [66, 67]. The involvement of 4CL in multiple metabolic branches may result in a distinct ex-
405 pression pattern that does not correlate strongly with flavonoid-specific genes. Due to the
406 phylogenetic divergence of the input species, multiple OCGs were generated for certain genes.
407 For instance, F3H was represented by three OCGs (0017, 0037, and 0069), corresponding to
408 different taxonomic groups: one cluster contained sequences from *Eudicotyledons*, another
409 from *Coniferopsida* and the third from species belonging to the order *Poales*.

410 The initial bait genes (*DFR*, *ANS*/*LDOX,* and *arGSTs*) were also retrieved as OCGs. Furthermore,
411 the OCGs annotated as UDP-dependent glycosyltransferase UGT78D2, along with the antho-
412 cyanin-decorating enzymes SCPL10, AT1, and O-methyltransferase 1 (OMT), were identified.
413 Two additional OCGs (0072 and 0145) were annotated with UDP-dependent glycosyltransfer-
414 ases UGT80B1 and UGT84A1, respectively. Although these enzymes have not been broadly
415 characterized in anthocyanin modification, the *Medicago truncatula* homolog *Mt*UGT84A1
416 has been reported to be linked to increased anthocyanin accumulation alongside its role in
417 shoot growth [68]. Further analysis of the sequences within these OCGs may reveal additional
418 glycosyltransferases pertinent to anthocyanin biosynthesis. However, the annotation process
419 is based on *Arabidopsis thaliana* reference sequences, which may explain certain limitations.
420 Genes or functions absent in *A. thaliana* cannot be annotated correctly, so that the annotation
421 of UGT80B1 and UGT84A1 might be an artifact thus suggesting that some other UGT lineages
422 could contribute to the diversity of anthocyanins. Similarly, a potential OCG containing se-
423 quences encoding F3'5'H was not identified as *A. thaliana* lacks this enzyme, highlighting the
424 importance of a multi-species reference database for the annotation.

425 Additionally, transcription factors associated with anthocyanin biosynthesis regulation, includ-
426 ing MYB90, MYB12, and the bHLH transcription factor bHLH42/TT8, were identified. Notably,
427 FLS was also detected, despite its primary role in flavonol biosynthesis rather than anthocya-
428 nin production. However, FLS; F3H and ANS/LDOX belong to the 2-oxoglutarate-dependent
429 dioxygenase (2ODD) superfamily whereby functional redundancy might occur [69].

430 **Proanthocyanidins**

431 PAs, also known as condensed tannins, contribute to plant defense and seed coat pigmenta-
432 tion. Their biosynthesis shares multiple enzymatic steps with anthocyanin biosynthesis, mak-
433 ing it challenging to delineate genes specific to this pathway. To assess CoExpPhylo's ability to
434 resolve this complexity, late stage PA biosynthetic genes, including *LAR* and *ANR* (also known
435 as *BANYLUS* (*BAN*)) were used as baits.

436 The initial bait gene *BAN* was detected three times among the final OCGs. As *A. thaliana* does
437 not have a LAR, no OCG was annotated accordingly. However, all genes encoding enzymes
438 involved in the straight PA biosynthesis, starting with PAL, were successfully retrieved within
439 annotated OCGs (Figure 4A). Notably, genes associated with closely related pathways, such as
440 arGSTs and FLS), were not retrieved, suggesting a certain specificity of CoExpPhylo in

distinguishing metabolic branches. Nonetheless, there is a moderate overlap between the analyzed pathways (Figure 4B). This can be explained as the metabolites share the majority of their upstream pathway and by the promiscuity of several enzymes, such as FLS, ANS, and multiple UFGTs [70–74]. Hence, the identification of the same candidate genes in multiple pathways does not indicate any errors, but is biologically plausible and expected.

Additionally, one OCG was annotated with the bHLH transcription factor bHLH42/TT8, which is involved in PA biosynthesis. The main MYB transcription factor regulating PA biosynthesis, MYB123, was not identified as a separate OCG. This highlights a limitation of CoExpPhylo in resolving individual lineages of large gene families such as MYB transcription factors. Due to high sequence similarity within these families, the delineation of OCGs may be influenced by the chosen similarity threshold, potentially leading to the merging of closely related sequences. Possibly, the OCG annotated with MYB5 (0058) contained MYB123 sequences along with other members of MYB subgroup 5. To achieve a more precise resolution of large transcription factor families, dedicated tools like MYB_annotator might show superior performance [75]. Furthermore, the WD40 transcription factor TTG1 was not captured as an OCG. This may be due to its diverse functional roles, spanning various developmental processes and specialized metabolism [63, 76–79] which lead to a constant expression pattern that is not suitable for detection via coexpression analysis.

Flavonols

Flavonols, a subclass of flavonoids, function as antioxidants and UV protectants in plants. They serve as precursors for multiple downstream metabolic pathways, making their biosynthetic gene network particularly intricate. To investigate the flavonol branch, FLS, a key enzyme converting dihydroflavonols into flavonols, was selected as a bait sequence. Since FLS competes with DFR for its substrate [80–82], its expression pattern provides a means to distinguish flavonol biosynthesis from related pathways.

All core genes of the flavonol biosynthesis pathway were retrieved within annotated OCGs. However, additional OCGs were identified containing sequences annotated with genes typically associated with anthocyanin biosynthesis, such as *DFR* and *ANS*/*LDOX*, or PA biosynthesis (e.g. *BAN*) which do not directly participate in flavonol production. This suggests that the separation of the flavonol and anthocyanin branches is not entirely distinct in the coexpression analysis, potentially reflecting their shared metabolic precursors and regulatory elements. Moreover, this overlap might be influenced by the diversity of the cell types included in the transcriptomic data. Previous studies have demonstrated cell type-specific expression patterns for different branches of flavonoid biosynthesis [83, 84].

In addition to enzymatic genes, transport and modification-related genes were identified. The transporter DTX35, implicated in flavonol transport, was retrieved as an OCG. Moreover, UDP-dependent glycosyltransferases UGT78D2, UGT84A1, and UGT84A2 were identified, suggesting potential roles in flavonol glycosylation. UGT78D2, an anthocyanin-glycosyltransferase, is present in *Caryophyllales* – an order lacking anthocyanins but having flavonols [85]. This

underlines their potential involvement in flavonols glycosylation, which was already reported in *Brassica napus* [86]. Regarding transcription factors, the bHLH regulator bHLH42/TT8 and MYB12 were captured as OCGs. However, MYB11 and MYB111, which together with MYB12 constitute subgroup 7 of the R2R3-MYB family [7], were not detected. The OCG annotated with MYB12 contained 336 sequences, indicating that the boundaries within this gene family might be blurred, potentially leading to the merging of multiple MYB subgroup 7 members into a single OCG.

Proof of concept: application to carotenoid biosynthesis

To further evaluate the performance of CoExpPhylo, we applied it to the carotenoid biosynthesis pathway, focusing on lutein and zeaxanthin branches. Using down-stream biosynthetic genes as input baits, the pipeline successfully detected multiple OCGs corresponding to known enzymes, as well as potential candidate genes involved in the pathway (Figure 4C).

Carotenoid biosynthesis starts from the precursor geranylgeranyl diphosphate (GGPP), which is converted to phytoene by phytoene synthase (PSY) [87]. Phytoene is subsequently desaturated by phytoene desaturase (PDS) and ζ-carotene desaturase (ZDS), forming lycopene [88, 89]. Lycopene cyclases, lycopene β-cyclase (β-LCY) and lycopene ε-cyclase (ε-LCY), catalyze the formation of α-carotene and β-carotene, key intermediates for xanthophyll biosynthesis [90]. Lutein biosynthesis proceeds via hydroxylation of α-carotene catalyzed by lutein deficient 5 (LUT5)/CYP97A3 and lutein deficient 1 (LUT1)/CYP97C1, while zeaxanthin is produced by hydroxylation of β-carotene, involving enzymes such as the β-carotene hydroxylases BCH1 and BCH2 [91–95].

Lutein

Focusing on the lutein branch of carotenoid biosynthesis, which plays a central role in photoprotection and light harvesting in plants, we evaluated CoExpPhylo's performance using LUT1/CYP97C1 as bait. All known genes of the lutein pathway (PSY, PDS, ZDS, CRTISO, β-LCY, ε-LCY, LUT5/CYP97A3, and LUT1/CYP97C1) were recovered (Additional File 6), demonstrating the tool's ability to reconstruct the main biosynthetic route.

No well-defined transcription factors could be identified in our analysis, as there is little consensus on transcriptional regulation of carotenoid biosynthesis across species and tissues [96]. Regulatory mechanisms are highly context-dependent, which makes detection through coexpression and phylogenetic approaches particularly challenging. Carotenoids, although sometimes associated with secondary traits such as pigmentation, are considered central metabolites due to their essential roles in photosynthesis and photoprotection, particularly in the case of lutein [97]. This underlines that while CoExpPhylo is particularly well-suited for specialized metabolism with more distinct pathway boundaries, it is also applicable to central metabolism, though in such cases, more stringent filtering strategies may be required.

Zeaxanthin

Focusing on the zeaxanthin branch, we recovered all known biosynthetic genes except CRTISO and BCH2, although both BCH1 and BCH2 were included as baits (Additional File 7). Interestingly, CoExpPhylo also identified LUT5/CYP97A3, LUT1/CYP97C1, and ε-LCY, which are primarily associated with lutein biosynthesis. This cross-detection likely reflects the fact that lutein biosynthetic enzymes are expressed constitutively across diverse conditions, causing their transcriptional profiles to co-vary with the more dynamically regulated zeaxanthin genes in our heterogeneous dataset.

Moreover, we detected a large number of clusters (706 OCGs), reflecting the fact that carotenoids, and especially xanthophylls, serve as precursors for a wide range of other compounds, including the plant hormones abscisic acid and strigolactones as well as various apocarotenoids [98, 99]. This metabolic branching makes distinguishing the core biosynthetic genes more challenging compared to specialized pathways.

Tool benchmarking

CoExpPhylo allows the user to select either MAFFT [43, 44] or MUSCLE [45] for the multiple sequence alignment in Step 5. To evaluate the impact of the alignment method on the phylogenetic analysis, several PSS were processed using both tools. A representative dataset comprising 43 sequences revealed key differences between the alignments (Additional File 8). Five larger groups could be identified that contained the same sequences but were positioned differently and the internal branches within these groups were arranged differently. The preservation of major groupings shows that both tools are consistent with the overall evolutionary structure but differ in detailed phylogenetic distances, i.e. the connections between the groups are interpreted differently. These discrepancies stem from the different algorithms and scoring systems that both alignment tools use.

In contrast, a comparison of the different available tools for tree inference from these alignments – FastTree [46, 47], RAxML [48] or IQ-TREE [49] – revealed greater variability in sequence placement. Although the overall tree topology remained consistent, topological details varied across different tree-building approaches (Additional File 9). Notably, the phylogenetic differences introduced by alternative alignment methods exceeded those introduced by different tree inference tools.

The choice of tree inference software significantly influenced runtime performance (Figure 5). The utilization of MUSCLE instead of MAFFT for the global alignment step leads to a runtime increase of 9.5 minutes or 12.5 minutes, depending on the execution mode. Replacing FastTree with another phylogenetic tree inference tool has a stronger impact on the runtime performance: The utilization of IQ-TREE increased the computing time by over three and four hours, respectively. Executing MAFFT in combination with RAxML-NG takes over 140 hours, amounting to almost six days (Figure 5).

Parallel processing mitigated the impact of large OCGs on runtime. As all OCGs were processed simultaneously, the largest cluster determined the overall computation time. This effect was particularly pronounced for IQ-TREE and RAxML-NG, where the largest OCG (containing 212 sequences) accounted for 77.7% and 99% of the total runtime, respectively. Given that the dataset included 20 OCGs with over 1,000 sequences each, computational demands would increase substantially for larger datasets. Thus, the computational time would increase profoundly. Overall, a change in the global alignment method does not affect the runtime as severely as replacing FastTree by either IQ-TREE or RAxML-NG for these example OCGs. However, utilization of MAFFT on all OCGs of the used dataset does have a significant impact on the runtime as the global alignment of the three largest OCGs (> 2,000 sequences) took over 100 hours (data not shown).

In summary, the combination of MAFFT for global alignment and FastTree for phylogenetic tree construction represents the most efficient approach. Given that the biological results obtained with different methods show no substantial differences, but computational time increases dramatically with OCG size for alternative tools, the use of MAFFT and FastTree ensures an optimal balance between accuracy and performance.

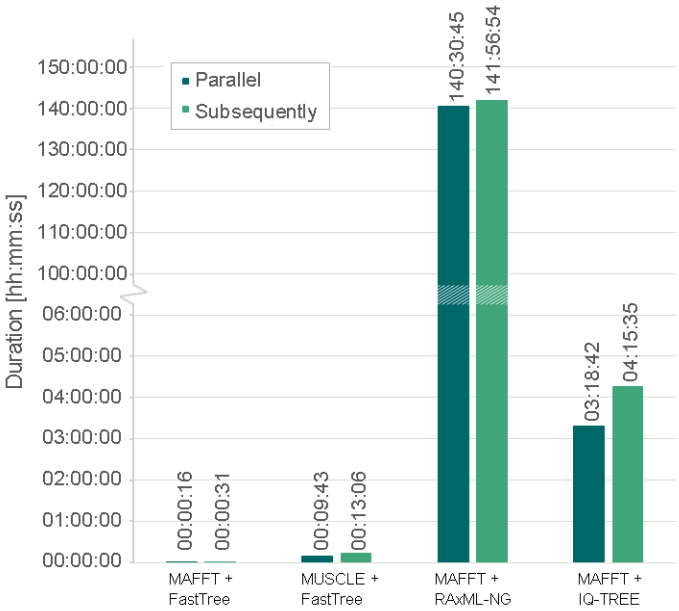

Figure 5: Comparison of the runtime performance of global alignment and phylogenetic tree construction. Shown are the runtime performances of the global alignment step combined with the construction of phylogenetic trees. Three representative sequence collections were selected (containing 43, 113, and 212 sequences) and processed in parallel (dark green) or subsequently (light green). The global alignment was conducted either with MAFFT v7.490 [43, 44] or MUSCLE v3.8.1551 [45], the phylogenetic tree was constructed with FastTree v2.1.11 [46, 47], RAxML-NG v1.2.2 [48], or IQ-TREE v2.0.7 [49]. The vertical axis was split between 6 and 100 hours in order to visualize all data points.

Conclusions

The discovery of multiple pathways based on a single bait gene in each indicates that CoExpPhylo can identify promising candidate genes of specialized biosynthetic pathways across

multiple species, providing a solid basis for further functional and comparative analyses. By integrating coexpression analysis with phylogenetic relationships, the pipeline enables a broader, evolutionary-informed perspective on pathway organization and gene function. In addition to core enzymatic genes, CoExpPhylo can detect a range of transcription factors and transporters, offering a more comprehensive view of pathway regulation, even though it may not capture all regulatory players.

A key strength of CoExpPhylo is its ability to identify functionally related genes even in cases where coexpression pattern would be diffuse. This makes it a powerful tool for expanding pathway annotations beyond well-characterized model species. The generated OCGs offer a data-driven basis for highlighting candidate genes with potential involvement in a biosynthetic pathway, independent of prior pathway knowledge. However, these candidates should be considered as a starting point for further experimental validation. By offering a systematic and large-scale approach to gene discovery, CoExpPhylo helps to bridge the gap between computational predictions and functional characterization, addressing one of the major challenges in identifying novel genes with specific biochemical functions.

While CoExpPhylo effectively resolves many pathway components, opportunities for further refinement remain. For example, large gene families, such as MYB transcription factors, may exhibit high sequence similarity, leading to broader clustering in some cases. Furthermore, in some cases, genes from closely related metabolic pathways were grouped within the same OCGs. This highlights the need for additional functional validation to precisely determine pathway specificity, particularly for multifunctional enzymes or regulatory proteins. Additionally, annotation quality is dependent on the reference database used, which may limit the classification of genes absent in *Arabidopsis thaliana*. Expanding reference data to include multiple species and refining the clustering approach for highly homologous sequences could further enhance the resolution of functional gene groups. Furthermore, incorporating a protein-protein interaction predictor would enhance the functionality of the program for pathways that form a metabolon. Accurate protein-protein interaction predictions could underline the likelihood of physical interaction, thereby supporting the functional relevance of identified candidate genes. In such cases, integrating predicted protein-protein interactions would help to differentiate between genes that are merely co-expressed or homologous and those that may directly cooperate at the protein level within the same metabolic pathway.

Despite these considerations, CoExpPhylo represents a valuable tool for biosynthetic pathway exploration, enabling the identification of both conserved and lineage-specific genes. By facilitating candidate gene discovery across diverse plant species, it offers new opportunities to investigate specialized metabolism with a systematic and scalable approach.

617 ## Availability and requirements

618 **Project name:** CoExpPhylo

619 **Project home page:** https://github.com/bpucker/CoExpPhylo

620 **Operating system(s):** Tested on Ubuntu but should work on all Unix-like operating systems

621 **Programming language:** Python, Bash

622 **Other requirements:** Python 3.0 or higher

623 **License:** GNU GPL

624 **Any restrictions to use by non-academics:** not applicable

625 ## Abbreviations

3GT	Anthocyanidin-3-O-Glycosyltransferase
4CL	4-Coumarate:CoA Ligase
ANR	Anthocyanidin Reductase
ANS	Anthocyanidin Synthase
arGST	Anthocyanin-related Glutathion S-Transferase
BAN	BANYLUS
BCH	Beta carotenoid hydroxylase
β-LCY	Lycopene β -cyclase
C4H	Cinnamate 4-Hydroxylase
CDS	Coding Sequence
CHI	Chalcone Isomersae
CHS	Chalcone Synthase
CRTISO	Carotenoid isomerase
CYP97	Cytochrome P450, family 97
DFR	Dihydroflavonol 4-Reductase
ε-LCY	Lycopene ε -cyclase
F3'5'H	Flavonoid 3’,5'-Hydroxylase
F3'H	Flavonoid 3’-Hydroxylase
F3H	Flavanone 3-Hydroxylase
FLS	Flavonol Synthase
GGPP	Geranylgeranyl diphosphate
LAR	Leucoanthocyanidin Reductase
LDOX	Leucoanthocyanidin Dioxygenase
LUT	Lutein deficient
NCBI	National Center for Biotechnology Information
OCG	Orthologous Coexpressed Group
OMT	O-Methyltransferase
PA	Proanthocyanidin

PAL	Phenylalanine Ammonia-Lyase
PDS	Phytoene desaturase
PSY	Phytoene synthase
SRA	Sequence Read Archive
TPM	Transcripts Per Million
TT	Transparent Testa
UFGT	UDP-Glucose:Flavonoid Glucosyltransferase
ZDS	ζ-carotene desaturase

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have read and approved the final manuscript.

Availability of data and materials

The authors declare that all datasets used in this paper are available online [51, 52] and the Python code is published on GitHub: https://github.com/bpucker/CoExpPhylo.

Competing interests

The authors declare no competing interests.

Funding

Not applicable.

Authors' contributions

NG and BP planned the study and wrote the software. NG conducted the bioinformatic analyses. NG and BP interpreted the results and wrote the manuscript.

Acknowledgements

This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). We also thank all members of the research group Plant Biotechnology and Bioinformatics for discussion and support. Open Access funding enabled and organized by Project DEAL and the University of Bonn.

Additional files

Additional File 1

Additional_File_1.csv

Title: Species used for analysis and their taxonomic ranking.

Description: Shown are all species and their respective order and family used in this analysis. CDS and TPM files were collected from two data publications [51, 52].

Additional File 2

Additional_File_2.pdf

Title: Distribution of OCG sizes at different similarity and length thresholds.

Description: Shown are the number of sequences per OCG and their occurrence at different similarity and length cutoffs. Starting from default values (A), the similarity cutoff was reduced to 75 (C) and 70 (E), whereas the length cutoff was reduced to 90 (B) and 80 (D).

Additional File 3

Additional_File_3.txt

Title: Functional annotation of OCGs – Anthocyanins.

Description: Shown are the results of CoExpPhylo with respect to anthocyanins. The analysis was based on two transcriptomic data sets [51, 52] with *DFR*, *ANS*, *Bz2*, and *TT19* as bait genes. Samples with a high expression (>5 TPM) of *LAR* and/or *ANR* were excluded. The analysis was conducted with default parameters. The functional annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and and corresponding information [56].

Additional File 4

Additional_File_4.txt

Title: Functional annotation of OCGs – Proanthocyanidins.

Description: Shown are the results of CoExpPhylo with respect to proanthocyanidins. The analysis was based on two transcriptomic data sets [51, 52] with *LAR* and *ANR* as bait genes. The analysis was conducted with default parameters. The functional annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and and corresponding information [56].

Additional File 5

Additional_File_5.txt

Title: Functional annotation of OCGs – Flavonols.

Description: Shown are the results of CoExpPhylo with respect to proanthocyanidins. The analysis was based on two transcriptomic data sets [51, 52] with *FLS* as the bait gene. The analysis was conducted with default parameters. The functional annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and and corresponding information [56].

Additional File 6

Additional_File_6.txt

Title: Functional annotation of OCGs – Lutein.

Description: Shown are the results of CoExpPhylo with respect to lutein. The analysis was based on two transcriptomic data sets [51, 52] with *LUT1/CYP97C1* as bait genes. The analysis was conducted with default parameters. The functional annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and and corresponding information [56].

Additional File 7

Additional_File_7.txt

Title: Functional annotation of OCGs – Zeaxanthin.

Description: Shown are the results of CoExpPhylo with respect to zeaxanthin. The analysis was based on two transcriptomic data sets [51, 52] with *BCH1* and *BCH2* as the bait genes. The analysis was conducted with default parameters. The functional annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and and corresponding information [56].

Additional File 8

Additional_File_8.pdf

Title: Phylogenetic trees retrieved through different alignment methods.

Description: Shown are phylogenetic trees of one selected OCG retrieved from CoExpPhylo annotated with Chalcone-isomerase like. The analysis was based on two transcriptomic data sets [51, 52] with *DFR*, *ANS*, *Bz2*, and *TT19* as bait genes. The global alignment was conducted with either MAFFT [43, 44] or MUSCLE [45]. The tree was inferred through FastTree [46, 47] and visualized in iTOL [50]. Different colors of strips next to the labels indicate major

phylogenetic groups. Sequences marked by green colored labels were collected via coexpression analysis. Trees are midpoint rooted.

Additional File 9

Additional_File_9.pdf

Title: Phylogenetic trees retrieved through different alignment methods.

Description: Shown are phylogenetic trees of one selected OCG retrieved from CoExpPhylo annotated with Chalcone-isomerase like. The analysis was based on two transcriptomic data sets [51, 52] with *DFR*, *ANS*, *Bz2*, and *TT19* as bait genes. The global alignment was conducted with MAFFT [43, 44] The tree was inferred either through FastTree [46, 47], RAxML [48] or IQ-TREE [49] and visualized in iTOL [50]. Sequences marked by green colored labels were collected via coexpression analysis. Trees are midpoint rooted.

References

1. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. Nat Plants. 2021;7:1571–8. https://doi.org/10.1038/s41477-021-01031-8.

2. Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ. 2020;8:e9954. https://doi.org/10.7717/peerj.9954.

3. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50:D387–90. https://doi.org/10.1093/nar/gkab1053.

4. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol Clifton NJ. 2019;1962:161–77. https://doi.org/10.1007/978-1-4939-9173-0_9.

5. Pucker B. Functional Annotation – How to Tackle the Bottleneck in Plant Genomics. Preprints. 2024. https://doi.org/10.20944/preprints202402.0645.v1.

6. Kliebenstein DJ, Osbourn A. Making new molecules – evolution of pathways for novel metabolites in plants. Curr Opin Plant Biol. 2012;15:415–23. https://doi.org/10.1016/j.pbi.2012.05.005.

7. 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. Plant J. 2007;50:660–77. https://doi.org/10.1111/j.1365-313X.2007.03078.x.

8. Stracke R, Jahns O, Keck M, Tohge T, Niehaus K, Fernie AR, et al. Analysis of PRODUCTION OF FLAVONOL GLYCOSIDES-dependent flavonol glycoside accumulation in Arabidopsis thaliana plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside accumulation. New Phytol. 2010;188:985–1000. https://doi.org/10.1111/j.1469-8137.2010.03421.x.

9. Gonzalez A, Zhao M, Leavitt JM, Lloyd AM. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. Plant J Cell Mol Biol. 2008;53:814–27. https://doi.org/10.1111/j.1365-313X.2007.03373.x.

10. Lloyd A, Brockman A, Aguirre L, Campbell A, Bean A, Cantero A, et al. Advances in the MYB–bHLH–WD Repeat (MBW) Pigment Regulatory Model: Addition of a WRKY Factor and Co-option of an Anthocyanin MYB for Betalain Regulation. Plant Cell Physiol. 2017;58:1431–41. https://doi.org/10.1093/pcp/pcx075.

11. Field B, Osbourn AE. Metabolic Diversification—Independent Assembly of Operon-Like Gene Clusters in Different Plants. Science. 2008;320:543–7. https://doi.org/10.1126/science.1154990.

12. Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, et al. A Papaver somniferum 10-Gene Cluster for Synthesis of the Anticancer Alkaloid Noscapine. Science. 2012;336:1704–8. https://doi.org/10.1126/science.1220757.

13. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters – from genetics to genomics. New Phytol. 2016;211:771–89. https://doi.org/10.1111/nph.13981.

14. Hakim SE, Choudhary N, Malhotra K, Peng J, Bültemeier A, Arafa A, et al. Phylogenomics and metabolic engineering reveal a conserved gene cluster in Solanaceae plants for withanolide biosynthesis. Nat Commun. 2025;16:6367. https://doi.org/10.1038/s41467-025-61686-1.

15. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. 2009;32:1633–51. https://doi.org/10.1111/j.1365-3040.2009.02040.x.

16. Gillis J, Pavlidis P. "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. PLoS Comput Biol. 2012;8:e1002444. https://doi.org/10.1371/journal.pcbi.1002444.

17. Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, et al. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 2015;43 Database issue:D996-1002. https://doi.org/10.1093/nar/gku1053.

18. Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, et al. Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci. 2019;116:2344–53. https://doi.org/10.1073/pnas.1817074116.

19. Zhai J, Tang Y, Yuan H, Wang L, Shang H, Ma C. A Meta-Analysis Based Method for Prioritizing Candidate Genes Involved in a Pre-specific Function. Front Plant Sci. 2016;7. https://doi.org/10.3389/fpls.2016.01914.

20. Winkel-Shirley B. Flavonoid Biosynthesis. A Colorful Model for Genetics, Biochemistry, Cell Biology, and Biotechnology. Plant Physiol. 2001;126:485–93.

21. Grotewold E. The genetics and biochemistry of floral pigments. Annu Rev Plant Biol. 2006;57:761–80. https://doi.org/10.1146/annurev.arplant.57.032905.105248.

22. Gu K-D, Wang C-K, Hu D-G, Hao Y-J. How do anthocyanins paint our horticultural products? Sci Hortic. 2019;249:257–62. https://doi.org/10.1016/j.scienta.2019.01.034.

23. Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR. The Arabidopsis TDS4 gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. Plant J. 2003;35:624–36. https://doi.org/10.1046/j.1365-313X.2003.01834.x.

24. Kitamura S, Shikazono N, Tanaka A. TRANSPARENT TESTA 19 is involved in the accumulation of both anthocyanins and proanthocyanidins in Arabidopsis. Plant J. 2004;37:104–14. https://doi.org/10.1046/j.1365-313X.2003.01943.x.

25. Pollastri S, Tattini M. Flavonols: old compounds for old roles. Ann Bot. 2011;108:1225–33. https://doi.org/10.1093/aob/mcr234.

26. Emiliani J, Grotewold E, Ferreyra MLF, Casati P. Flavonols Protect Arabidopsis Plants against UV-B Deleterious Effects. Mol Plant. 2013;6:1376–9. https://doi.org/10.1093/mp/sst021.

27. Naik J, Tyagi S, Rajput R, Kumar P, Pucker B, Bisht NC, et al. Flavonols affect the interrelated glucosinolate and camalexin biosynthetic pathways in Arabidopsis thaliana. J Exp Bot. 2024;75:219–40. https://doi.org/10.1093/jxb/erad391.

28. Grünig N, Horz JM, Pucker B. Diversity and Ecological Functions of Anthocyanins. 2024. https://doi.org/10.20944/preprints202408.2272.v1.

29. Jiang N, Grotewold E. Flavonoids and Derived Anthocyanin Pigments in Plants-Structure, Distribution, Function, and Methods for Quantification and Characterization. Cold Spring Harb Protoc. 2024. https://doi.org/10.1101/pdb.top108516.

30. Tohge T, de Souza LP, Fernie AR. Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. J Exp Bot. 2017;68:4013–28. https://doi.org/10.1093/jxb/erx177.

31. 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;6:927–38. https://doi.org/10.1038/s41929-023-01018-y.

32. Frommann J-F, Pucker B, Sielmann LM, Müller C, Weisshaar B, Stracke R, et al. Metabolic fingerprinting reveals roles of Arabidopsis thaliana BGLU1, BGLU3, and BGLU4 in glycosylation of various flavonoids. Phytochemistry. 2025;231:114338. https://doi.org/10.1016/j.phytochem.2024.114338.

33. Cazzonelli CI. Carotenoids in nature: insights from plants and beyond. Funct Plant Biol. 2011;38:833–47. https://doi.org/10.1071/FP11192.

34. Lu S, Li L. Carotenoid Metabolism: Biosynthesis, Regulation, and Beyond. J Integr Plant Biol. 2008;50:778–85. https://doi.org/10.1111/j.1744-7909.2008.00708.x.

35. Cunningham FXJr, Gantt E. GENES AND ENZYMES OF CAROTENOID BIOSYNTHESIS IN PLANTS. Annu Rev Plant Biol. 1998;49 Volume 49, 1998:557–83. https://doi.org/10.1146/annurev.arplant.49.1.557.

36. Zhou X, Rao S, Wrightstone E, Sun T, Lui ACW, Welsch R, et al. Phytoene Synthase: The Key Rate-Limiting Enzyme of Carotenoid Biosynthesis in Plants. Front Plant Sci. 2022;13. https://doi.org/10.3389/fpls.2022.884720.

833 37. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
834 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–
835 72. https://doi.org/10.1038/s41592-019-0686-2.

836 38. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array
837 programming with NumPy. Nature. 2020;585:357–62. https://doi.org/10.1038/s41586-020-
838 2649-2.

839 39. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function
840 using NetworkX. 2008.

841 40. Plotly Technologies Inc. Collaborative data science. 2015.

842 41. Tange O. GNU Parallel 20210822 ('Kabul'). 2021.
843 https://doi.org/10.5281/zenodo.5233953.

844 42. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using
845 DIAMOND. Nat Methods. 2021;18:366–8. https://doi.org/10.1038/s41592-021-01101-x.

846 43. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improve-
847 ments in Performance and Usability. Mol Biol Evol. 2013;30:772–80.
848 https://doi.org/10.1093/molbev/mst010.

849 44. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fou-
850 rier transform. Nucleic Acids Res. 2002;30:3059–66. https://doi.org/10.1093/nar/gkf436.

851 45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
852 Nucleic Acids Res. 2004;32:1792–7. https://doi.org/10.1093/nar/gkh340.

853 46. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for
854 Large Alignments. PLOS ONE. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.

855 47. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with
856 Profiles instead of a Distance Matrix. Mol Biol Evol. 2009;26:1641–50.
857 https://doi.org/10.1093/molbev/msp077.

858 48. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-
859 friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35:4453–
860 5. https://doi.org/10.1093/bioinformatics/btz305.

861 49. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al.
862 IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.
863 Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015.

864 50. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree
865 display and annotation tool. Nucleic Acids Res. 2024;52:W78–82.
866 https://doi.org/10.1093/nar/gkae268.

51. Pucker B, Fiene N, Choudhary N, Borchert M, Khatun N. Collection of plant gene expression data. 2024. https://doi.org/10.24355/dbbs.084-202409160820-0.

52. Pucker B, Grünig N, Khatun N. Gene Expression Analysis Across Plantae. 2025. https://doi.org/10.24355/dbbs.084-202501230512-0.

53. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7. https://doi.org/10.1038/nbt.3519.

54. Pucker B, Iorizzo M. Apiaceae FNS I originated from F3H through tandem gene duplication. PLOS ONE. 2023;18:e0280155. https://doi.org/10.1371/journal.pone.0280155.

55. Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML. Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform. J Proteome Res. 2015;14:1880–7. https://doi.org/10.1021/pr501286b.

56. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89:789–804. https://doi.org/10.1111/tpj.13415.

57. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40:D1202–10. https://doi.org/10.1093/nar/gkr1090.

58. Pucker B, Reiher F, Schilbert HM. Automatic Identification of Players in the Flavonoid Biosynthesis with Application on the Biomedicinal Plant Croton tiglium. Plants. 2020;9:1103. https://doi.org/10.3390/plants9091103.

59. Rempel A, Choudhary N, Pucker B. KIPEs3: Automatic annotation of biosynthesis pathways. PLOS ONE. 2023;18:e0294342. https://doi.org/10.1371/journal.pone.0294342.

60. Seitz C, Ameres S, Forkmann G. Identification of the molecular basis for the functional difference between flavonoid 3’-hydroxylase and flavonoid 3’,5’-hydroxylase. FEBS Lett. 2007;581:3429–34. https://doi.org/10.1016/j.febslet.2007.06.045.

61. Ramsay NA, Glover BJ. MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. Trends Plant Sci. 2005;10:63–70. https://doi.org/10.1016/j.tplants.2004.12.011.

62. Hichri I, Barrieu F, Bogs J, Kappel C, Delrot S, Lauvergeat V. Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. J Exp Bot. 2011;62:2465–83. https://doi.org/10.1093/jxb/erq442.

63. Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L. The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. Plant Cell. 2001;13:2099–114. https://doi.org/10.1105/tpc.010098.

64. Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C. Activation Tagging Identifies a Conserved MYB Regulator of Phenylpropanoid Biosynthesis. Plant Cell. 2000;12:2383–93. https://doi.org/10.1105/tpc.12.12.2383.

65. Stracke R, Werber M, Weisshaar B. The *R2R3-MYB* gene family in *Arabidopsis thaliana*. Curr Opin Plant Biol. 2001;4:447–56. https://doi.org/10.1016/S1369-5266(00)00199-0.

66. Li M, Guo L, Wang Y, Li Y, Jiang X, Liu Y, et al. Molecular and biochemical characterization of two 4-coumarate: CoA ligase genes in tea plant (Camellia sinensis). Plant Mol Biol. 2022;109:579–93. https://doi.org/10.1007/s11103-022-01269-6.

67. Liu T, Yao R, Zhao Y, Xu S, Huang C, Luo J, et al. Cloning, Functional Characterization and Site-Directed Mutagenesis of 4-Coumarate: Coenzyme A Ligase (4CL) Involved in Coumarin Biosynthesis in Peucedanum praeruptorum Dunn. Front Plant Sci. 2017;8. https://doi.org/10.3389/fpls.2017.00004.

68. Wang X, Wang J, Cui H, Yang W, Yu B, Zhang C, et al. The UDP-glycosyltransferase MtUGT84A1 regulates anthocyanin accumulation and plant growth *via* JA signaling in *Medicago truncatula*. Environ Exp Bot. 2022;201:104972. https://doi.org/10.1016/j.envexpbot.2022.104972.

69. Farrow SC, Facchini PJ. Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. Front Plant Sci. 2014;5. https://doi.org/10.3389/fpls.2014.00524.

70. Zhang J-R, Trossat-Magnin C, Bathany K, Negroni L, Delrot S, Chaudière J. Oxidative Transformation of Dihydroflavonols and Flavan-3-ols by Anthocyanidin Synthase from Vitis vinifera. Molecules. 2022;27:1047. https://doi.org/10.3390/molecules27031047.

71. J. Turnbull J, J. Sobey W, T. Aplin R, Hassan A, L. Firmin J, J. Schofield C, et al. Are anthocyanidins the immediate products of anthocyanidin synthase? Chem Commun. 2000;0:2473–4. https://doi.org/10.1039/B007594I.

72. Tohge T, Nishiyama Y, Hirai MY, Yano M, Nakajima J, Awazuhara M, et al. Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants overexpressing an MYB transcription factor. Plant J. 2005;42:218–35. https://doi.org/10.1111/j.1365-313X.2005.02371.x.

73. 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. 2021;12. https://doi.org/10.3389/fpls.2021.733762.

74. Martens S, Forkmann G, Britsch L, Wellmann F, Matern U, Lukacin R. Divergent evolution of flavonoid 2-oxoglutarate-dependent dioxygenases in parsley. FEBS Lett. 2003;544:93–8. https://doi.org/10.1016/s0014-5793(03)00479-4.

75. Pucker B. Automatic identification and annotation of MYB gene family members in plants. BMC Genomics. 2022;23:220. https://doi.org/10.1186/s12864-022-08452-5.

76. Debeaujon I, Léon-Kloosterziel KM, Koornneef M. Influence of the Testa on Seed Dormancy, Germination, and Longevity in Arabidopsis1. Plant Physiol. 2000;122:403–14. https://doi.org/10.1104/pp.122.2.403.

77. Galway ME, Masucci JD, Lloyd AM, Walbot V, Davis RW, Schiefelbein JW. The *TTG* Gene Is Required to Specify Epidermal Cell Fate and Cell Patterning in the *Arabidopsis* Root. Dev Biol. 1994;166:740–54. https://doi.org/10.1006/dbio.1994.1352.

78. Li C, Zhang B, Chen B, Ji L, Yu H. Site-specific phosphorylation of TRANSPARENT TESTA GLABRA1 mediates carbon partitioning in Arabidopsis seeds. Nat Commun. 2018;9:571. https://doi.org/10.1038/s41467-018-03013-5.

79. Tsuchiya Y, Nambara E, Naito S, McCourt P. The FUS3 transcription factor functions through the epidermal regulator TTG1 during embryogenesis in Arabidopsis. Plant J. 2004;37:73–81. https://doi.org/10.1046/j.1365-313X.2003.01939.x.

80. Choudhary N, Pucker B. Conserved amino acid residues and gene expression patterns associated with the substrate preferences of the competing enzymes FLS and DFR. PLOS ONE. 2024;19:e0305837. https://doi.org/10.1371/journal.pone.0305837.

81. 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;131:259–68. https://doi.org/10.1023/A:1024018729349.

82. 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. 2016;6. https://doi.org/10.3389/fpls.2015.01257.

83. Zhan X, Qiu T, Zhang H, Hou K, Liang X, Chen C, et al. Mass spectrometry imaging and single-cell transcriptional profiling reveal the tissue-specific regulation of bioactive ingredient biosynthesis in Taxus leaves. Plant Commun. 2023;4. https://doi.org/10.1016/j.xplc.2023.100630.

84. Ntelkis N, Goossens A, Šola K. Cell type-specific control and post-translational regulation of specialized metabolism: opening new avenues for plant metabolic engineering. Curr Opin Plant Biol. 2024;81:102575. https://doi.org/10.1016/j.pbi.2024.102575.

85. Pucker B, Walker-Hale N, Dzurlic J, Yim WC, Cushman JC, Crum A, et al. Multiple mechanisms explain loss of anthocyanins from betalain-pigmented Caryophyllales, including repeated wholesale loss of a key anthocyanidin synthesis enzyme. New Phytol. 2024;241:471–89. https://doi.org/10.1111/nph.19341.

86. Chen W, Miao Y, Ayyaz A, Huang Q, Hannan F, Zou H-X, et al. Anthocyanin accumulation enhances drought tolerance in purple-leaf *Brassica napus*: Transcriptomic, metabolomic, and physiological evidence. Ind Crops Prod. 2025;223:120149. https://doi.org/10.1016/j.indcrop.2024.120149.

87. Camara B. [32] Plant phytoene synthase complex: Component enzymes, immunology, and biogenesis. In: Methods in Enzymology. Academic Press; 1993. p. 352–65. https://doi.org/10.1016/0076-6879(93)14079-X.

88. Hirschberg J. Carotenoid biosynthesis in flowering plants. Curr Opin Plant Biol. 2001;4:210–8. https://doi.org/10.1016/S1369-5266(00)00163-1.

89. Chamovitz D, Pecker I, Sandmann G, Böger P, Hirschberg J. Cloning a Gene Coding for Norflurazon Resistance in Cyanobacteria. Z Für Naturforschung C. 1990;45:482–6. https://doi.org/10.1515/znc-1990-0531.

90. Cazzaniga S, Li Z, Niyogi KK, Bassi R, Dall'Osto L. The Arabidopsis szl1 Mutant Reveals a Critical Role of β-Carotene in Photosystem I Photoprotection1[C][W]. Plant Physiol. 2012;159:1745–58. https://doi.org/10.1104/pp.112.201137.

91. Tian L, Musetti V, Kim J, Magallanes-Lundback M, DellaPenna D. The Arabidopsis LUT1 locus encodes a member of the cytochrome p450 family that is required for carotenoid epsilon-ring hydroxylation activity. Proc Natl Acad Sci U S A. 2004;101:402–7. https://doi.org/10.1073/pnas.2237237100.

92. Fiore A, Dall'Osto L, Fraser PD, Bassi R, Giuliano G. Elucidation of the β-carotene hydroxylation pathway in *Arabidopsis thaliana*. FEBS Lett. 2006;580:4718–22. https://doi.org/10.1016/j.febslet.2006.07.055.

93. Tian L, Magallanes-Lundback M, Musetti V, DellaPenna D. Functional Analysis of β- and ε-Ring Carotenoid Hydroxylases in Arabidopsis. Plant Cell. 2003;15:1320–32. https://doi.org/10.1105/tpc.011403.

94. Tian L, DellaPenna D. Characterization of a second carotenoid β-hydroxylase gene from Arabidopsis and its relationship to the LUT1 locus. Plant Mol Biol. 2001;47:379–88. https://doi.org/10.1023/A:1011623907959.

95. Sun Z, Gantt E, Cunningham FX. Cloning and Functional Analysis of the β-Carotene Hydroxylase of Arabidopsis thaliana*. J Biol Chem. 1996;271:24349–52. https://doi.org/10.1074/jbc.271.40.24349.

96. Stanley L, Yuan Y-W. Transcriptional Regulation of Carotenoid Biosynthesis in Plants: So Many Regulators, So Little Consensus. Front Plant Sci. 2019;10. https://doi.org/10.3389/fpls.2019.01017.

97. Dall'Osto L, Lico C, Alric J, Giuliano G, Havaux M, Bassi R. Lutein is needed for efficient chlorophyll triplet quenching in the major LHCII antenna complex of higher plants and effective photoprotection in vivounder strong light. BMC Plant Biol. 2006;6:32. https://doi.org/10.1186/1471-2229-6-32.

98. Nisar N, Li L, Lu S, Khin NC, Pogson BJ. Carotenoid Metabolism in Plants. Mol Plant. 2015;8:68–82. https://doi.org/10.1016/j.molp.2014.12.007.

99. Schwartz SH, Tan BC, Gage DA, Zeevaart JA, McCarty DR. Specific oxidative cleavage of carotenoids by VP14 of maize. Science. 1997;276:1872–4. https://doi.org/10.1126/science.276.5320.1872.