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 anthocyanin, proanthocyanidin, and flavonol biosynthesis, 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, 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 plants 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 as recently reported for withanolide biosynthesis in *Solanacae* or thalianol biosynthesis in *Arabidopsis thaliana* [11, 12]. 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 [13]. 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 [14]. 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 [15, 16]. Anthocyanins confer colors to fruits and flowers [17], proanthocyanidins (PAs) are responsible for seed coat pigmentation [18, 19] and flavonols are known as UV-response compounds [20, 21] and have been associated with regulatory functions [22]. Due to a plethora of modification reactions catalyzed by numerous (promiscuous) enzymes, there are thousands of different flavonoid derivatives in plants [23, 24]. Despite decades of intense research on the flavonoid biosynthesis [15, 16, 25], there is still potential for new discoveries as demonstrated by recent publications [26, 27].

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. The potential is demonstrated by analyses on the anthocyanin, proanthocyanidin, and flavonol biosynthesis.

Implementation

CoExpPhylo was written in Python3 using SciPy v1.13.0 [28], NumPy v1.26.4 [29], Networkx v3.3 [30] and Plotly v5.22.0 [31] as well as the utility GNU parallel 20210822 [32] through bash scripting. CoExpPhylo utilizes different external tools: Several local alignments are conducted via DIAMOND v2.0.14.152 [33], global alignments are executed using MAFFT v7.490 [34, 35] or optionally by MUSCLE v3.8.1551 [36] and phylogenetic trees are inferred by FastTree v2.1.11 [37, 38], RAxML-NG v1.2.2 [39] or IQ-TREE v2.0.7 [40]. Optionally, the final tree files can be uploaded to iTOL [41].

Input data collection

The analysis is conducted based on transcriptomic datasets of 240 species from various orders (Figure 1, Additional File 1) [42, 43]. 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 [44] and subsequently merged into a single count table per species with custom Python scripts [45]. 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). If the gene IDs did not contain any transcript information, transcripts of one gene were determined and reduced via isoform_purger.py v0.21 (https://github.com/bpucker/PBBtools).

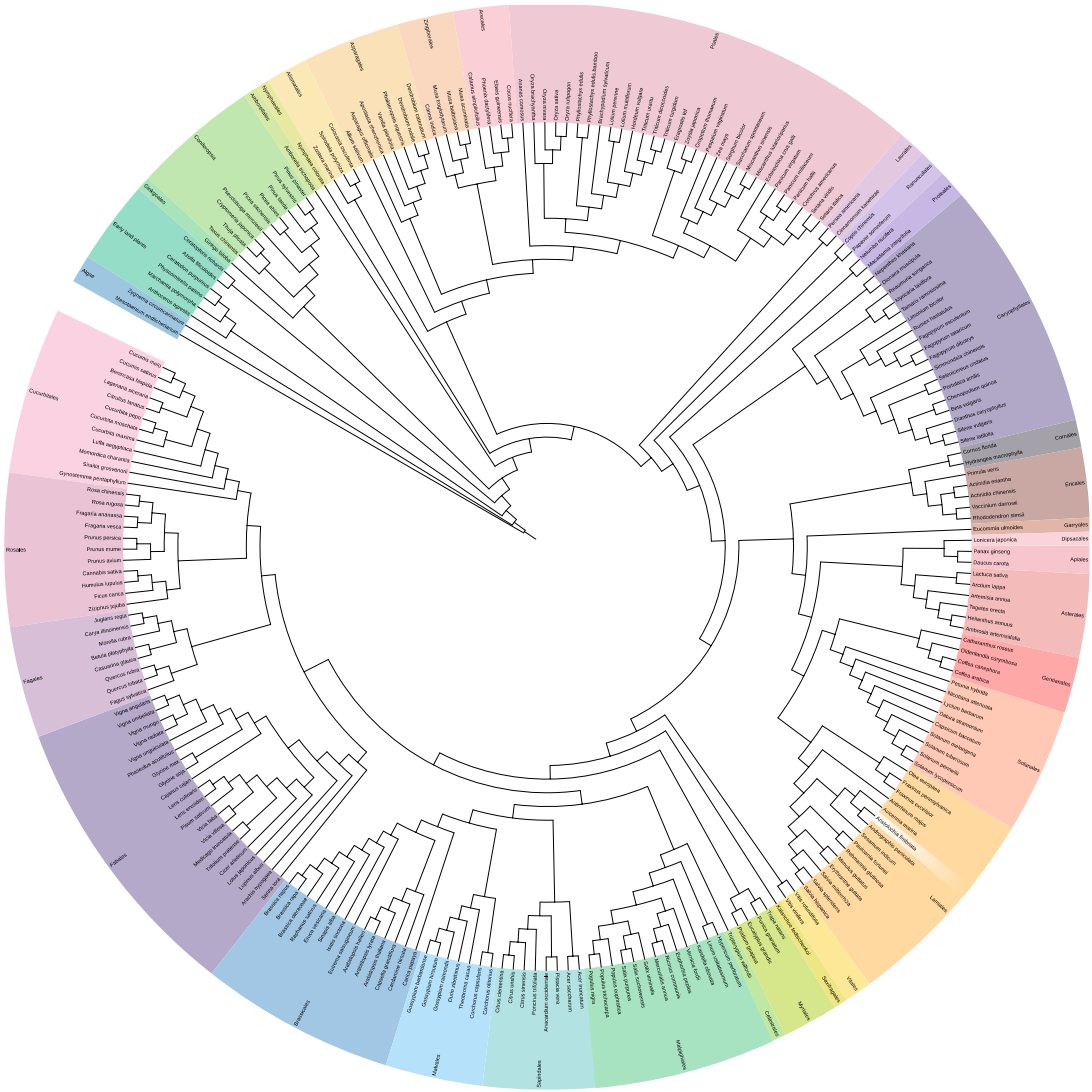

Figure 1: Taxonomic grouping of species included in this study.
Colored ranges represent the taxonomic orders of the species. For better readability, the classification of early land plants and *Coniferopsida* has been simplified. *Aristolochia fimbriata* appears to be misplaced, as it belongs to the *Piperales*, which would be placed close to the *Laurales*. The figure was generated using iTOL [41].

A crucial input for the analysis is a config file that contains all the information about all input files. The columns in this file need to be comma-separated. Each row describes one data set. The columns are: ID, TPM file, CDS file and name of file with IDs of bait sequences. Each row should represent one dataset. The ID column usually contains the species names which are later used as a prefix for all sequences of the respective species to avoid ambiguities when

combining sequences of different species in one file. The TPM column contains paths to count tables where the first column contains the gene IDs and the first row contains the sample names. All other fields in this table are gene expression values. The CDS column contains the paths to a multiple FASTA file with coding sequences for the respective dataset. The fourth column contains the path to a bat sequence ID file. This file contains the IDs of the gene candidate sequences that should be used as baits. As the research aims to find unknown genes of the anthocyanin biosynthesis, any late players of the pathway are unequivocally appropriate bait candidates. Preferably, any genes that are not part of the biosynthesis of any other group should be used. These predefined genes should be orthologs. The bait ID files should be basic text files with one sequence ID given per line.

To annotate each OCG properly, Araport11 *Arabidopsis thaliana* Col-0 reference polypeptide sequences and annotation terms derived from TAIR10 and Araport11 [46, 47] 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.

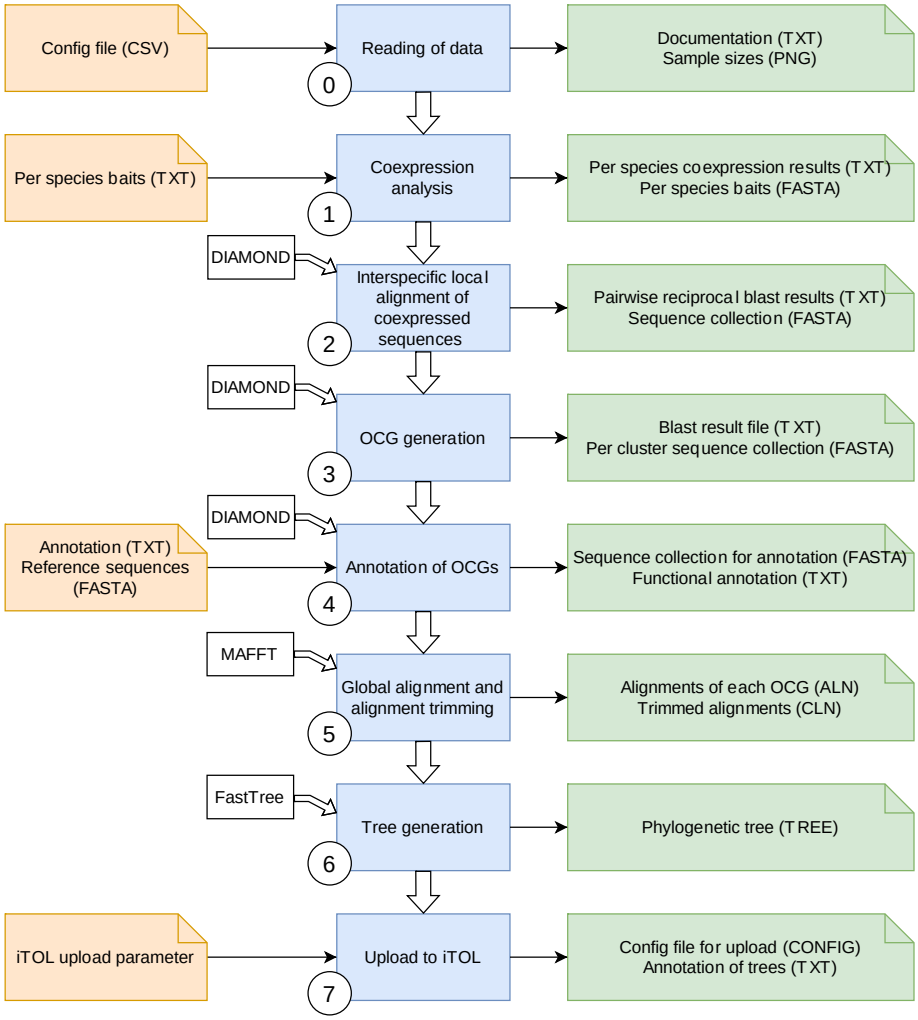

Figure 2: Workflow of CoExpPhylo.
Shown is a schematic overview of the pipeline of coexp_phylo.py. The workflow steps are represented in blue, input files and their file extension are denoted in orange, whereas the output files and their file extensions are colored in green. White boxes indicate the 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 [28] in Python to calculate the Spearman's rank correlation coefficient r_s and the corresponding adjusted p-value. The default correlation coefficient cutoff is set to 0.7 but can be adjusted using the argument --r. By specifying the argument --p, the corresponding p-value can be varied, the default is 0.05. 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 coexpression 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 [33]. 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: Firstly, the e-value of the alignment must not exceed 10^{-5}. Secondly, the bit score must be greater than 100, thirdly, the length score must surpass 100, and fourthly, the similarity score must be greater than 82. Those values can be adjusted by setting --evalue, --scorecut, --lencut as well as --simcut, respectively. However, all collected sequences are written into one FASTA file.

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 [30]. 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 de fined 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 [33] 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 [34, 35] or optionally with MUSCLE [36]. 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 [37, 38] is utilized with the options -wag and -nosupport. Alternatively, RAxML-NG [39] using the model "LG+8G+F" or IQTREE [40] can be applied for this step. To enable the upload to common phylogenetic tree visualization programs, especially the batch upload to iTOL [41], 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 [41]. 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 1). 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 [41] 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 1: 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 [41]

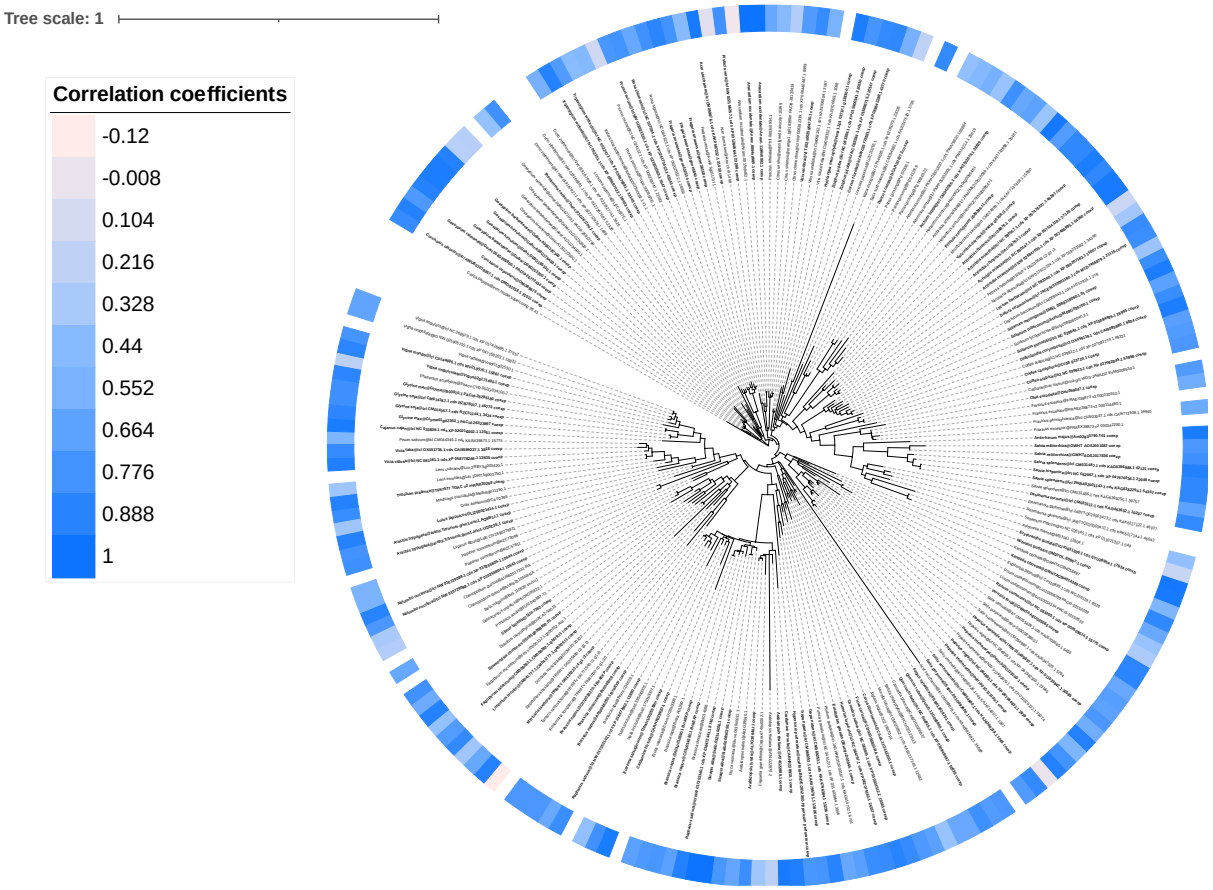

Figure 3: Exemplary representation of the visualization of one Step via iTOL [41]. Shown is the result of Orthologous Coexpressed Group (OCG) 0009 retrieved through CoExpPhylo with respect to the anthocyanin biosynthesis synthesis. CoExpPhylo was conducted based on the transcriptomic data of 239 species [42, 43], 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 [48, 49]. 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 [34, 35]. The trees were constructed with FastTree [37, 38].

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 4).

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) [15, 50]. 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) [15, 26]. 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, 51]. The MYB component determines the target gene, forming a complex with bHLH42 and TTG1 [52]. 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 [53]. MYB75, MYB90, MYB113, and MYB114 are responsible for anthocyanin biosynthesis regulation [9, 54, 55].

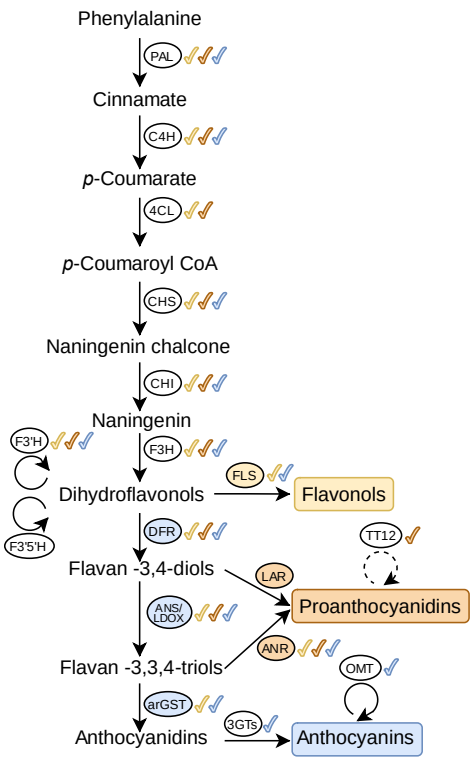

Figure 4: Schematic overview of different branches of the flavonoid biosynthesis and the detection via CoExpPhylo.

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. Enzymes with a colored background were used as input baits for analysis of the respective biosynthetic branch. Colored ticks indicate retrieved Orthologous Coexpressed Groups in the respective branch.

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 2) 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 functional roles beyond flavonoid biosynthesis or potential limitations in coexpression-based detection. Due to the phylogenetic divergence of the input species, multiple OCGs were generated for certain genes. For instance, F3H was represented by three OCGs (0017, 0037, and 0069), corresponding to different taxonomic groups: one cluster contained sequences from *Eudicotyledons*, another from *Coniferopsida* and the third from species belonging to the order *Poales*.

368 The initial bait genes (*DFR*, *ANS*/*LDOX,* and *arGSTs*) were also retrieved as OCGs. Furthermore,
369 the OCGs annotated as UDP-dependent glycosyltransferase UGT78D2, along with the
370 anthocyanin-decorating enzymes SCPL10, AT1, and O-methyltransferase 1 (OMT), were
371 identified. Two additional OCGs (0072 and 0145) were annotated with UDP-dependent
372 glycosyltransferases UGT80B1 and UGT84A1, respectively. While these UGTs have not yet
373 been reported to be involved in anthocyanin modification, further analysis of the sequences
374 within these OCGs may reveal additional glycosyltransferases pertinent to anthocyanin
375 biosynthesis. However, the annotation process is based on *Arabidopsis thaliana* reference
376 sequences, which may explain certain limitations. Genes or functions absent in *A. thaliana*
377 cannot be annotated correctly, so that the annotation of UGT80B1 and UGT84A1 might be an
378 artifact thus suggesting that some other UGT lineages could contribute to the diversity of
379 anthocyanins. Similarly, a potential OCG containing sequences encoding F3'5'H was not
380 identified as *A. thaliana* lacks this enzyme, highlighting the importance of a multi-species
381 reference database for the annotation.

382 Additionally, transcription factors associated with anthocyanin biosynthesis regulation,
383 including MYB90, MYB12, and the bHLH transcription factor bHLH42/TT8, were identified.
384 Notably, FLS was also detected, despite its primary role in flavonol biosynthesis rather than
385 anthocyanin production. However, FLS; F3H and ANS/LDOX belong to the 2-oxoglutarate-
386 dependent dioxygenase (2ODD) superfamily whereby functional redundancy might occur
387 [56].

388 **Proanthocyanidins**

389 PAs, also known as condensed tannins, contribute to plant defense and seed coat
390 pigmentation. Their biosynthesis shares multiple enzymatic steps with anthocyanin
391 biosynthesis, making it challenging to delineate genes specific to this pathway. To assess
392 CoExpPhylo's ability to resolve this complexity, late stage PA biosynthetic genes, including *LAR*
393 and *ANR* (also known as *BANYLUS* (*BAN*)) were used as baits.

394 The initial bait gene *BAN* was detected three times among the final OCGs (Additional File 3).
395 As *A. thaliana* does not have a LAR, no OCG was annotated accordingly. However, all genes
396 encoding enzymes involved in the straight PA biosynthesis, starting with PAL, were successfully
397 retrieved within annotated OCGs (Figure 4). Notably, genes associated with closely related
398 pathways, such as arGSTs and FL), were not retrieved, suggesting a certain specificity of
399 CoExpPhylo in distinguishing metabolic branches. Additionally, one OCG was annotated with
400 the bHLH transcription factor bHLH42/TT8, which is involved in PA biosynthesis.

401 The main MYB transcription factor regulating PA biosynthesis, MYB123, was not identified as
402 a separate OCG. This highlights a limitation of CoExpPhylo in resolving individual lineages of
403 large gene families such as MYB transcription factors. Due to high sequence similarity within
404 these families, the delineation of OCGs may be influenced by the chosen similarity threshold,
405 potentially leading to the merging of closely related sequences. Possibly, the OCG annotated
406 with MYB5 (0058) contained MYB123 sequences along with other members of MYB subgroup

407 5. To achieve a more precise resolution of large transcription factor families, dedicated tools
408 like MYB_annotator might show superior performance [57]. Furthermore, the WD40
409 transcription factor TTG1 was not captured as an OCG. This may be due to its diverse
410 functional roles, spanning various developmental processes and specialized metabolism [53,
411 58–61] which lead to a constant expression pattern that is not suitable for detection via
412 coexpression analysis.

413 **Flavonols**

414 Flavonols, a subclass of flavonoids, function as antioxidants and UV protectants in plants. They
415 serve as precursors for multiple downstream metabolic pathways, making their biosynthetic
416 gene network particularly intricate. To investigate the flavonol branch, FLS, a key enzyme
417 converting dihydroflavonols into flavonols, was selected as a bait sequence. Since FLS
418 competes with DFR for its substrate [62–64], its expression pattern provides a means to
419 distinguish flavonol biosynthesis from related pathways.

420 All core genes of the flavonol biosynthesis pathway were retrieved within annotated OCGs
421 (Additional File 4). However, additional OCGs were identified containing sequences annotated
422 with genes typically associated with anthocyanin biosynthesis, such as *DFR* and *ANS*/*LDOX*, or
423 PA biosynthesis (e.g. *BAN*) which do not directly participate in flavonol production. This
424 suggests that the separation of the flavonol and anthocyanin branches is not entirely distinct
425 in the coexpression analysis, potentially reflecting their shared metabolic precursors and
426 regulatory elements. Moreover, this overlap might be influenced by the diversity of the cell
427 types included in the transcriptomic data. Previous studies have demonstrated cell type-
428 specific expression patterns for different branches of flavonoid biosynthesis [65, 66].

429 In addition to enzymatic genes, transport and modification-related genes were identified. The
430 transporter DTX35, implicated in flavonol transport, was retrieved as an OCG. Moreover, UDP-
431 dependent glycosyltransferases UGT78D2, UGT84A1, and UGT84A2 were identified,
432 suggesting potential roles in flavonol glycosylation. Regarding transcription factors, the bHLH
433 regulator bHLH42/TT8 and MYB12 were captured as OCGs. However, MYB11 and MYB111,
434 which together with MYB12 constitute subgroup 7 of the R2R3-MYB family [7], were not
435 detected. The OCG annotated with MYB12 contained 336 sequences, indicating that the
436 boundaries within this gene family might be blurred, potentially leading to the merging of
437 multiple MYB subgroup 7 members into a single OCG.

438 Tool benchmarking

439 CoExpPhylo allows the user to select either MAFFT [34, 35] or MUSCLE [36] for the multiple
440 sequence alignment in Step 5. To evaluate the impact of the alignment method on the
441 phylogenetic analysis, several PSS were processed using both tools. A representative dataset
442 comprising 43 sequences revealed key differences between the alignments (Additional File
443 5). Five larger groups could be identified that contained the same sequences but were
444 positioned differently and the internal branches within these groups were arranged
445 differently. The preservation of major groupings shows that both tools are consistent with the

446 overall evolutionary structure but differ in detailed phylogenetic distances, i.e. the
447 connections between the groups are interpreted differently. These discrepancies stem from
448 the different algorithms and scoring systems that both alignment tools use.

449 In contrast, a comparison of the different available tools for tree inference from these
450 alignments – FastTree [37, 38], RAxML [39] or IQ-TREE [40] – revealed greater variability in
451 sequence placement. Although the overall tree topology remained consistent, topological
452 details varied across different tree-building approaches (Additional File 6). Notably, the
453 phylogenetic differences introduced by alternative alignment methods exceeded those
454 introduced by different tree inference tools.

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

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

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

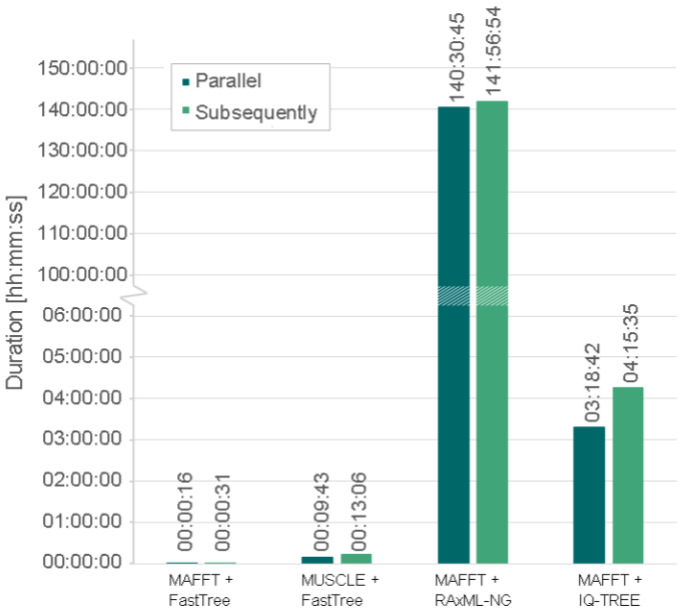

Figure 5: Comparison of the runtime performance of global alignment and phylogenetic tree construction. Shown are the runtime performances of the global alignment step combined with the construction of phylogenetic trees. Three representative sequence collections were selected (containing 43, 113, and 212 sequences) and processed in parallel (dark green) or subsequently (light green). The global alignment was conducted either with MAFFT v7.490 [34, 35] or MUSCLE v3.8.1551 [36], the phylogenetic tree was constructed with FastTree v2.1.11 [37, 38], RAxML-NG v1.2.2 [39], or IQ-TREE v2.0.7 [40]. The vertical axis was split between 6 and 100 hours in order to visualize all data points.

Conclusions

These results demonstrate that CoExpPhylo effectively identifies key components of specialized biosynthetic pathways across multiple species. 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 successfully retrieves transcription factors and transporters, providing a more comprehensive view of pathway regulation.

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 provide a solid foundation for identifying candidate genes within a biosynthetic pathway. 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.

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.

Availability and requirements

Project name: CoExpPhylo

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

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

Programming language: Python, Bash

Other requirements: Python 3.0 or higher

License: GNU GPL

Any restrictions to use by non-academics: not applicable

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
C4H	Cinnamate 4-Hydroxylase
CDS	Coding Sequence
CHI	Chalcone Isomersae
CHS	Chalcone Synthase
DFR	Dihydroflavonol 4-Reductase
F3'5'H	Flavonoid 3’,5'-Hydroxylase
F3'H	Flavonoid 3’-Hydroxylase
F3H	Flavanone 3-Hydroxylase
FLS	Flavonol Synthase
LAR	Leucoanthocyanidin Reductase
LDOX	Leucoanthocyanidin Dioxygenase
NCBI	National Center for Biotechnology Information
OCG	Orthologous Coexpressed Group
OMT	O-Methyltransferase
PA	Proanthocyanidin
PAL	Phenylalanine Ammonia-Lyase
SRA	Sequence Read Archive
TPM	Transcripts Per Million
TT	Transparent Testa
UFGT	UDP-Glucose:Flavonoid Glucosyltransferase

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 [42, 43] 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

Titel: 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 [42, 43].

Additional File 2

Additional_File_2.txt

Titel: 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 [42, 43] 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 annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and annotations [46].

Additional File 3

Additional_File_3.txt

Titel: 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 [42, 43] with *LAR* and *ANR* as bait genes. The analysis was conducted with default parameters. The annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and annotations [46].

Additional File 4

Additional_File_4.txt

Titel: 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 [42, 43] with *FLS* as the bait gene. The analysis was conducted with default parameters. The annotations refer to Araport11 *A. thaliana* Col-0 reference polypeptide sequences and annotations [46].

Additional File 5

Additional_File_5.pdf

Titel: 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 [42, 43] with *DFR*, *ANS*, *Bz2*, and *TT19* as bait genes. The global alignment was conducted with either MAFFT [34, 35] or MUSCLE [36]. The tree was inferred through FastTree [37, 38] and visualized in iTOL [41]. 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 6

Additional_File_6.pdf

Titel: 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 [42, 43] with *DFR*, *ANS*, *Bz2*, and *TT19* as bait genes. The global alignment was conducted with MAFFT [34, 35] The tree was inferred either through FastTree [37, 38], RAxML [39] or IQ-TREE [40] and visualized in iTOL [41]. Sequences marked by green colored labels were collected via coexpression analysis. Trees are midpoint rooted.

606 # References

607 1. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20
608 years of plant genome sequencing. Nat Plants. 2021;7:1571–8.

609 2. Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential
610 risks and rewards. PeerJ. 2020;8:e9954.

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

613 4. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron
614 Position Conservation and RNA-seq Data. Methods Mol Biol Clifton NJ. 2019;1962:161–77.

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

617 6. Kliebenstein DJ, Osbourn A. Making new molecules – evolution of pathways for novel
618 metabolites in plants. Curr Opin Plant Biol. 2012;15:415–23.

619 7. Stracke R, Ishihara H, Huep G, Barsch A, Mehrtens F, Niehaus K, et al. Differential regulation
620 of closely related R2R3-MYB transcription factors controls flavonol accumulation in different
621 parts of the Arabidopsis thaliana seedling. Plant J. 2007;50:660–77.

622 8. Stracke R, Jahns O, Keck M, Tohge T, Niehaus K, Fernie AR, et al. Analysis of PRODUCTION
623 OF FLAVONOL GLYCOSIDES-dependent flavonol glycoside accumulation in Arabidopsis
624 thaliana plants reveals MYB11-, MYB12- and MYB111-independent flavonol glycoside
625 accumulation. New Phytol. 2010;188:985–1000.

626 9. Gonzalez A, Zhao M, Leavitt JM, Lloyd AM. Regulation of the anthocyanin biosynthetic
627 pathway by the TTG1/bHLH/Myb transcriptional complex in Arabidopsis seedlings. Plant J Cell
628 Mol Biol. 2008;53:814–27.

629 10. Lloyd A, Brockman A, Aguirre L, Campbell A, Bean A, Cantero A, et al. Advances in the
630 MYB–bHLH–WD Repeat (MBW) Pigment Regulatory Model: Addition of a WRKY Factor and
631 Co-option of an Anthocyanin MYB for Betalain Regulation. Plant Cell Physiol. 2017;58:1431–
632 41.

633 11. Field B, Osbourn AE. Metabolic Diversification—Independent Assembly of Operon-Like
634 Gene Clusters in Different Plants. Science. 2008;320:543–7.

635 12. Hakim SE, Choudhary N, Malhotra K, Peng J, Arafa A, Bültemeier A, et al. Phylogenomics
636 and metabolic engineering reveal a conserved gene cluster in Solanaceae plants for
637 withanolide biosynthesis. 2024;:2024.09.27.614867.

638 13. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, et al. Co-expression
639 tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell
640 Environ. 2009;32:1633–51.

641 14. Gillis J, Pavlidis P. “Guilt by Association” Is the Exception Rather Than the Rule in Gene
642 Networks. PLoS Comput Biol. 2012;8:e1002444.

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

645 16. Grotewold E. The genetics and biochemistry of floral pigments. Annu Rev Plant Biol.
646 2006;57:761–80.

647 17. Gu K-D, Wang C-K, Hu D-G, Hao Y-J. How do anthocyanins paint our horticultural products?
648 Sci Hortic. 2019;249:257–62.

649 18. Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, Ashton AR. The Arabidopsis TDS4 gene
650 encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin
651 synthesis and vacuole development. Plant J. 2003;35:624–36.

652 19. Kitamura S, Shikazono N, Tanaka A. TRANSPARENT TESTA 19 is involved in the accumulation
653 of both anthocyanins and proanthocyanidins in Arabidopsis. Plant J. 2004;37:104–14.

654 20. Pollastri S, Tattini M. Flavonols: old compounds for old roles. Ann Bot. 2011;108:1225–33.

655 21. Emiliani J, Grotewold E, Ferreyra MLF, Casati P. Flavonols Protect Arabidopsis Plants against
656 UV-B Deleterious Effects. Mol Plant. 2013;6:1376–9.

657 22. Naik J, Tyagi S, Rajput R, Kumar P, Pucker B, Bisht NC, et al. Flavonols affect the interrelated
658 glucosinolate and camalexin biosynthetic pathways in Arabidopsis thaliana. J Exp Bot.
659 2024;75:219–40.

660 23. Grünig N, Horz JM, Pucker B. Diversity and Ecological Functions of Anthocyanins. 2024.

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

664 25. Tohge T, de Souza LP, Fernie AR. Current understanding of the pathways of flavonoid
665 biosynthesis in model and crop plants. J Exp Bot. 2017;68:4013–28.

666 26. Eichenberger M, Schwander T, Hüppi S, Kreuzer J, Mittl PRE, Peccati F, et al. The catalytic
667 role of glutathione transferases in heterologous anthocyanin biosynthesis. Nat Catal.
668 2023;6:927–38.

669 27. Frommann J-F, Pucker B, Sielmann LM, Müller C, Weisshaar B, Stracke R, et al. Metabolic
670 fingerprinting reveals roles of Arabidopsis thaliana BGLU1, BGLU3, and BGLU4 in glycosylation
671 of various flavonoids. Phytochemistry. 2025;231:114338.

672 28. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
673 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–
674 72.

675 29. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array
676 programming with NumPy. Nature. 2020;585:357–62.

677 30. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function
678 using NetworkX. 2008.

679 31. Plotly Technologies Inc. Collaborative data science. 2015.

680 32. Tange O. GNU Parallel 20210822 ('Kabul'). 2021.

681 33. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using
682 DIAMOND. Nat Methods. 2021;18:366–8.

683 34. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
684 Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.

685 35. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast
686 Fourier transform. Nucleic Acids Res. 2002;30:3059–66.

687 36. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
688 Nucleic Acids Res. 2004;32:1792–7.

689 37. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for
690 Large Alignments. PLOS ONE. 2010;5:e9490.

691 38. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with
692 Profiles instead of a Distance Matrix. Mol Biol Evol. 2009;26:1641–50.

693 39. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-
694 friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35:4453–
695 5.

696 40. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al.
697 IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.
698 Mol Biol Evol. 2020;37:1530–4.

699 41. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree
700 display and annotation tool. Nucleic Acids Res. 2024;52:W78–82.

701 42. Pucker B, Fiene N, Choudhary N, Borchert M, Khatun N. Collection of plant gene expression
702 data. 2024.

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

705 44. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq
706 quantification. Nat Biotechnol. 2016;34:525–7.

707 45. Pucker B, Iorizzo M. Apiaceae FNS I originated from F3H through tandem gene duplication.
708 PLOS ONE. 2023;18:e0280155.

709 46. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a
710 complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 2017;89:789–
711 804.

712 47. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis
713 Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res.
714 2012;40:D1202–10.

715 48. Pucker B, Reiher F, Schilbert HM. Automatic Identification of Players in the Flavonoid
716 Biosynthesis with Application on the Biomedicinal Plant Croton tiglium. Plants. 2020;9:1103.

717 49. Rempel A, Choudhary N, Pucker B. KIPEs3: Automatic annotation of biosynthesis
718 pathways. PLOS ONE. 2023;18:e0294342.

719 50. Seitz C, Ameres S, Forkmann G. Identification of the molecular basis for the functional
720 difference between flavonoid 3'-hydroxylase and flavonoid 3',5'-hydroxylase. FEBS Lett.
721 2007;581:3429–34.

722 51. Ramsay NA, Glover BJ. MYB-bHLH-WD40 protein complex and the evolution of cellular
723 diversity. Trends Plant Sci. 2005;10:63–70.

724 52. Hichri I, Barrieu F, Bogs J, Kappel C, Delrot S, Lauvergeat V. Recent advances in the
725 transcriptional regulation of the flavonoid biosynthetic pathway. J Exp Bot. 2011;62:2465–83.

726 53. Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L. The Arabidopsis TT2 gene encodes an
727 R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation
728 in developing seed. Plant Cell. 2001;13:2099–114.

729 54. Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C. Activation Tagging Identifies a Conserved
730 MYB Regulator of Phenylpropanoid Biosynthesis. Plant Cell. 2000;12:2383–93.

731 55. Stracke R, Werber M, Weisshaar B. The *R2R3-MYB* gene family in *Arabidopsis thaliana*.
732 Curr Opin Plant Biol. 2001;4:447–56.

733 56. Farrow SC, Facchini PJ. Functional diversity of 2-oxoglutarate/Fe(II)-dependent
734 dioxygenases in plant metabolism. Front Plant Sci. 2014;5.

735 57. Pucker B. Automatic identification and annotation of MYB gene family members in plants.
736 BMC Genomics. 2022;23:220.

737 58. Debeaujon I, Léon-Kloosterziel KM, Koornneef M. Influence of the Testa on Seed
738 Dormancy, Germination, and Longevity in Arabidopsis1. Plant Physiol. 2000;122:403–14.

739 59. Galway ME, Masucci JD, Lloyd AM, Walbot V, Davis RW, Schiefelbein JW. The *TTG* Gene Is
740 Required to Specify Epidermal Cell Fate and Cell Patterning in the *Arabidopsis* Root. Dev Biol.
741 1994;166:740–54.

742 60. Li C, Zhang B, Chen B, Ji L, Yu H. Site-specific phosphorylation of TRANSPARENT TESTA
743 GLABRA1 mediates carbon partitioning in Arabidopsis seeds. Nat Commun. 2018;9:571.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.