FamNet: A Framework to Identify Multiplied Modules Driving Pathway Expansion in Plants[1]

Colin Ruprecht, Amelie Mendrinna, Takayuki Tohge, Arun Sampathkumar, Sebastian Klie, Alisdair R. Fernie, Zoran Nikoloski, Staffan Persson[2], and Marek Mutwil[2]*

Max Planck Institute for Molecular Plant Physiology, 14476 Potsdam, Germany (C.R., T.T, S.K., A.R.F., Z.N., M.M.), School of Biosciences and Australian Research Council Centre of Excellence in Plant Cell Walls, University of Melbourne, Parkville, Victoria 3010, Australia (A.M., S.P.); and Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125 (A.S.)

ORCID IDs: 0000-0002-5993-0953 (C.R.), 0000-0002-6380-0203 (A.M.); 0000-0003-1703-0137 (A.S.); 0000-0003-2671-6763 (Z.N.); 0000-0002-6377-5132 (S.P.); 0000-0002-7848-0126 (M.M.).

Gene duplications generate new genes that can acquire similar but often diversified functions. Recent studies of gene coexpression networks have indicated that, not only genes, but also pathways can be multiplied and diversified to perform related functions in different parts of an organism. Identification of such diversified pathways, or modules, is needed to expand our knowledge of biological processes in plants and to understand how biological functions evolve. However, systematic explorations of modules remain scarce, and no user-friendly platform to identify them exists. We have established a statistical framework to identify modules and show that approximately one-third of the genes of a plant's genome participate in hundreds of multiplied modules. Using this framework as a basis, we implemented a platform that can explore and visualize multiplied modules in coexpression networks of eight plant species. To validate the usefulness of the platform, we identified and functionally characterized pollen- and root-specific cell wall modules that multiplied to confer tip growth in pollen tubes and root hairs, respectively. Furthermore, we identified multiplied modules involved in secondary metabolite synthesis and corroborated them by metabolite profiling of tobacco (*Nicotiana tabacum*) tissues. The interactive platform, referred to as FamNet, is available at http://www.gene2function.de/famnet.html.

Transcriptionally associated genes tend to be involved in related biological processes (Usadel et al., 2009). Transcriptional associations, termed coexpression, have been used extensively to infer gene functions in many model organisms (Stuart et al., 2003; Yu et al., 2003; Persson et al., 2005; Itkin et al., 2013). Several Web-based tools have been developed to allow users to exploit such relationships (Mutwil et al., 2010; Obayashi et al., 2011; Lee et al., 2015). Some of these tools offer the possibility to extend the analyses to species that only recently have emerged as tractable systems for genetic engineering, such as several plant crop species (Ficklin and Feltus, 2011; Movahedi et al., 2011; Mutwil et al., 2011; Tzfadia et al., 2012). Coexpression patterns may also be conserved across species barriers (Stuart et al., 2003; Bergmann et al., 2004). Such conserved coexpressed patterns can be used to transfer knowledge obtained from a well-investigated model species to other organisms (e.g. crop plants), as is possible via several Web tools (Mutwil et al., 2011; Ruprecht et al., 2011; Tzfadia et al., 2012; Park et al., 2013). Furthermore, conserved coexpression patterns tend to be enriched for biologically relevant relationships and can be used to improve predictions (Movahedi et al., 2011; Hansen et al., 2014).

Generally, scientists apply classification schemes to associate gene products with functions. For example, genes and proteins may be associated with a family, a metabolic pathway, subcellular localization, and a protein complex. These classification schemes make it possible to define biological hierarchies and to communicate advances within specific research fields. While classifications are instructive for gene products that are associated with known biological functions, they do not allow for inferences of genes and proteins that lack functional description. Coexpressed gene neighborhoods, as functional biological units, can associate uncharacterized genes with biological functions (Aoki et al., 2007; Langfelder and Horvath, 2008; Heyndrickx and Vandepoele, 2012; Kanehisa et al., 2016). Prominent examples where this approach has been used

[1] This work was supported by the Max-Planck Gesellschaft (to C.R., A.M., T.T., A.R.F., S.K., Z.N., and M.M.), the University of Melbourne (R@MAP grant to S.P.), the European Commission's Directorate General for Research within the 7th Framework Program (FP7/2007-2013 grant no. 270089 [MULTIBIOPRO] to C.R., T.T., A.R.F., and S.P.), and ERA-CAPS (grant EVOREPRO to M.M).

[2] These authors contributed equally to the article.

* Address correspondence to mutwil@mpimp-golm.mpg.de.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Marek Mutwil (mutwil@mpimp-golm.mpg.de).

M.M. and S.P. conceived the project; C.R. and A.M. performed the knockout experiments; T.T. and A.R.F. performed the metabolomic analysis; S.K., Z.N., and M.M. conceived the bioinformatic analyses; M.M. performed the bioinformatic analyses; M.M., S.P., A.S., and C.R. wrote the article with help from all authors.

www.plantphysiol.org/cgi/doi/10.1104/pp.15.01281

include primary and secondary wall cellulose production (Brown et al., 2005; Persson et al., 2005; Ruprecht et al., 2011) and secondary metabolite production (Tohge et al., 2007; Yonekura-Sakakibara et al., 2008; Itkin et al., 2013) in plants as well as cholesterol biosynthesis (Langfelder and Horvath, 2008) and cell proliferation (Shi et al., 2010) in mouse and human breast carcinoma, respectively.

Recently, several reports have touched upon the notion that related coexpressed gene neighborhoods appear multiple times in an organism. For instance, the primary wall cellulose synthesis neighborhood contains several genes for which close homologs appear in the secondary wall cellulose synthesis neighborhood (Ruprecht et al., 2011). Similarly, a coexpressed gene neighborhood in Arabidopsis (*Arabidopsis thaliana*) is responsible for a specialized phenolic pathway during pollen development (Matsuno et al., 2009), and genes in this neighborhood have close homologs that form coexpressed neighborhoods that participate in phenolic pathways in other parts of the plant (Ehlting et al., 2008). This suggests that coexpressed gene neighborhoods have been duplicated, or even multiplied, and subfunctionalized or neofunctionalized during evolution. We refer to such multiplied gene neighborhoods as multiplied modules. A major obstacle to identify multiplied modules has been to label the genes in an appropriate manner. The multiplication of modules was investigated in 17 fungal genomes (He and Zhang, 2005; Conant and Wolfe, 2006; Wapinski et al., 2007), where genes across the whole genome were grouped into families as an indicator of functional relatedness. However, genes from different families might harbor the same protein domains that have analogous functions (Kummerfeld and Teichmann, 2005); consequently, using only gene families might not detect functionally related modules. Proteins can be labeled by protein domains via the Pfam database (Punta et al., 2012) and through families, for example via the PLAZA database (Proost et al., 2009). An alternative route, therefore, may be to use both families and domains to label gene products, with the aim to detect multiplied modules.

To capture plant-specific modules that might have related functions, our method combines protein domain and gene family labels. We used these labels and developed a statistical pipeline, which detected hundreds of multiplied modules. Furthermore, we established a Web tool, FamNet, that allows the user to retrieve conserved and multiplied modules across and within eight plant species. We used FamNet to identify, and functionally characterize, multiplied modules involved in secondary metabolism and in cell wall biosynthesis in tip-growing cells. Our findings suggest that multiplied modules indeed may perform related, but specialized, functions.

RESULTS AND DISCUSSION

A Statistical Pipeline to Detect Multiplied Modules

We have shown that several homologous gene pairs are present in the coexpressed gene neighborhoods of primary and secondary wall cellulose synthesis (Ruprecht et al., 2011). This discovery led to the question, Are homologous genes typically found in multiple coexpressed neighborhoods, or modules, and if so, how can we detect such modules? Attempts to identify gene modules based on coexpression networks have often been based on clustering algorithms that produce different clusters depending on the network properties and parameter settings (Mao et al., 2009). Here, we developed a statistical pipeline to systematically detect coexpressed gene neighborhoods with common PLAZA gene families and Pfam protein domain labels within and across eight plant species: Arabidopsis, rice (*Oryza sativa*), *Medicago truncatula*, poplar (*Populus tremula*), barley (*Hordeum vulgare*), soybean (*Glycine max*), tobacco (*Nicotiana tabacum*), and wheat (*Triticum* spp.; for details, see Supplemental Methods S1). Our pipeline consists of two main parts: (1) identification of conserved transcriptional associations of gene family and protein domain labels, and (2) mapping of these conserved associations onto coexpressed gene neighborhoods to find multiplied neighborhoods in genome-wide coexpression networks. These similar gene neighborhoods were then termed gene modules.

The assumptions behind the first part of the pipeline are that functionally related labels (i.e. gene families and Pfam domains) should be coexpressed and that the coexpression relationships should be conserved across species. We assigned the labels to genes, and therefore, any gene can be associated with multiple labels. While the labels used in this study are sequence based, our pipeline allows the inclusion of any type of label, such as ontology, protein structure information, and others. To identify the transcriptional association of labels, we transformed coexpressed gene neighborhoods into label coexpression networks (Fig. 1, A and B; Supplemental Methods S1). We then permuted the gene-label assignments to obtain associated labels in the eight plants (Fig. 1C; Supplemental Methods S1). As conserved coexpression relationships are better estimates for true biological relationships (Mutwil et al., 2011; Heyndrickx and Vandepoele, 2012; Hansen et al., 2014), we only retained coexpressed label associations found in at least two species to ensure the robustness of the associations (Fig. 1D; Supplemental Methods S1). We termed the conserved label association network the ELA network (Supplemental Data S2). The ELA network represents conserved associations between gene families and protein domains and can reveal functional associations between these labels. Figure 1, E to G, shows three ELA regions specific to labels involved in cell wall biosynthesis, photosynthesis, and ribosome biogenesis. The ELA region of the Cu-oxidase_2 label associated with lignin production during cell wall formation identified several other labels involved in cell wall biosynthesis, such as COBRA, DUF579, and various carbohydrate-active enzymes (CBMs, glycosyl hydrolases, and others; Fig. 1E; Ruprecht et al., 2011). The ELA region of the nascent polypeptide-associated complex (NAC) contains labels that are structural components of ribosomes, ribosome assembly, and

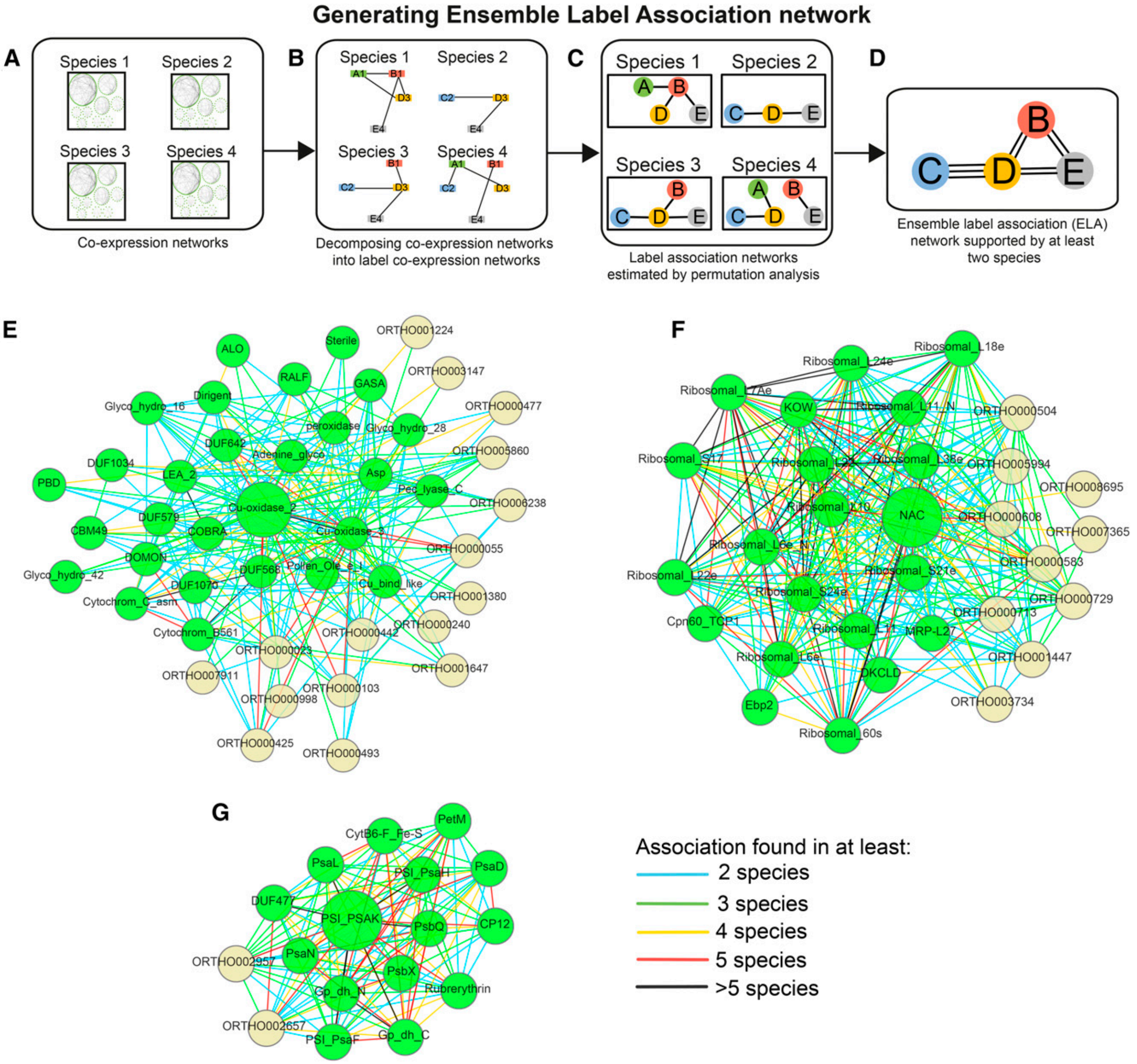


Figure 1. Generating the Ensemble Label Association (ELA) network. A, Coexpression networks derived from the PlaNet platform are used as input. Each gene may be assigned multiple labels. B, The gene coexpression networks are decomposed into label coexpression networks, where nodes represent labels assigned to genes and edges represent coexpression relationships between the labels. C, Associations between labels are detected for each species by a label-based node cover and permutation test. The result is label association networks, where nodes represent labels and edges represent associations between the labels. D, Label association networks are combined into an ELA network. The number of edges (associations) that are conserved across the different species is determined. In this example, labels C and D are connected in three species (species 2, 3, and 4). E, ELA of Cu-oxidase_2. F, ELA of PSI_PSAK. G, ELA of NAC. Green and yellow nodes represent Pfam and PLAZA labels, respectively. Edges show in how many species an association was found.

translation factors (Ebp2, MRP-L27, and Cpn60_TCP1; Fig. 1F). Another example, the PSI label PSI_PSAK, revealed other components of the photosystem, such as PSI (PSI and PSA labels) and PSII (PSB labels; Fig. 1G). Therefore, this part of the pipeline established conserved label associations across eight plant species. The ELA network is used to define valid labels when estimating similarities of modules by only using label combinations found in the ELA network, as described below.

Next, we mapped the conserved label associations (ELA) to the gene coexpression network with the aim to find modules. Importantly, we removed genes that were not supported by the ELA network, as they represented nonconserved associations (Fig. 2, A and B; Supplemental Methods S1). As genes in our pipeline can be associated with multiple labels, it is likely that neighborhood similarities are overestimated if only the number of shared labels is used for counting. For example, simple label counting would return the same result when comparing two neighborhoods if (1) each contains one gene with labels ABC or (2) each contains three genes with single labels D, E, and F. While both examples indicate three labels in common for the neighborhoods, the outcome of (1) is due to the number of labels assigned to the genes. To avoid this potential bias, we iteratively binned genes that were associated with the same labels into what we refer to as label co-occurrences (Fig. 2C; Supplemental Methods S1). Label co-occurrences were subsequently counted and used to represent neighborhood similarities (Supplemental Methods S1). To test which neighborhood pairs are significantly similar, we permuted gene-label associations 1,000 times to estimate the empirical P value for each pair (Fig. 2D; Supplemental Methods S1). Gene neighborhood pairs that were significantly similar ($P <$ 0.01) were then referred to as multiplied modules (Fig. 2E; Supplemental Methods S1). We note that since label co-occurrences greedily bin genes that have at least one protein domain or gene family in common into one unit, the metric tends to underestimate the similarity of modules. The multiplied modules are available as

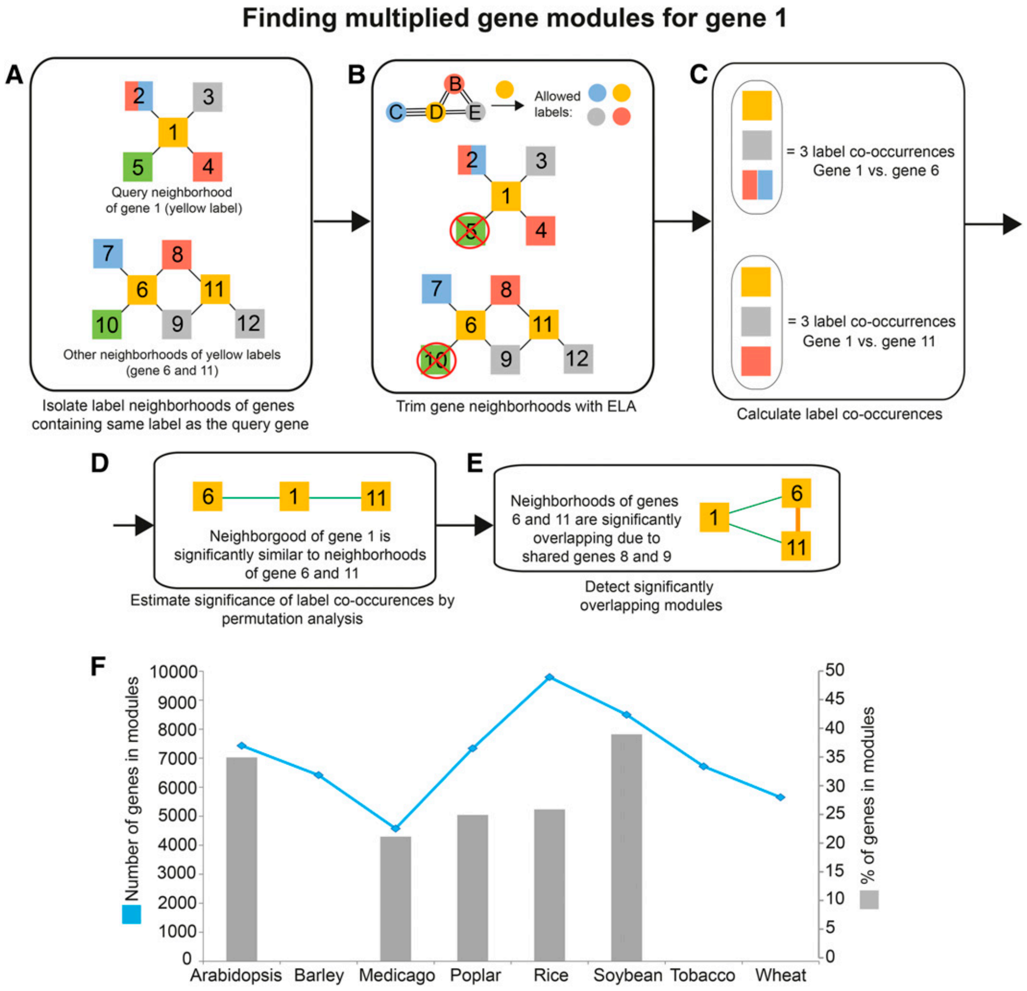

Finding multiplied gene modules for gene 1
A
2
3
1
5
4
Query neighborhood of gene 1 (yellow label)
7
8
6
11
10
9
12
Other neighborhoods of yellow labels (gene 6 and 11)
Isolate label neighborhoods of genes containing same label as the query gene
B
B
C
D
E
Allowed labels:
2
3
1
5
4
7
8
6
11
10
9
12
Trim gene neighborhoods with ELA
C
= 3 label co-occurrences
Gene 1 vs. gene 6
= 3 label co-occurrences
Gene 1 vs. gene 11
Calculate label co-occurences
D
6
1
11
Neighborgood of gene 1 is significantly similar to neighborhoods of gene 6 and 11
Estimate significance of label co-occurences by permutation analysis
E
Neighborhoods of genes 6 and 11 are significantly overlapping due to shared genes 8 and 9
1
6
11
Detect significantly overlapping modules
F
Number of genes in modules
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
% of genes in modules
50
45
40
35
30
25
20
15
10
5
0
Arabidopsis
Barley
Medicago
Poplar
Rice
Soybean
Tobacco
Wheat

Figure 2. Detecting similar modules. The pipeline is exemplified by searching for similar modules to the neighborhood of gene 1. A, The neighborhood of the query gene 1 is first isolated. Nodes represent genes, edges represent coexpression relationships, and node colors indicate labels found in collected genes. Note that gene 2 has two labels, red and blue. Label neighborhoods of genes containing orange label (genes 6 and 11) are isolated. B, The neighborhoods are trimmed with ELA, where labels not supported by ELA are removed (Fig. 1D). C, Label co-occurrences found between the neighborhood of the query gene and label neighborhoods are calculated. As gene 2 contains two labels, genes 7 and 8 are collapsed into one label co-occurrence. D, The significance of found label co-occurrences is estimated by permutation analysis. Green edges indicate similar neighborhoods. E, Overlapping modules are identified. F, Total numbers and percentages of genes assigned to similar modules. The blue line (left y axis) denotes the numbers of genes assigned to modules. Gray bars (right y axis) represent the percentages of total genes found on the microarrays that are assigned to modules. Note that the percentages of genes for barley, wheat, and tobacco are missing due to the lack of comprehensive genome annotation of the microarrays.

Supplemental Data S3. Since many of these multiplied modules are overlapping in the coexpression network, we selected only nonoverlapping modules in a last step of the pipeline (Fig. 2F; Supplemental Methods S1).

Genome-Wide Analysis of Multiplied Gene Modules

We found that between 4,000 (*M. truncatula*; blue line in Fig. 2F) and 10,000 (rice; blue line in Fig. 2F) genes were associated with multiplied modules in the eight plants. This indicates that between 22% (*M. truncatula*; gray bars in Fig. 2F) to 38% (soybean; gray bars in Fig. 2F) of the genes in the genome of these species were part of the multiplied modules. These numbers are likely to be underestimates, as not all genes are represented on microarrays; typically, around 60% of the total genes in the genome of these species have corresponding probe sets (Mutwil et al., 2011). Also, not all cell types and tissues are covered by the expression data. For example, *M. truncatula* lacks microarrays capturing the transcriptome of pollen (Mutwil et al., 2011), and consequently, pollen-specific modules will not be detected in our study. Finally, since we only considered conserved label associations, the analysis disregards multiplied gene modules that are species specific. Nevertheless, our analysis revealed that a substantial portion of the genes in the eight plant genomes participate in the multiplied modules.

Next, we investigated the module sizes (i.e. how many label co-occurrences any two multiplied modules have in common; Fig. 3). As our pipeline does not use clustering but is based on neighborhoods, it is possible that some genes of one module also are included in another module. To estimate the number of nonoverlapping modules, we used a greedy heuristic based on sorting pairs of duplicated modules according to the number of label co-occurrences in descending order and collected the values of label co-occurrences when modules do not overlap with already collected modules (Supplemental Fig. S1). While this heuristic favors the selection of large modules, we found that approximately 80% of the multiplied modules were small (i.e. similar due to two to five common labels). However, we also identified modules that contained over 15 label co-occurrences (Fig. 3A; Supplemental Data S4), exemplified by large multiplied modules involved in the defense response in soybean (Fig. 3B), chromatin remodeling in rice (Supplemental Fig. S2), and ribosome biogenesis in tobacco (Supplemental Fig. S3). This demonstrates that large functionally related modules have been multiplied.

We also investigated the number of times the modules can be multiplied, termed module degree. Since some modules are overlapping, we again used a greedy heuristic to select nonoverlapping modules by sorting each module according to the degree in descending order (Supplemental Fig. S4). While this heuristic favors modules with a high degree, we observed that approximately 80% of them have been multiplied a few (less than five) times (Fig. 4A; Supplemental Data S5). However, we also found modules that were multiplied more than 20 times, such as modules related to protein degradation in Arabidopsis (Fig. 4B), metabolism in tobacco, and transcription in poplar (Supplemental Fig. S5). Taken together, these results support frequent module multiplications, which can lead to alternative pathways.

To evaluate if particular biological processes have been preferentially multiplied, we analyzed the modules by MapMan ontology enrichment analysis (Supplemental Data S6). We found that modules of high degree were enriched for regulatory processes, including transcriptional control, RNA processing, protein degradation, and receptor kinases (Fig. 5). Furthermore, the large number of cell wall-related modules indicates that plants have evolved multiple specialized pathways to produce, remodel, and degrade cell walls (Fig. 5). Interestingly, eukaryotic protein synthesis modules are also abundant, indicating that plants might employ diverse translational machineries.

The FamNet Platform as a Web-Based Tool to Search for Modules

To provide the research community with a platform to explore the multiplied modules, we established a Web-based database, coined FamNet, that is fully integrated in the PlaNet platform (Mutwil et al., 2011). We updated gene pages in PlaNet to indicate if a gene of interest participates in multiplied modules (Fig. 6A), while new label pages show the ELA network of any label of interest and indicate multiplied gene modules in which the label is present. The FamNet database enables viewing coexpression neighborhoods, expression profiles of genes, and Gene Ontology enrichment analyses of selected modules (Fig. 6A). We exemplify the usefulness of the FamNet platform below using multiplied cell wall modules and secondary metabolism-related modules.

Functional Characterization of Cell Wall Modules within Arabidopsis

Primary and secondary cell wall cellulose biosynthesis are multiplied modules found in higher plants (Persson et al., 2005). However, navigating to the gene page of primary cell wall multicopper oxidase (At1g41830) suggested that Arabidopsis contains many (13) cell wall-related modules similar to At1g41830, with at least 10 label co-occurrences (Fig. 6B). We chose four modules, centered around At1g41830, At5g05390, At3g13390, and At4g37160, for further analysis by selecting them from the gene module table, selecting ELA support (to remove genes not supported by ELA), and clicking Compare. Output from FamNet returned an expression profile analysis of module centers (Fig. 6C), which revealed that At5g05390 is expressed in stems and coexpressed with secondary wall-related genes. We also found that the At4g37160 module contained genes preferentially expressed in roots, while the

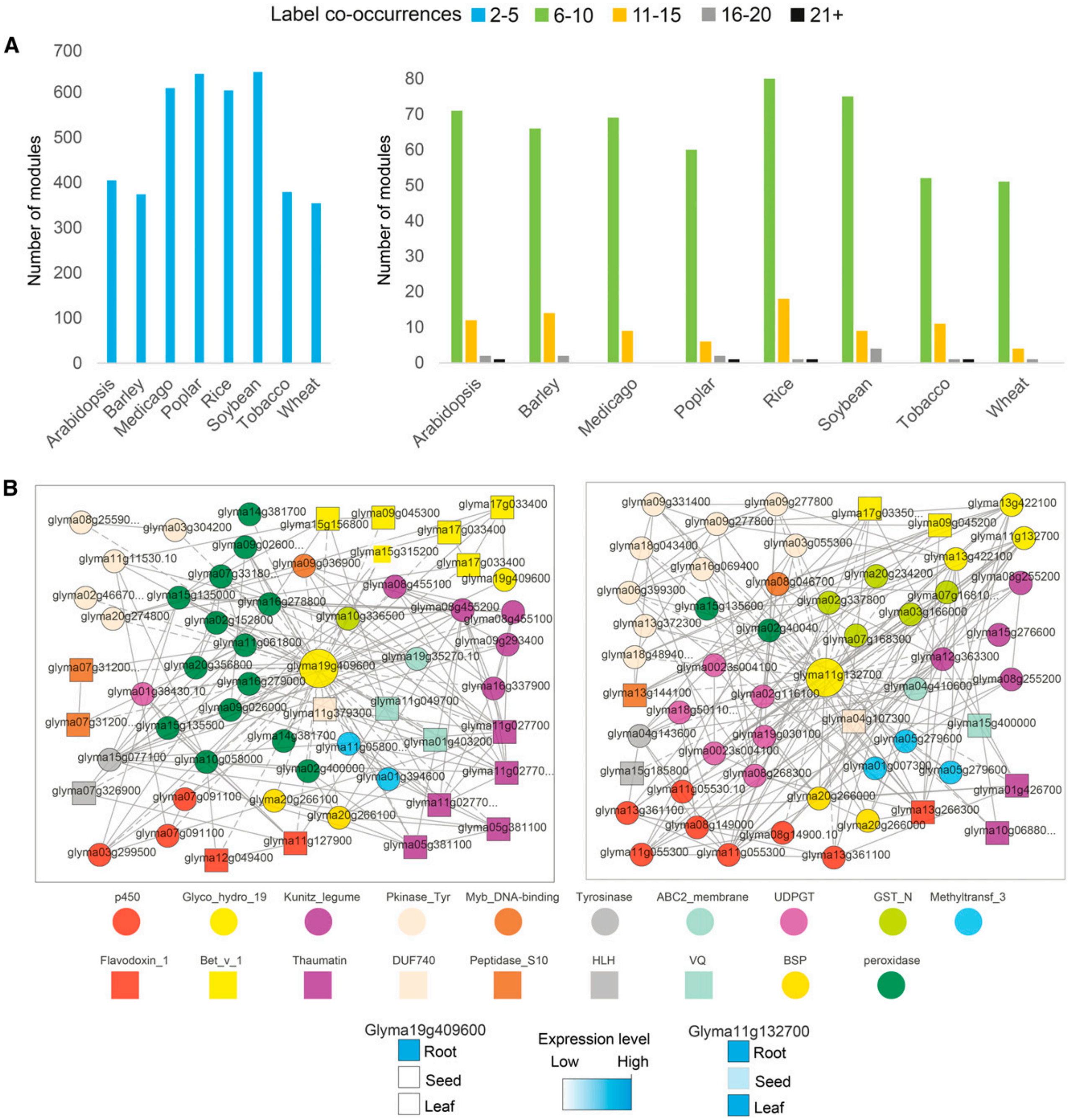


Figure 3. Distribution of label co-occurrences found between similar modules. A, Distribution of label co-occurrences between similar modules in the eight angiosperms. Blue bars (left chart) indicate modules similar due to two to five label co-occurrences. Green, orange, gray, and black bars (right chart) indicate modules similar due to a higher number of label co-occurrences. B, Example of two modules similar due to 19 label co-occurrences in soybean, with Glyma19g40960 and Glyma11g13270 used as module centers (large yellow nodes). The colored nodes represent label co-occurrences, while gray edges represent coexpression relationships. Expression profiles of the two module centers are found on the PlaNet home page. For simplicity, only Pfam labels are shown in the legend below.

At3g13390 module contained genes preferentially expressed in pollen (Fig. 6D).

To investigate the function of the pollen module further, we targeted a number of genes from this module using transfer DNA (T-DNA) insertion lines (Supplemental Data S7). Defective pollen has been reported for mutants corresponding to genes from the primary wall-related module (e.g. *CELLULOSE SYNTHASE A* genes; Persson et al., 2007), suggesting that the primary wall cellulose module is important for

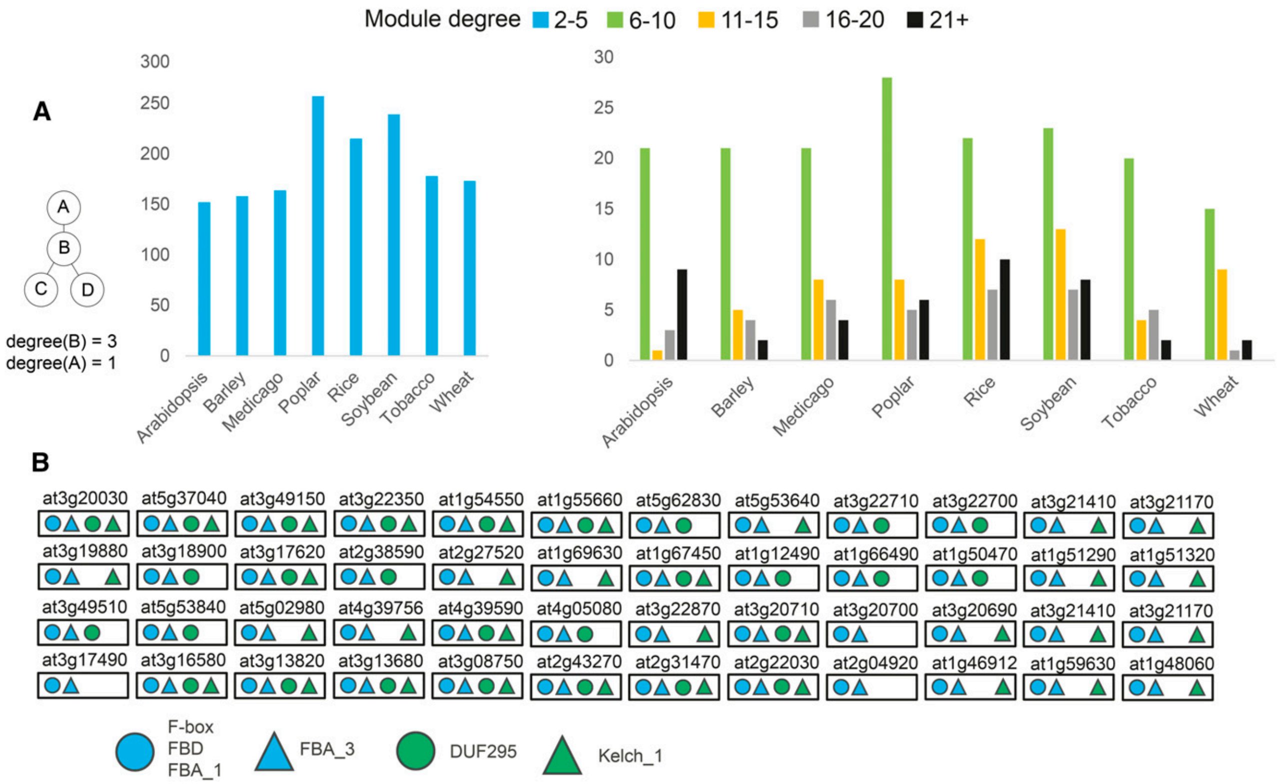
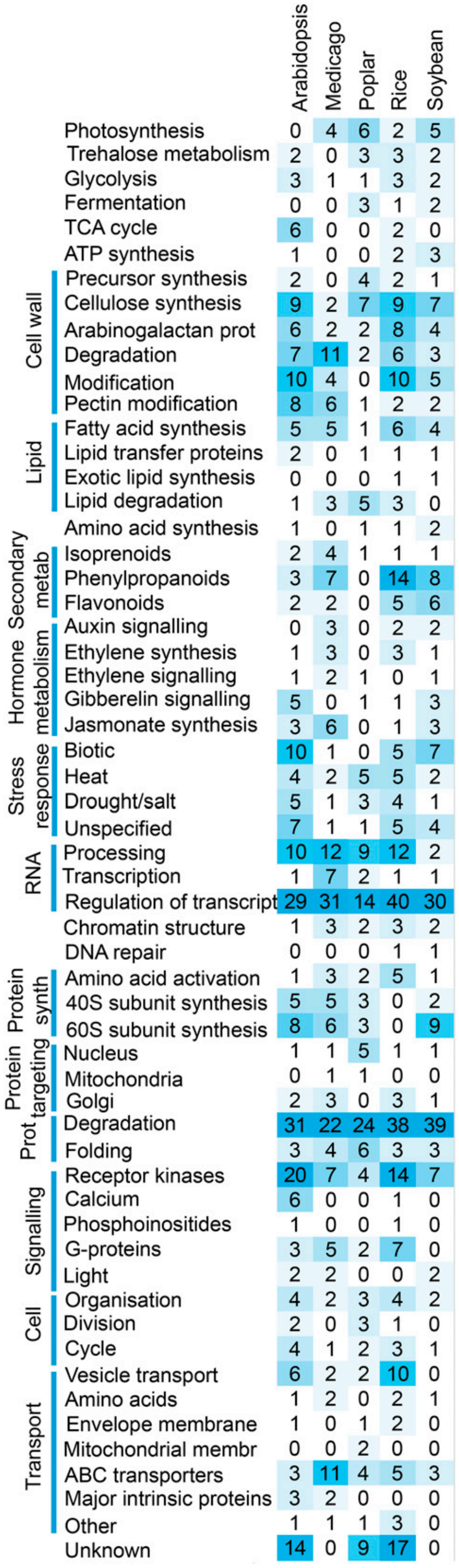


Figure 4. Distribution of module degrees. A, Module degree is defined as the number of times a representative module has been multiplied (see example at left). Blue bars (left chart) indicate modules multiplied two to five times. Green, orange, gray, and black bars (right chart) indicate modules with higher multiplication. B, Example of a highly multiplied protein degradation-related module from Arabidopsis. The Arabidopsis Genome Initiative code above each box indicates the gene used to generate the neighborhood. Colored shapes indicate the label co-occurrences shared between modules. For simplicity, gene identifiers and coexpression edges are omitted.

synthesis of the pollen wall. In contrast, none of the T-DNA mutants that corresponded to the pollen module displayed any defects in pollen morphology (Supplemental Fig. S6). However, T-DNA mutants corresponding to *COBL10*, *At4g39110*, and *At2g33420* displayed pollen tube growth-related phenotypes (Fig. 7A). COBL10 is a pollen-specific homolog of COBRA, which was recently associated with pollen tube formation (Li et al., 2013). We confirmed these results with a new T-DNA allele, *cobl10-4*, that also showed pollen tube growth defects (Fig. 7A). In contrast to the weak alleles in the previous report, *cobl10-4* showed no transmission of the T-DNA insert through pollen in reciprocal backcrosses (Fig. 7B). This phenotype could be complemented by introducing a genomic construct of *pCOBL10*::*COBL10* into *cobl10-4* (Fig. 7B), corroborating that COBL10 is essential for pollen tube growth. For the gene *At2g33420*, we found mutant lines with bulging pollen tubes (Fig. 7A). The function of At2g33420 is unknown, and based on its Pfam classification as a domain of unknown function (DUF810), we named it CELLULOSE-RELATED DUF810 (CRD1). To confirm this in vitro phenotype, we again performed reciprocal backcrosses, which revealed that two independent T-DNA mutant lines for *crd1* showed reduced transmission of the T-DNA insert through pollen (Fig. 7B). In addition, we could not obtain homozygous plants for a T-DNA mutant line corresponding to the gene *At4g39110*, and we found a segregation ratio of approximately 1:1 from a heterozygous parent plant (15 wild-type:19 heterozygous plants), suggesting gametophytic defects or lethality. This gene encodes for a receptor-like kinase that is a pollen-specific homolog of the putative cell wall integrity sensor THESEUS1 (McFarlane et al., 2014). Therefore, we named the gene *PIRITHIOUS1* (*PIR1*) according to a friend of Theseus in Greek mythology. To confirm pollen tube expression of the gene, we pollinated wild-type pistils with *pPIR*::*GUS* pollen (Supplemental Fig. S6). Furthermore, reciprocal backcrosses showed almost no transmission of the *PIR1* T-DNA insertion through the male gametophyte (Fig. 7B), which indicated pollen tube growth defects.

Our analysis of the root-specific cell wall module revealed that a T-DNA line corresponding to the RLK *PERK13* displayed bulging root hair tips (Fig. 7C), which we could complement by introducing a genomic *PERK13* construct into the mutant (Fig. 7C). These results suggested that this root-specific cell wall module is associated with root hair growth. Indeed, navigating



		Arabidopsis	Medicago	Poplar	Rice	Soybean
	Photosynthesis	0	4	6	2	5
	Trehalose metabolism	2	0	3	3	2
	Glycolysis	3	1	1	3	2
	Fermentation	0	0	3	1	2
	TCA cycle	6	0	0	2	0
	ATP synthesis	1	0	0	2	3
Cell wall	Precursor synthesis	2	0	4	2	1
	Cellulose synthesis	9	2	7	9	7
	Arabinogalactan prot	6	2	2	8	4
	Degradation	7	11	2	6	3
	Modification	10	4	0	10	5
	Pectin modification	8	6	1	2	2
Lipid	Fatty acid synthesis	5	5	1	6	4
	Lipid transfer proteins	2	0	1	1	1
	Exotic lipid synthesis	0	0	0	1	1
	Lipid degradation	1	3	5	3	0
	Amino acid synthesis	1	0	1	1	2
Secondary metab	Isoprenoids	2	4	1	1	1
	Phenylpropanoids	3	7	0	14	8
	Flavonoids	2	2	0	5	6
Hormone metabolism	Auxin signalling	0	3	0	2	2
	Ethylene synthesis	1	3	0	3	1
	Ethylene signalling	1	2	1	0	1
	Gibberelin signalling	5	0	1	1	3
	Jasmonate synthesis	3	6	0	1	3
Stress response	Biotic	10	1	0	5	7
	Heat	4	2	5	5	2
	Drought/salt	5	1	3	4	1
	Unspecified	7	1	1	5	4
RNA	Processing	10	12	9	12	2
	Transcription	1	7	2	1	1
	Regulation of transcript	29	31	14	40	30
	Chromatin structure	1	3	2	3	2
	DNA repair	0	0	0	1	1
Protein synth	Amino acid activation	1	3	2	5	1
	40S subunit synthesis	5	5	3	0	2
	60S subunit synthesis	8	6	3	0	9
Protein targeting	Nucleus	1	1	5	1	1
	Mitochondria	0	1	1	0	0
	Golgi	2	3	0	3	1
Prot	Degradation	31	22	24	38	39
	Folding	3	4	6	3	3
Signalling	Receptor kinases	20	7	4	14	7
	Calcium	6	0	0	1	0
	Phosphoinositides	1	0	0	1	0
	G-proteins	3	5	2	7	0
	Light	2	2	0	0	2
Cell	Organisation	4	2	3	4	2
	Division	2	0	3	1	0
	Cycle	4	1	2	3	1
Transport	Vesicle transport	6	2	2	10	0
	Amino acids	1	2	0	2	1
	Envelope membrane	1	0	1	2	0
	Mitochondrial membr	0	0	2	0	0
	ABC transporters	3	11	4	5	3
	Major intrinsic proteins	3	2	0	0	0
	Other	1	1	1	3	0
	Unknown	14	0	9	17	0

Figure 5. Gene Ontology analysis of multiplied modules for the five plants with comprehensive genome sequences. The values correspond to the number of times a given ontology term was enriched in the multiplied modules.

to the PlaNet gene page dedicated to PERK13 revealed enrichment for genes with annotated functions in cell wall development. To conclude, the identified pollen and root modules represent specialized cell wall synthesis machineries for pollen tube and root hair formation, respectively. We hypothesize that these two cellulose-related modules duplicated and subspecialized to confer tip growth in these cell types. These data indicate that our approach finds true biological modules that have duplicated and attained specialized functions.

Combining Metabolomics and Gene Modules: Secondary Metabolism

Coexpression has been a rewarding approach to increase our understanding of the structural pathways, and the possible regulatory machinery, governing the complexity of secondary metabolism (Tohge et al., 2007; Alejandro et al., 2012). For example, this approach has been used to find enzymes involved in distinct pathways, including steroidal glycoalkaloids (Itkin et al., 2013), flavonoid biosynthesis (Tohge et al., 2007; Yonekura-Sakakibara et al., 2007, 2008; Tohge and Fernie, 2010), as well as regulators of glucosinolate metabolism (Hirai et al., 2007) and a monolignol transporter (Alejandro et al., 2012).

Since we introduced the tobacco coexpression network in the PlaNet platform, we were especially interested to try to find gene modules related to secondary metabolism in this species. To this end, we queried FamNet using several labels that might be associated with secondary metabolism. These included chalcone synthase, chalcone isomerase, methyltransferase_2, and ABC transporter. While all of these labels generated many gene modules, here we exemplify FamNet label pages by using the methyltransferase_2 label, which contains 334 genes involved in the methylation of a range of metabolites (Fig. 8A). From the resulting ELA network, it is evident that many labels that are closely related to secondary metabolism are also present in the methyltransferase_2 network (e.g. P450, transferase, peroxidase, and methyltransferase_3; Fig. 8B). To investigate the modules that underpin the ELA network in tobacco, we went to the section Network showing similar modules containing the label. This network shows that the ELA network is supported by many modules in all eight plant species. In tobacco, there are nine modules for the methyltransferase_2 ELA that are similar to one another, with at least five label co-occurrences (Fig. 8C). While most of these modules are similar to each other (indicated by green edges), there are also several modules for which genes show overlapping expression patterns (yellow solid edges; Fig. 8C). To find out which genes make up these modules, we selected the module for which genes did not show any overlapping expression with other modules (i.e. module 203) and one representative module of the modules that did show overlapping expression

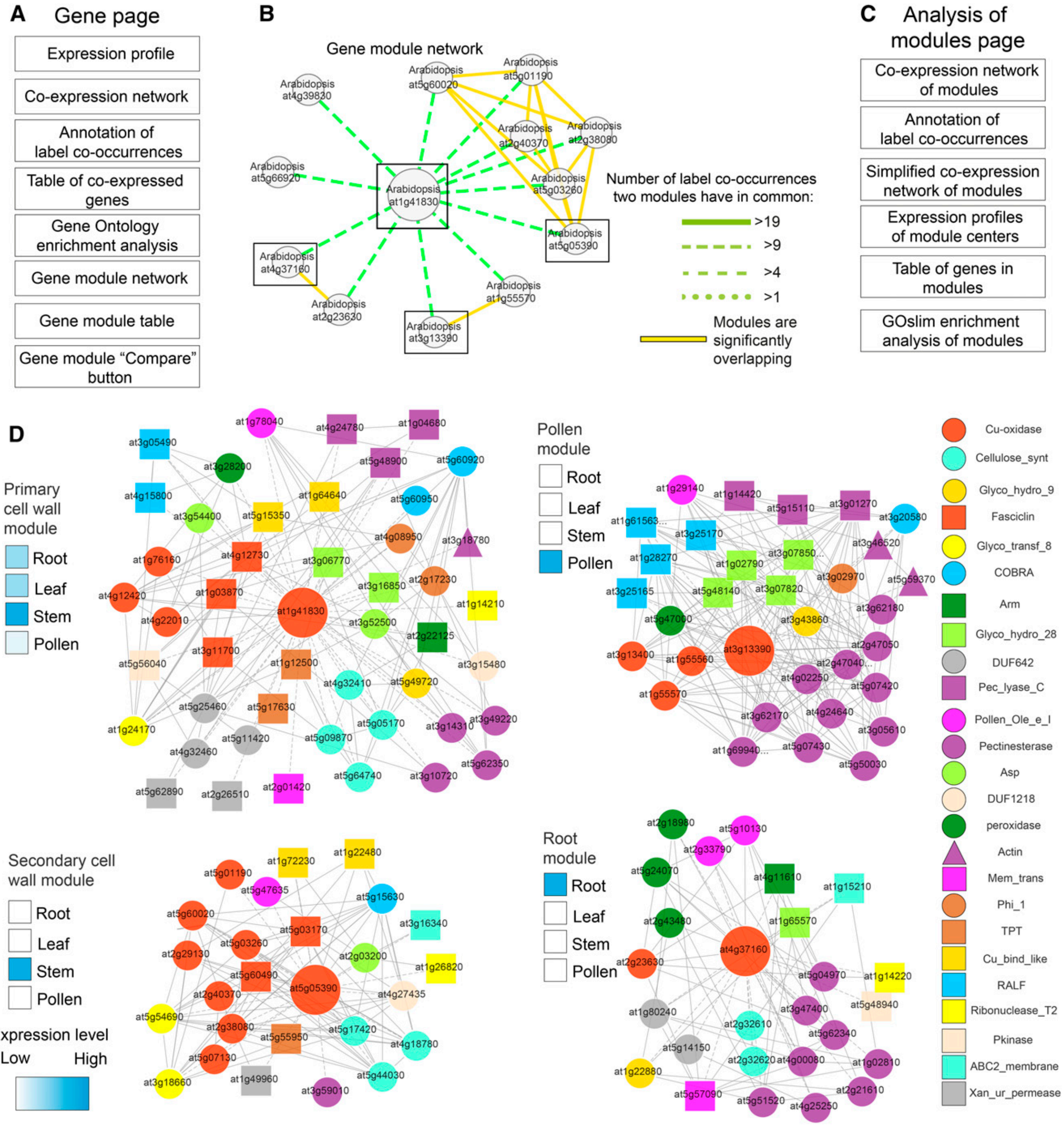


Figure 6. Cell wall biosynthetic modules occur multiple times in plants. A, Contents of new gene pages in PlaNet. B, Arabidopsis gene modules similar to the primary cell wall module centered around At1g41830 (large node). Green edges indicate similarity strength between modules as the number of shared label co-occurrences. The figure was generated by right clicking on the network and selecting Toggle similarity within one species and setting the label co-occurrence cutoff to 10. Boxes indicate modules that are displayed in detail below. C, Contents of the analysis of modules page. D, Coexpression networks of selected cell wall-related modules. Nodes and edges represent genes and coexpression relationships between genes, respectively. Colored shapes of the nodes depict label co-occurrences found in the four networks, as seen in the legend at right. Large nodes represent genes serving as module centers. Expression profiles of module center genes were estimated from expression profiles generated by FamNet and are depicted by heat maps to the left of each module.

patterns (i.e. modules 177, 6, and 175; Fig. 8C). We selected these from the Table containing the modules link, selected ELA support, and clicked Compare. FamNet indicated that the genes at the centers of these four modules have different gene expression profiles, with eb427179 expressed mainly in leaf and flower tissues,

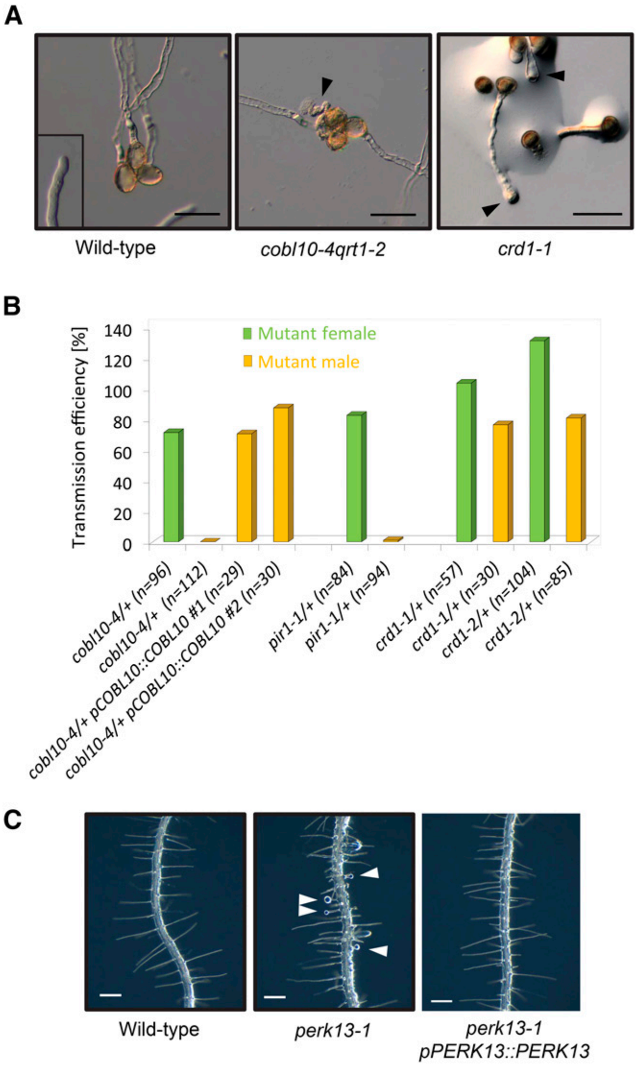


Figure 7. Pollen tube and root hair phenotypes of mutants from the pollen and root modules. A, Pollen module mutants (*cobl10-4* and *crd1-1*) show disrupted and bulging pollen tubes (arrowheads). Bars = 50 μm. B, Reciprocal backcrosses of pollen module mutants show that transmission of the T-DNA insertion through the male is completely abolished in *cobl10-4*, strongly reduced in *pir1-1*, and slightly reduced in both mutant alleles for *crd1*. Note that the phenotype of *cobl10-4* was complemented by introducing a genomic construct of *COBL10*. Transmission efficiency was calculated as heterozygous plants/wild-type plants × 100. C, Root module mutant (*perk13-1*) with bulging root hairs (arrowheads). Complementation of the *perk13-1* mutant used a genomic construct of *PERK13*. Bars = 200 μm.

c3748 in root and stem tissues, c4525 in roots, and c9634 expressed ubiquitously (Fig. 8D). Thus, the label methyltransferase_2 is present in nine tobacco modules that contain center genes with four different expression profiles.

In an attempt to associate the modules with metabolite contents, we first performed liquid chromatography-mass spectrometry (LC-MS) on plant extracts from 11 tissues, namely mature root, young leaf, mature leaf, senescent leaf, lower stem, upper stem, young silique, closed bud, open bud, flower, and mature seed of tobacco, as described (Tohge and Fernie, 2010; Supplemental Data S8 and S9). In total, 105 peaks were detected by LC-MS analysis; 14 of these could be associated with three different compound classes, hydroxycinnamates (chlorogenates), flavonoids (quercetin and kaempferol glycosides), and diterpenes (nicotianosides), that we annotated in tobacco tissues (Supplemental Fig. S7; Supplemental Data S8 and S9). Figure 9 and Supplemental Figure S7 show heat maps and the relative relationship for the different compounds and tissues. These data revealed that many compounds were preferentially accumulated in certain tissues. Most of the identified compounds were present at relatively high levels in leaves and in buds and flowers but at lower levels in mature roots, mature seeds, and stem tissues. The amounts appeared to increase with the maturity of the tissues. While this pattern generally holds true for the peaks detected by LC-MS, it is interesting that a number of compounds also are present exclusively at high levels in mature roots and seeds (Supplemental Fig. S7). Similar observations have been made previously for Arabidopsis (Lepiniec et al., 2006).

Then, to link the modules with metabolite profiles, we focused on a tissue in which a distinct profile of metabolites was evident. As we found many flavonols associated with floral tissues (Fig. 9; Supplemental Fig. S7), we decided that this could be an interesting and revealing example. Only genes from the eb427179 cluster of overlapping modules show strong expression in tobacco flowers (overlapping modules 112, 177, and 74). These overlapping modules include genes assigned to labels such as P450, transferase, and 2OG-FeII_Oxy (Fig. 8D). To get a closer estimate of the actual function of these modules, we manually investigated gene contents of the largest module 177, with gene eb427179 as its center (Fig. 10). We navigated to the eb427179 gene page by clicking on the link in the first table found on the methyltransferase_2 label page. While Gene Ontology enrichment analysis suggests that the module is involved in terpenoid metabolic processes (Supplemental Data S10), manual inspection of genes in this module revealed that 41 out of 150 genes are associated with flavonoid biosynthesis (Supplemental Data S11). In contrast, only four genes could be assigned to terpenoid biosynthesis. Moreover, many of these genes encode proteins that could facilitate a direct pathway for the synthesis of the flavonoids observed in the floral tissues of tobacco (Fig. 10). For example, we found all the genes corresponding to proteins that may convert 4-coumaroyl-CoA to a quercetin glycoside. These data are clearly in line with our metabolic estimates and support the notion that the detection of modules, together with metabolic profiling, may provide a means to discover genes associated with certain metabolic processes. We hypothesize that the discrepancy between functions predicted by Gene Ontology enrichment and those derived by manual inspection of the module contents are due to incomplete/erroneous Gene Ontology annotations. Our results are further

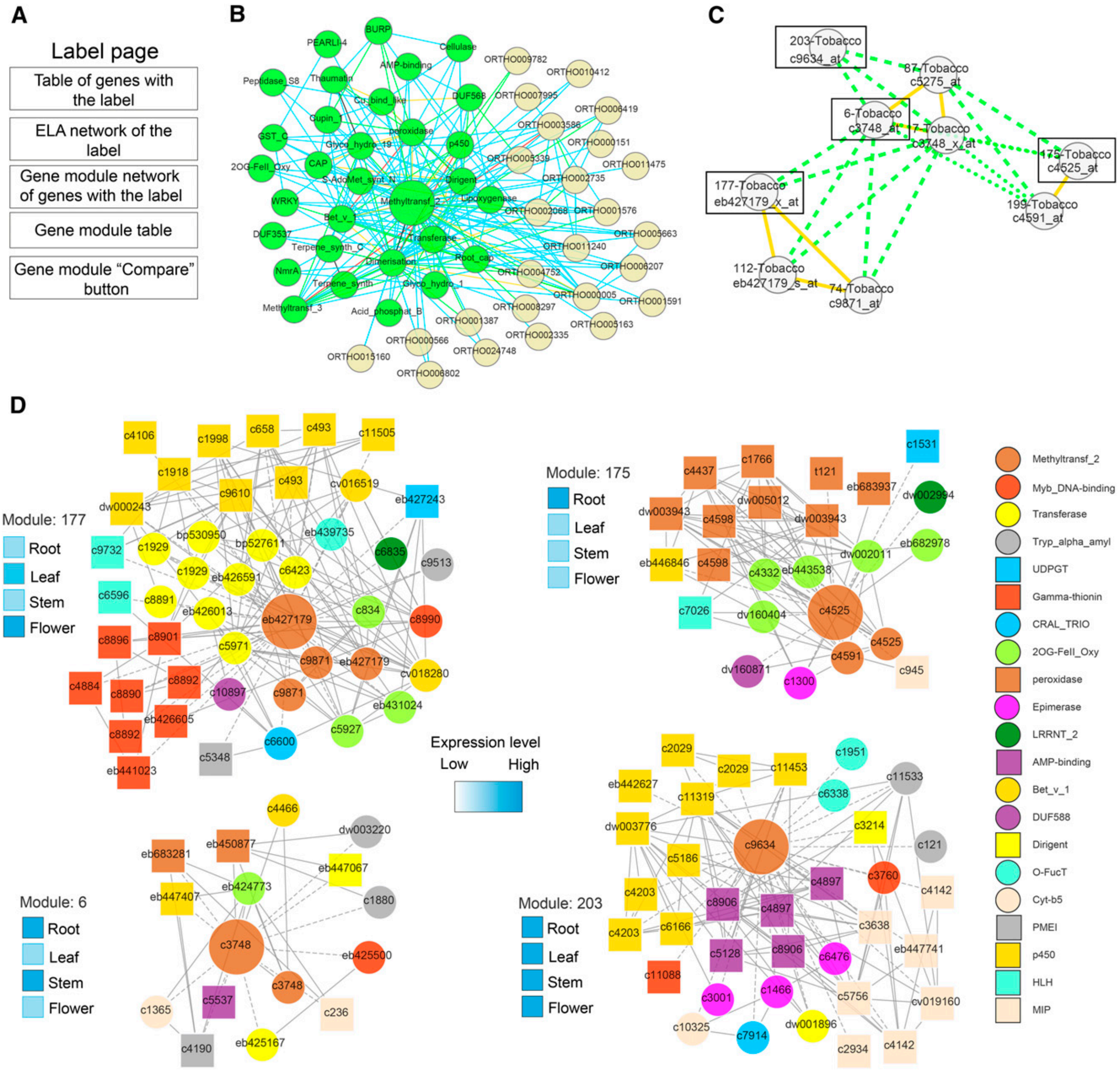


Figure 8. Secondary metabolism-related modules in tobacco based on analysis of the methyltransferase_2 label. A, Contents of label pages. B, ELA network of methyltransferase_2. Nodes represent labels, and colored edges indicate in how many species an association was found (as in Fig. 1). C, Tobacco gene modules that contain the methyltransferase_2 label. Nodes and edges are described in Figure 6. Boxes indicate modules that are displayed in detail below. Tobacco modules were highlighted by clicking on Toggle internal similarities and toggling all other species off. D, Putative flavonol-related modules in tobacco. Nodes represent genes, and the colored shapes of the nodes indicate the label co-occurrence that the respective gene is associated with. Gray edges indicate coexpression relationships between the genes. Annotation of the label co-occurrences is at right. Expression profiles of module center genes are depicted by heat maps to the left of each module.

supported by looking at genes that are supported by the ELA network (Fig. 10B; function Toggle nodes supported by ELA by right clicking on Coexpression network). This function removed approximately 100 nodes (indicated as grayed-out, transparent nodes/edges in Fig. 10B) but retained flavonoid biosynthesis-related genes, with the exceptions of c3378 and c4146 (chalcone isomerases). Hence, we show how ELA can be used to trim coexpression networks and to highlight conserved associations. However, this procedure might also lead to the removal of relevant functions of a module, as seen with the chalcone isomerases. Based on these results, we suggest that the overlapping modules 112, 177, and 74, with genes preferentially expressed in flowers, represent a floral flavonoid pathway in tobacco.

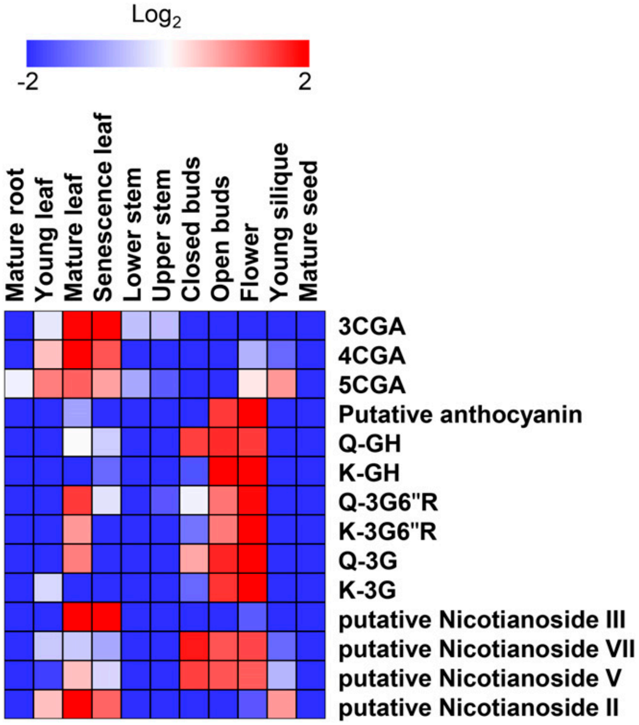


Figure 9. Heat map visualization of secondary metabolite contents analyzed by LC-MS in tobacco tissues. The analysis was conducted with three independent biological replicates. Metabolite identification and annotation were performed using standard compounds from the literature and coelution profiles with tomato pericarp extracts. The relative peak area was normalized by an average value and shown with logarithmic scale. Fold change is visualized by color: red (high) and blue (low). 3CGA, 3-Caffeoylquinate; 4CGA, cryptochlorogenate; 5CGA, neochlorogenate; Q, quercetin; G, Glc; R, Rha; H, hexose.

To see if similar modules also are present in other dicot species, we identified the modules most similar to the floral tobacco modules in Arabidopsis. We did this by navigating to the gene page of eb427179 using probe set identifier EB427179_s_at in PlaNet (Fig. 10B). Under the heading Gene module network, we selected modules from Arabidopsis that were linked to the *EB427179* tobacco module (Supplemental Fig. S8A, blue connections). These included modules centered around *At1g76790*, *At1g21100*, *At1g21130*, *At5g54160*, *At5g53810*, and *At5g37170*, of which the latter two were overlapping modules (Supplemental Fig. S8A). Interestingly, only the overlapping modules centered around *At5g53810* and *At5g37170* contained genes that clearly were expressed in Arabidopsis flowers (Supplemental Fig. S8B). Closer examination revealed that these modules contained genes that were similar to the putative floral tobacco flavonol module and contained genes annotated as MYB transcription factors, cytochrome P450, methyltransferase, and UDP glucosyltransferase (Supplemental Fig. S8B). Therefore, it appears that tobacco and Arabidopsis both contain flavonoid-related flower-expressed modules.

While our data illustrate the power of finding commonalities within and across species for the methyltransferase_2-related modules, it is important to note that it is useful to try different centers (genes) of the modules to optimize the module content when comparing them across different species and/or within one species. This is because the coexpressed gene neighborhoods are different between homologous genes, and to obtain a complete picture of the similarities in coexpressed gene neighborhoods, it is advisable to use multiple starting points for any given process (i.e. several different genes or labels). For example, in the case of secondary metabolism, one could assess the ELA networks, and subsequent gene modules, for methyltransferases, chalcone synthases, and glycosyltransferases and then compare the output from these to capture a broader picture of the process. These analyses may then inform targeted reverse genetics approaches to test the predictions and thus act as powerful tools for both gene functional annotation and metabolic engineering.

How Are Modules Multiplied?

We investigated how multiplied modules are generated. Duplication of genetic material can be divided into large-scale duplications (LSDs; duplication of the whole genome or of chromosomal segments) and small-scale duplications (SGDs; single gene duplications; Maere et al., 2005). The majority of plant species have undergone at least one, and in many cases several, LSD event(s) in the form of genome duplications and/or triplications (Bowers et al., 2003). LSD events can lead to pathway multiplication in plants, as proposed for six putative modules in Arabidopsis (Blanc and Wolfe, 2004). However, multiple subsequent SGD events also could generate modules (Fig. 11A). To determine whether the LSD or SGD events preferentially multiply modules, we first considered that LSD-duplicated genes can belong to three different classes in terms of modules (Fig. 11B). The across two modules class represents LSD gene pairs found in two similar modules and, thus, would support an LSD-based generation. The within a module and not in module classes represent LSD pairs found either together in one module or not in similar modules, which would reject LSD-based generation (Fig. 11B; Supplemental Fig. S9). By counting the number of the three classes, we found that only 13% of LSD gene pairs were associated with the across two modules class, indicating that LSD events were not the predominant mechanism for module generation (Fig. 11C).

To further corroborate this finding, we determined the bias of the distribution of the three LSD classes by switch randomization analysis (Supplemental Fig. S10). We found that the largest difference between observed and permuted networks was associated with the within a module class, as the number of LSD gene pairs belonging to this class decreased by 51% (Fig. 11C). This indicates that LSD-generated gene pairs tend to retain the expression profiles and, thus, connectivity in the coexpression networks. Conversely, the across two modules class decreased by only 7%, indicating that LSD gene pairs are rarely used to generate modules. Interestingly, ontology analysis of the

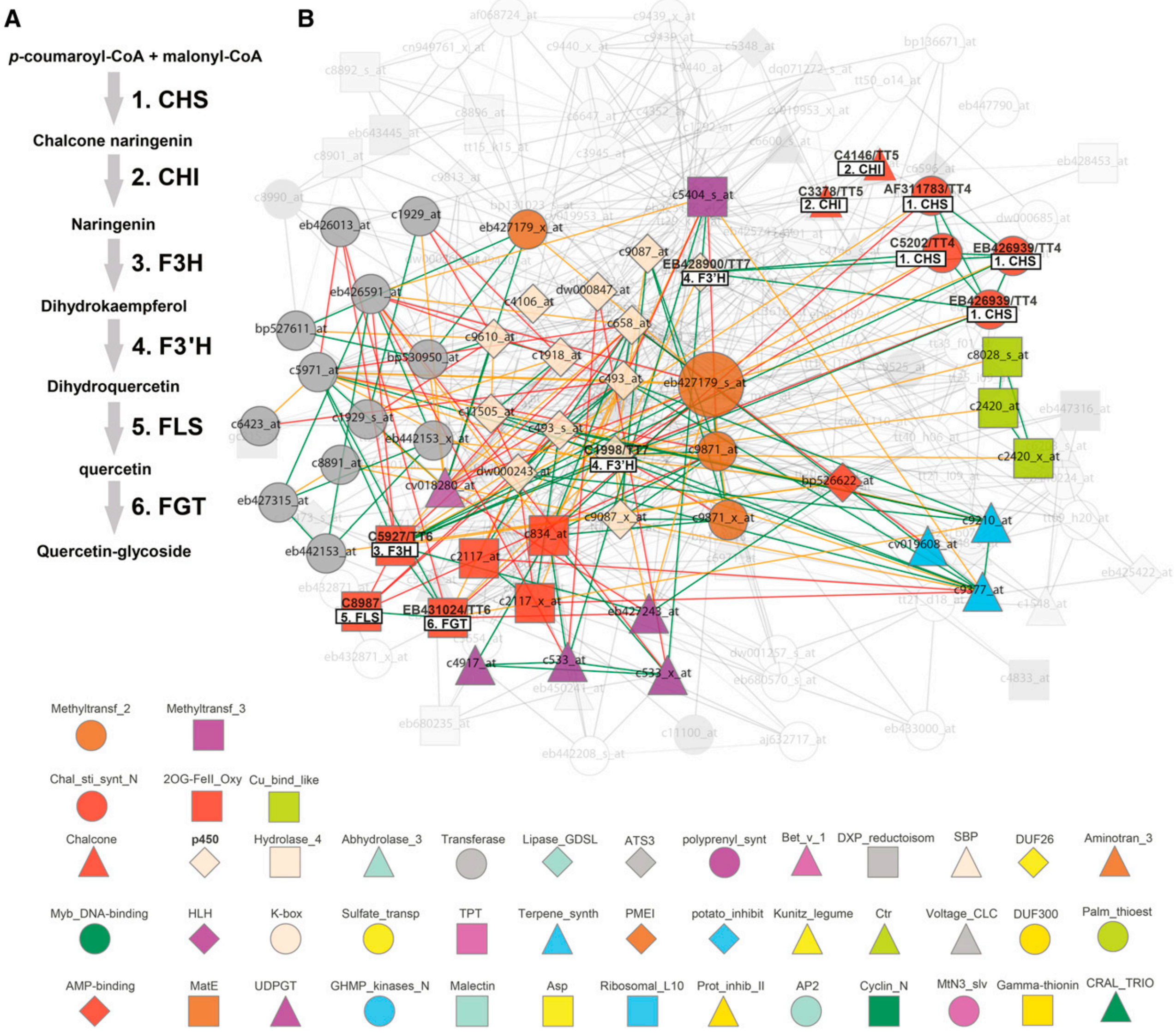


Figure 10. Scheme and coexpression network for a putative flavonol synthesis pathway in tobacco flowers. A, Outline of a potential flavonoid synthesis pathway for tobacco (based on metabolites measured in Fig. 7 and Supplemental Fig. S6). CHS, Chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; FLS, flavonol synthase; FGT, flavonol glycosyl transferase. B, Coexpression network of EB427179 (large node) corresponding to tobacco methyltransferase_2 module 177 (Fig. 8). Nodes depict genes (probe set identifiers are associated with nodes), and edges depict coexpression relationships as outlined by Mutwil et al. (2011). Colored shapes of the nodes indicate the label co-occurrence that the respective gene is associated with. Genes that correspond to enzymes in the flavonoid pathway scheme (A) are highlighted in boldface and are associated with respective boxes. The grayed-out, transparent part of the network represents nodes that are not supported by the ELA.

few modules enriched for LSD gene pairs revealed that they were preferentially dedicated to the biogenesis of eukaryotic ribosomes (Fig. 11, D and E). Taken together, the low number of LSD gene pairs in the across modules class, together with the modest decrease of the class in permuted networks, suggest that multiple SGD events are major contributors for the generation of modules in plants.

CONCLUSION

Coexpression has emerged as an important tool to rapidly infer functional relatedness among genes. These types of analyses are largely done on a gene-by-gene basis, in which a query gene for a certain biological process is used to obtain other genes that may be involved in the same process. More recently, similarities in coexpression patterns across species have become a focus (Langfelder and Horvath, 2008; Ficklin and Feltus, 2011; Mutwil et al., 2011; Heyndrickx and Vandepoele, 2012); however, instead of the gene-based approach, we here exploited the idea that sets of genes, or modules, have related functions. By analyzing such modules, we constructed FamNet, which goes beyond the gene-by-gene approach to look at transcriptional

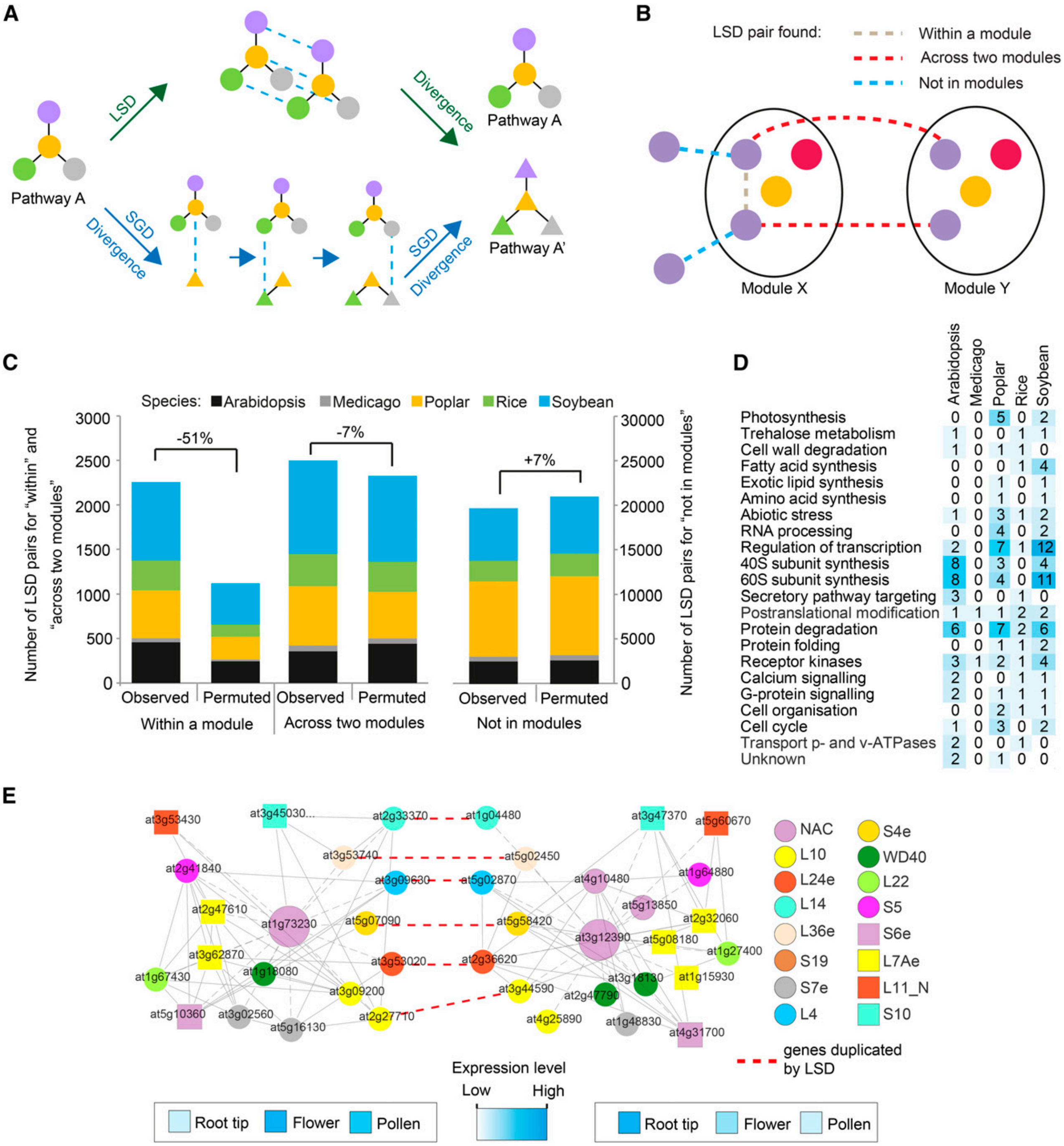


Figure 11. Gene modules are not likely generated through large-scale gene duplication events. A, Two possible models for multiplying biological pathways. Colored shapes and edges represent genes and functional relationships between genes, respectively. Blue dashed edges depict recently duplicated genes. B, Three LSD types that can occur between two similar modules. Colored shapes represent gene families. Gray, red, and blue edges depict LSD-generated gene pairs that were retained within the same module, found across the modules, and not found within both modules, respectively. C, Colors and heights of the bars represent species and numbers of LSD-generated genes. Numbers denote percentage change between the observed and the average of permuted networks. D, Ontology analysis of modules enriched significantly for LSD gene pairs. E, LSD-enriched Arabidopsis modules involved in eukaryotic ribosome biosynthesis. Colored shapes represent label co-occurrences (key shown at right). Gray and red dashed edges represent coexpression relationships and LSD gene pairs, respectively. Heat maps represent expression levels of the module centers, genes At1g73230 and At3g12390.

associations between gene labels. The inclusion of multiple species in the FamNet platform allows for better accuracy due to conserved associations between labels. The combination between the FamNet platform and the gene-based network tool PlaNet (Mutwil et al., 2011) will provide plant biologists with a versatile toolbox to explore conserved coexpressed relationships, which might facilitate rapid knowledge transfer within and across species.

MATERIALS AND METHODS

Generation of Coexpression Networks for Tobacco

The 144 microarrays comprising different tissues and environmental perturbation of tobacco (*Nicotiana tabacum*) were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). The microarrays were RMA normalized with Affymetrix Power Tools (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx) with command line apt-probeset-summarize.exe -a rma -d ATCTOBa520488.cdf -o tobaccoRMA-cel-files cel_files.txt. The normalized expression values were used to generate a Highest Reciprocal Rank network with HRRnetworkCreator.py script downloaded from http://gene2function.de/download.html (Mutwil et al., 2011). The coexpression networks are available at http://gene2function.de/download.html.

Assignment of Pfam and PLAZA Labels to Genes and Probe Sets

Fasta sequences of Pfam-A release 27 were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam27.0m (Finn et al., 2006). Protein-coding sequences of genes for Arabidopsis (*Arabidopsis thaliana*), *Medicago truncatula*, poplar (*Populus tremula*), rice (*Oryza sativa*), and soybean (*Glycine max*) were blasted against the Pfam-A database with e-value cutoff of 10^{-5}. For barley (*Hordeum vulgare*), wheat (*Triticum* spp.), and tobacco, translated representative sequences, as provided by Affymetrix, were used to BLAST against the Pfam-A database (http://www.affymetrix.com/catalog/131517/AFFY/Wheat-Genome-Array#1_3). HOM and ORTH gene labels for Arabidopsis, *M. truncatula*, poplar, rice, and soybean were downloaded from PLAZA (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v2_5/download/index; Proost et al., 2009).

Coexpression Networks

The networks, except for tobacco, were downloaded from PlaNet (http://gene2function.de/download.html; Mutwil et al., 2011). The networks, together with MapMan ontologies and Pfam and PLAZA labels, can be downloaded from PlaNet (www.gene2function.de/download.html). A table summarizing properties such as density and Pfam and PLAZA annotations is available as Supplemental Data S1.

Identification of Gene Modules

The pipeline is explained in detail in Supplemental Methods S1, which consists of two main sections: (1) how the ELA network is generated (section 1) and (2) how ELA is used to detect multiplied gene modules (section 2). To generate the ELA network, the gene coexpression networks are first transformed into label coexpression networks (section 1.1). Then, label associations are calculated from the label coexpression network of each species (section 1.2). The information from the different species are finally combined to generate the ELA network (section 1.3). To detect multiplied modules, first the nonconserved coexpression relationships between genes are removed using the ELA (section 2.1). Then, the similarities of neighborhoods are estimated based on counting of label co-occurrences, and the significance of neighborhood similarity is calculated by permutation analysis (section 2.2). Finally, similar neighborhoods are summarized to arrive at gene modules (section 2.3), and overlapping modules are detected (section 2.4).

Estimating the Distribution of Label Co-occurrences between Gene Modules

To obtain a global collection of label co-occurrences from nonoverlapping gene modules, we used a greedy heuristic, which sorted similar duplicated module pairs according to the number of label co-occurrences, in decreasing order (Supplemental Fig. S1). If at least one of the modules contains genes that have not been collected before, the value of the shared label co-occurrence in the module pair is collected. The heuristic (1) sorts all module pairs according to their label co-occurrence value. (2) For each module pair, each module is compared with the takenGenes set. (3) If genes from one or two of the modules are not in takenGenes, the label co-occurrence value is collected, and genes from either one or both modules are added to the takenGenes set. (4) Steps 2 and 3 are repeated for each module pair obtained in step 1. The heuristic is exemplified in Supplemental Figure S1, and the result, which was used to generate Figure 3A, is shown in Supplemental Data S1.

Estimating the Distribution of Module Degrees

We used a greedy heuristic to estimate the degree (i.e. the copy number) of gene modules (Supplemental Fig. S4). Similar to the previous section, the heuristic (1) sorts all module centers in descending order according to the number of other modules they are similar to. (2) The content of each module center is compared with the takenGenes set. (3) If genes from a module center are not in takenGenes, the module degree value is collected, and genes from the module center, together with genes from similar modules, are added to the takenGenes set. (4) Steps 2 and 3 are repeated for each module center obtained in step 1. The heuristic is exemplified in Supplemental Figure S4 and was used to generate Figure 4A. Similar to the estimation of the distribution of similarity strength between gene modules, the greedy heuristic returns a lower bound of the actual number of modules. The results used to generate Figure 4A are shown in Supplemental Data S5.

Estimating Functional Ontologies of Module Pairs

We used MapMan ontologies to investigate the functional enrichment of multiplied modules (Klie and Nikoloski, 2012). The empirical *P* value of ontological term enrichment is conducted by first estimating the number of ontologies present in each module, followed by shuffling Gene Ontology assignments 1,000 times. The empirical *P* value is given by the proportion of scores from the shuffling that are larger than the score from the original network. Finally, the analysis estimates which enriched MapMan terms are shared between two modules and assigns shared ontologies with the modules. The results from this analysis can be found in Supplemental Data S6. Figure 5 was generated by counting ontology terms of duplicated modules, where module selection is the same as used to generate Figure 3A (see above). To emphasize more complex modules, we used a label co-occurrence cutoff of 5 (i.e. two modules share at least five label co-occurrences). The number of enriched ontologies for cutoffs of 2, 5, and 10 can be found in Supplemental Data S6.

Plant Material, Growth Conditions, and Mutant Analysis

Seeds for all mutant lines were obtained from the Nottingham Arabidopsis Stock Centre (http://arabidopsis.info) and are all in the Columbia-0 background (Supplemental Table S7). Primers for genotyping are listed in Supplemental Table S5. Mutants from the pollen module were first grown on Murashige and Skoog medium containing 1% (w/w) Suc for 2 weeks and then transferred to standard soil (Einheitserde GS90; Gebrüder Patzer) and grown in a greenhouse under a 16-h-light/8-h-dark regime at 21°C (day) and 17°C (night). Pollen tube growth assays were performed as described previously (Boavida and McCormick, 2007). Observations of pollen tubes were carried out with a BX61 microscope (Olympus) equipped with differential interference contrast microscopy using a 10× objective. Imaging was carried out with a ColorviewIII digital camera (Olympus) controlled with cell^P software from Olympus. Mutants from the root module were grown on Murashige and Skoog medium containing 120 mM Suc for 10 d under long-day conditions (16 h of light/8 h of dark). Note that the phenotype of the *perk13* mutant is conditional on 120 mM Suc.

Metabolite Profiling and Data Analysis

Secondary metabolite analysis by LC-MS was performed as described (Tohge and Fernie, 2010). Obtained chromatographic data were processed using Xcalibur 2.1 software (Thermo Fisher Scientific). The obtained peak matrix was normalized using the internal standard isovitexin (CASRN; 29702-25-8). Metabolite identification and annotation were performed using standard compounds (3-caffeoylquinate, rutin, kaempferol-3-*O*-rutinoside, quercetin-3-*O*-glucoside, and kaempferol-3-*O*-glucoside), spectral data described in the literature (Niggeweg et al., 2004; Jassbi et al., 2008; Heiling et al., 2010; Bedoya et al., 2012; Onkokesung et al., 2012), and coelution profiles with tomato (*Solanum lycopersicum*) pericarp extracts (Rohrmann et al., 2011).

Estimating Types of LSD Genes in Modules

We used the Plant Genome Duplication Database (PGDD) to retrieve genes duplicated by LSDs for each of the five sequenced species (http://chibba.agtec.uga.edu/duplication/; Lee et al., 2013). The LSDs encompass genome and chromosome segment duplications and contain gene pairs that were found to be generated by LSD. We defined three types of relationships that LSD gene pairs can have: (1) both of the two LSD genes are found in two similar modules; (2) both genes are found in the same module; and (3) the LSD gene pairs cannot be assigned to type 1 and 2. It is important to note that a gene pair can be present in multiple modules and, therefore, can have multiple LSD relationships (Supplemental Fig. S9). Here, we have set the order of relationships as $1 > 2 > 3$. For example, if an LSD gene pair is determined to be both within a module and across two modules (such as genes 2 and 4; Supplemental Fig. S9), the analysis assigns the LSD pair to the within a module relationship. The rationale behind setting this order is 2-fold. First, since LSD relationships are investigated for each module pair, and the maximum number of genes in a module usually does not exceed 50, the majority of LSD gene pairs are always assigned to type 3 for each module pair comparison. Consequently, if relationship 1 or 2 is detected, it has higher precedence over relationship 3. Second, relationship 1 represents an LSD gene pair that is coexpressed to some degree (genes 2 and 4 are connected via gene 3; Supplemental Fig. S9) Hence, there is an uncertainty whether the gene pair is part of the same module (module C) or two similar modules with very close expression profiles (modules A and E). Here, we choose $1 > 2$, to select the more conservative scenario. Using this strategy, we counted the number of the three LSD types for the five sequenced plant species. The outcome of this analysis is shown in Figure 11.

Switch Randomization of LSD Types to Determine the Significance of LSD Type Distribution

In this section we aim to investigate if there is a bias in the distribution of the within, across, and not in modules edges described in the previous section. Permuting the LSD edges should indicate if LSD gene pairs are preferentially found within, across, and not in modules. To do this, we have employed switch randomization analysis of the LSD edges with two constraints: (1) LSD genes must belong to the same family, and (2) edges have to be shuffled to other members of the family (Supplemental Fig. S10). The permutation analysis was repeated 1,000 times, and the number of within, across, and not in modules relationships was noted for each permutation (Supplemental Fig. S10B). The average of the analysis was used to generate the permuted data bars in Figure 11C.

PERK13 Mutant Complementation

For the complementation of the perk13 phenotype, the PERK13 gene, including the endogenous promoter, was cloned into the Gateway-compatible pMDC99 vector using following primers: CACCGGTCACACGTTTGATGGTTG; rev: TCAGTAGCGCCGGTTATTGAAG. The construct was introduced into the Arabidopsis mutant via the floral dip method (Clough and Bent, 1998), and positive transformants were screened using the hygromycin resistance marker. The T3 generation was analyzed for complementation of the mutant phenotype.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Estimating genome-wide distribution of label co-occurrences between gene modules.

Supplemental Figure S2. An example of large gene modules involved in chromatin remodeling in rice.

Supplemental Figure S3. An example of large gene modules involved in ribosome biosynthesis in tobacco.

Supplemental Figure S4. Estimating the distribution of representative module degrees.

Supplemental Figure S5. Examples of frequently multiplied modules in plants.

Supplemental Figure S6. Mutants from the pollen cell wall module show normal pollen.

Supplemental Figure S7. Hierarchical clustering analysis of LC-MS metabolite profile of tobacco tissues.

Supplemental Figure S8. EB427179-like gene modules in Arabidopsis.

Supplemental Figure S9. Genes can be present in multiple modules and have multiple LSD relationships.

Supplemental Figure S10. Counting and estimating the significance of large-scale duplicated genes (LSD) in modules.

Supplemental Data S1. Properties of the microarray data and coexpression networks.

Supplemental Data S2. ELA network.

Supplemental Data S3. Multiplied modules.

Supplemental Data S4. Distribution of similarity strength values (in label co-occurrences) between modules.

Supplemental Data S5. Degree versus number of modules in the eight analyzed species.

Supplemental Data S6. MapMan ontology terms enriched between multiplied modules.

Supplemental Data S7. T-DNA insertion information about the selected genes from the pollen- and root-specific cell wall modules.

Supplemental Data S8. Metabolite reporting guidelines (checklist).

Supplemental Data S9. Recommendations for gas chromatography-mass spectrometry and LC-MS.

Supplemental Data S10. Gene Ontology analysis of EB427179_s_at.

Supplemental Data S11. Functional annotation of module eb427179_s_at.

Supplemental Methods S1. Description of algorithms used in the FamNet database.

Received August 14, 2015; accepted January 7, 2016; published January 11, 2016.

LITERATURE CITED

Alejandro S, Lee Y, Tohge T, Sudre D, Osorio S, Park J, Bovet L, Lee Y, Geldner N, Fernie AR, et al (2012) AtABCG29 is a monolignol transporter involved in lignin biosynthesis. Curr Biol **22:** 1207–1212

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol **48:** 381–390

Bedoya LC, Martínez F, Orzáez D, Daròs JA (2012) Visual tracking of plant virus infection and movement using a reporter MYB transcription factor that activates anthocyanin biosynthesis. Plant Physiol **158:** 1130–1138

Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol **2:** E9

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678

Boavida LC, McCormick S (2007) Temperature as a determinant factor for increased and reproducible in vitro pollen germination in Arabidopsis thaliana. Plant J **52:** 570–582

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR (2005) Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell **17:** 2281–2295

Clough SJ, Bent AF (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J **16:** 735–743

Conant GC, Wolfe KH (2006) Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol **4:** e109

Ehlting J, Sauveplane V, Olry A, Ginglinger JF, Provart NJ, Werck-Reichhart D (2008) An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in Arabidopsis thaliana. BMC Plant Biol **8:** 47

Ficklin SP, Feltus FA (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol **156:** 1244–1256

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al (2006) Pfam: clans, web tools and services. Nucleic Acids Res **34:** D247–D251

Hansen BO, Vaid N, Musialak-Lange M, Janowski M, Mutwil M (2014) Elucidating gene function and function evolution through comparison of co-expression networks of plants. Front Plant Sci **5:** 394

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169:** 1157–1164

Heiling S, Schuman MC, Schoettner M, Mukerjee P, Berger B, Schneider B, Jassbi AR, Baldwin IT (2010) Jasmonate and ppHsystemin regulate key malonylation steps in the biosynthesis of 17-hydroxygeranyllinalool diterpene glycosides, an abundant and effective direct defense against herbivores in *Nicotiana attenuata*. Plant Cell **22:** 273–292

Heyndrickx KS, Vandepoele K (2012) Systematic identification of functional plant modules through the integration of complementary data sources. Plant Physiol **159:** 884–901

Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, Araki R, Sakurai N, Suzuki H, Aoki K, et al (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci USA **104:** 6478–6483

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R, et al (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science **341:** 175–179

Jassbi AR, Gase K, Hettenhausen C, Schmidt A, Baldwin IT (2008) Silencing geranylgeranyl diphosphate synthase in *Nicotiana attenuata* dramatically impairs resistance to tobacco hornworm. Plant Physiol **146:** 974–986

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res **44:** D457–D462

Klie S, Nikoloski Z (2012) The choice between MapMan and Gene Ontology for automated gene function prediction in plant science. Front Genet **3:** 115

Kummerfeld SK, Teichmann SA (2005) Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet **21:** 25–30

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:** 559

Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, Shim JE, Shim H, Kim H, Kim C, et al (2015) 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 **43:** D996–D1002

Lee TH, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res **41:** D1152–D1158

Lepiniec L, Debeaujon I, Routaboul JM, Baudry A, Pourcel L, Nesi N, Caboche M (2006) Genetics and biochemistry of seed flavonoids. Annu Rev Plant Biol **57:** 405–430

Li S, Ge FR, Xu M, Zhao XY, Huang GQ, Zhou LZ, Wang JG, Kombrink A, McCormick S, Zhang XS, et al (2013) Arabidopsis COBRA-LIKE 10, a GPI-anchored protein, mediates directional growth of pollen tubes. Plant J **74:** 486–497

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102:** 5454–5459

Mao L, Van Hemert JL, Dash S, Dickerson JA (2009) Arabidopsis gene coexpression network and its functional modules. BMC Bioinformatics **10:** 346

Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard JE, Pollet B, Hehn A, Heintz D, Ullmann P, et al (2009) Evolution of a novel phenolic pathway for pollen development. Science **325:** 1688–1692

McFarlane HE, Döring A, Persson S (2014) The cell biology of cellulose synthesis. Annu Rev Plant Biol **65:** 69–94

Movahedi S, Van de Peer Y, Vandepoele K (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol **156:** 1316–1330

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell **23:** 895–910

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S (2010) Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol **152:** 29–43

Niggeweg R, Michael AJ, Martin C (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. Nat Biotechnol **22:** 746–754

Obayashi T, Nishida K, Kasahara K, Kinoshita K (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. Plant Cell Physiol **52:** 213–219

Onkokesung N, Gaquerel E, Kotkar H, Kaur H, Baldwin IT, Galis I (2012) MYB8 controls inducible phenolamide levels by activating three novel hydroxycinnamoyl-coenzyme A:polyamine transferases in *Nicotiana attenuata*. Plant Physiol **158:** 389–407

Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG (2013) Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. PLOS Comput Biol **9:** e1002957

Persson S, Paredez A, Carroll A, Palsdottir H, Doblin M, Poindexter P, Khitrov N, Auer M, Somerville CR (2007) Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. Proc Natl Acad Sci USA **104:** 15566–15571

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA **102:** 8633–8638

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell **21:** 3718–3731

Punta M, Coggill P, Eberhardt R, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al (2012) The Pfam protein families databases. Nucleic Acids Res **40:** D290–D301

Rohrmann J, Tohge T, Alba R, Osorio S, Caldana C, McQuinn R, Arvidsson S, van der Merwe MJ, Riaño-Pachón DM, Mueller-Roeber B, et al (2011) Combined transcription factor profiling, microarray analysis and metabolite profiling reveals the transcriptional control of metabolic shifts occurring during tomato fruit development. Plant J **68:** 999–1013

Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S (2011) Large-scale co-expression approach to dissect secondary cell wall formation across plant species. Front Plant Sci **2:** 23

Shi Z, Derow CK, Zhang B (2010) Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst Biol **4:** 74

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302:** 249–255

Tohge T, Fernie AR (2010) Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat Protoc **5:** 1210–1227

Tohge T, Yonekura-Sakakibara K, Niida R, Watanabe-Takahashi A, Saito K (2007) Phytochemical genomics in Arabidopsis thaliana: a case study for functional identification of flavonoid biosynthesis genes. Pure Appl Chem **79:** 811–823

Tzfadia O, Amar D, Bradbury LMT, Wurtzel ET, Shamir R (2012) The MORPH algorithm: ranking candidate genes for membership in *Arabidopsis* and tomato pathways. Plant Cell **24:** 4389–4406

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ **32:** 1633–1651

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature **449:** 54–61

Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, Niida R, Watanabe-Takahashi A, Inoue E, Saito K (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. Plant Cell **20:** 2160–2176

Yonekura-Sakakibara K, Tohge T, Niida R, Saito K (2007) Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. J Biol Chem **282:** 14932–14941

Yu H, Luscombe NM, Qian J, Gerstein M (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet **19:** 422–427