The Plant Cell, Vol. 29: 944–959, May 2017, www.plantcell.org © 2017 ASPB.

LARGE-SCALE BIOLOGY ARTICLE

A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants

Jennifer H. Wisecaver,[a] Alexander T. Borowsky,[a] Vered Tzin,[b] Georg Jander,[c] Daniel J. Kliebenstein,[d] and Antonis Rokas[a,1]

[a] Department of Biological Sciences, Vanderbilt University, Nashville, Tennessee 37235
[b] French Associates Institute for Agriculture and Biotechnology of Drylands, Jacob Blaustein Institute for Desert Research, Ben Gurion University, Sede-Boqer Campus 84990, Israel
[c] Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853
[d] Department of Plant Sciences, University of California-Davis, Davis, California 95616

ORCID IDs: 0000-0001-6843-5906 (J.H.W.); 0000-0002-5912-779X (V.T.); 0000-0002-9675-934X (G.J.); 0000-0001-5759-3175 (D.J.K.); 0000-0002-7248-6551 (A.R.)

Plants produce diverse specialized metabolites (SMs), but the genes responsible for their production and regulation remain largely unknown, hindering efforts to tap plant pharmacopeia. Given that genes comprising SM pathways exhibit environmentally dependent coregulation, we hypothesized that genes within a SM pathway would form tight associations (modules) with each other in coexpression networks, facilitating their identification. To evaluate this hypothesis, we used 10 global coexpression data sets, each a meta-analysis of hundreds to thousands of experiments, across eight plant species to identify hundreds of coexpressed gene modules per data set. In support of our hypothesis, 15.3 to 52.6% of modules contained two or more known SM biosynthetic genes, and module genes were enriched in SM functions. Moreover, modules recovered many experimentally validated SM pathways, including all six known to form biosynthetic gene clusters (BGCs). In contrast, bioinformatically predicted BGCs (i.e., those lacking an associated metabolite) were no more coexpressed than the null distribution for neighboring genes. These results suggest that most predicted plant BGCs are not genuine SM pathways and argue that BGCs are not a hallmark of plant specialized metabolism. We submit that global gene coexpression is a rich, largely untapped resource for discovering the genetic basis and architecture of plant natural products.

INTRODUCTION

Plants, being sessile and therefore at the mercy of their surroundings, harbor many adaptations that facilitate their interaction with and management of their environment. One such adaptation is the ability to produce a vast array of specialized metabolites (SMs), bioactive compounds that are not essential for growth and reproduction but rather have important ecological roles to combat pathogens, herbivores, and competitors; attract pollinators and seed dispersers; and resist abiotic stress including fluctuations in temperature, salinity, and water availability (Hartmann, 2007). Humans exploit the SM diversity of plants for medicines and other natural products; to this end, thousands of plant-derived SMs have been isolated and biochemically characterized (Raskin et al., 2002). Yet the genes responsible for the production and regulation of most SMs across land plants are unknown, which ultimately limits their potential utility in agricultural, pharmaceutical, and biotechnological applications (McChesney et al., 2007; Wurtzel and Kutchan, 2016).

[1] Address correspondence to antonis.rokas@vanderbilt.edu.
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.plantcell.org) is: Antonis Rokas (antonis.rokas@vanderbilt.edu).
www.plantcell.org/cgi/doi/10.1105/tpc.17.00009

Given their biomedical and agricultural relevance, it is perhaps surprising that the constituent genes and pathways involved in biosynthesis of most plant SMs are unknown (De Luca et al., 2012). There are two explanations for why this is so; first, SM pathways are highly variable in the number and functions of genes they contain (Hartmann, 2007; D'Auria and Gershenzon, 2005). Second, consistent with their involvement in the production of ecologically specialized bioactive molecules, SM genes typically exhibit narrower taxonomic distributions compared with genes involved in core metabolism, and SM genes are fast evolving both in terms of sequence divergence and rate of gene family diversification and display extensive functional divergence (Pichersky and Lewinsohn, 2011; Chae et al., 2014; Mukherjee et al., 2015). The consequence of this lack of evolutionary and functional conservation is that traditional sequence homology metrics for inferring gene function (e.g., Eisen, 1998) are weak predictors of SM pathway composition and function.

Network biology offers a promising alternative for identifying SM pathways and their constituent genes. Because SM pathways exist at the interface of organisms and their environments, the genes within an SM pathway share a common regulatory network that tightly controls the "where" (e.g., in what tissues) and "when" (e.g., in response to which ecological conditions) of SM production (Tohge and Fernie, 2012; Hartmann, 2007; Grotewold, 2005). Therefore, gene coexpression data, as a proxy for

coregulation, have been particularly effective in identifying the constituent genes that make up many SM pathways (Lau and Sattely, 2015; Rajniak et al., 2015; Yonekura-Sakakibara et al., 2008; Sawada et al., 2009; Hirai et al., 2007; Maeda et al., 2011; Naoumkina et al., 2010; Fridman and Pichersky, 2005; Itkin et al., 2016; Boachon et al., 2015; Sohrabi et al., 2015). Furthermore, given the availability of data from hundreds to thousands of individual gene expression experiments, integrative global coexpression networks have the power to predict SM pathways and genes in a high-throughput fashion (Horan et al., 2008; Mentzen and Wurtele, 2008; Mao et al., 2009). However, as measuring gene coexpression on a large scale was, until recently, a costly and labor-intensive undertaking, the hundreds (or more) of global gene expression studies in diverse conditions required for global coexpression network analyses currently exist for only a small minority of plant species (Aoki et al., 2016b; Hiss et al., 2014; Zimmermann et al., 2008; Ehlting et al., 2008; Usadel et al., 2009).

Another attribute that is characteristic of SM pathways found in bacteria and fungi is that they can physically colocate in the genome, forming biosynthetic gene clusters (BGCs) (Osbourn, 2010). As expected of SM pathways, genes within these microbial BGCs are coregulated and display strong signatures of coexpression, a pattern that holds true for functionally characterized as well as for putative BGCs in these genomes (Yu et al., 2011; Lawler et al., 2013; Gibbons et al., 2012a, 2012b; Lind et al., 2016; Andersen et al., 2013). As the proximity of genes on chromosomes is far easier to measure than their coexpression across multiple experimental conditions, bioinformatic algorithms strongly rely this "clustering" of genes to predict SM pathways in microbial genomes (Weber et al., 2015; Khaldi et al., 2010; Cimermancic et al., 2014). Thus, thousands of microbial BGCs have been predicted and hundreds validated (i.e., connected to known products), suggesting that gene proximity is informative for SM pathway identification, at least in these organisms (Hadjithomas et al., 2015). Nevertheless, the number of SM pathways in bacteria and fungi that do not (or only partially) form BGCs is unknown (Bradshaw et al., 2013; Sanchez et al., 2011; Lo et al., 2012).

In plants, most characterized SM pathways (e.g., glucosinolate biosynthesis) are not clustered, and their genes are distributed across the genome (Kliebenstein and Osbourn, 2012). More recently, however, nearly two dozen BGCs responsible for the production of SM defensive compounds have been identified and functionally characterized from 15 plant species (Nützmann et al., 2016), raising the possibility that gene proximity could also be used for predicting plant SM pathways (Medema and Osbourn, 2016). To this end, computational searches based on gene clustering similar to those developed for fungal and bacterial genomes postulate the existence of dozens to hundreds more BGCs across a wide variety of plant genomes (Boutanaev et al., 2015; Chae et al., 2014; Castillo et al., 2013; Schlapfer et al., 2017). However, the vast majority of these putative plant BGCs has not been functionally validated, and the fraction of plant SM pathways that form BGCs is unclear.

We hypothesized that plant SM pathways are coexpressed, independently of being organized into BGCs, in line with their ecological roles that typically require strong temporal and spatial coregulation (Tohge and Fernie, 2012; Hartmann, 2007; Grotewold, 2005). To test our hypothesis, we developed a gene coexpression network-based approach for plant SM pathway discovery (Figure 1). Using data from 10 meta-analyses of global coexpression that collectively contain 21,876 microarray or RNA-seq experiments across eight plant species, we identified dozens to hundreds of modules of coexpressed genes containing SM biosynthetic genes (e.g., cytochrome P450s, terpene synthases, and chalcone synthases) in each species, including many experimentally validated SM pathways and all validated BGCs in these species. In contrast, genes predicted to be in BGCs based on their physical proximity did not exhibit significantly different coexpression patterns than their nonclustered neighbors. Our results cast doubt on the general utility of approaches for SM pathway identification based on gene proximity in the absence of functional data and suggest that global gene coexpression data, when in abundance, are very powerful in the high-throughput identification of plant SM pathways.

RESULTS

Network Analysis Identifies Small, Overlapping Modules of Coexpressed Genes in Global Coexpression Networks

Given that SM pathway genes are often coregulated in response to specific environmental conditions, we hypothesized that genes from a given SM pathway would form tight associations (modules) with each other in gene coexpression networks. To identify modules of coexpressed SM genes, we accessed three microarray- and seven RNA-seq-based coexpression data sets from ATTED-II (Aoki et al., 2016b) and ALCOdb (Aoki et al., 2016a) for eight Viridiplantae species: *Arabidopsis thaliana*, field mustard (*Brassica rapa*), *Chlamydomonas reinhardtii*, soybean (*Glycine max*), rice (*Oryza sativa* Japonica group), poplar (*Populus trichocarpa*), tomato (*Solanum lycopersicum*), and maize (*Zea mays*) (Supplemental Data Set 1). Each data set consisted of a meta-analysis of hundreds to thousands of experiments measuring global patterns of gene expression in a wide variety of tissues, environmental conditions, and developmental stages. The number of experiments varied in each data set, from 172 in the *C. reinhardtii* RNA-seq data set (Aoki et al., 2016a) to 15,275 in the Arabidopsis microarray-based set (Aoki et al., 2016b). Pairwise measurements of gene coexpression were specified as mutual ranks (MRs; Obayashi and Kinoshita, 2009) (calculated as the geometric mean of the rank of the Pearson's correlation coefficient [PCC] of gene A to gene B and of the PCC rank of gene B to gene A; Figure 1B). For each data set, we constructed five MR-based networks, each using a different coexpression threshold for assigning edge weights (connections) between nodes (genes) in the network (Figure 1C). Networks were ordered based on size (i.e., number of nodes and edges), such that N1 and N5 indicated the smallest and largest networks, respectively.

To discover coexpressed gene modules in the eight model plants, we employed the graph-clustering method ClusterONE (Nepusz et al., 2012), which allowed genes to belong to multiple modules (Figure 1D). This attribute is biologically realistic; many plant metabolic pathways are nonlinear, containing multiple branch points and alternative end products (e.g., terpenoid biosynthesis pathways; Guo et al., 2016; Lodeiro et al., 2007).

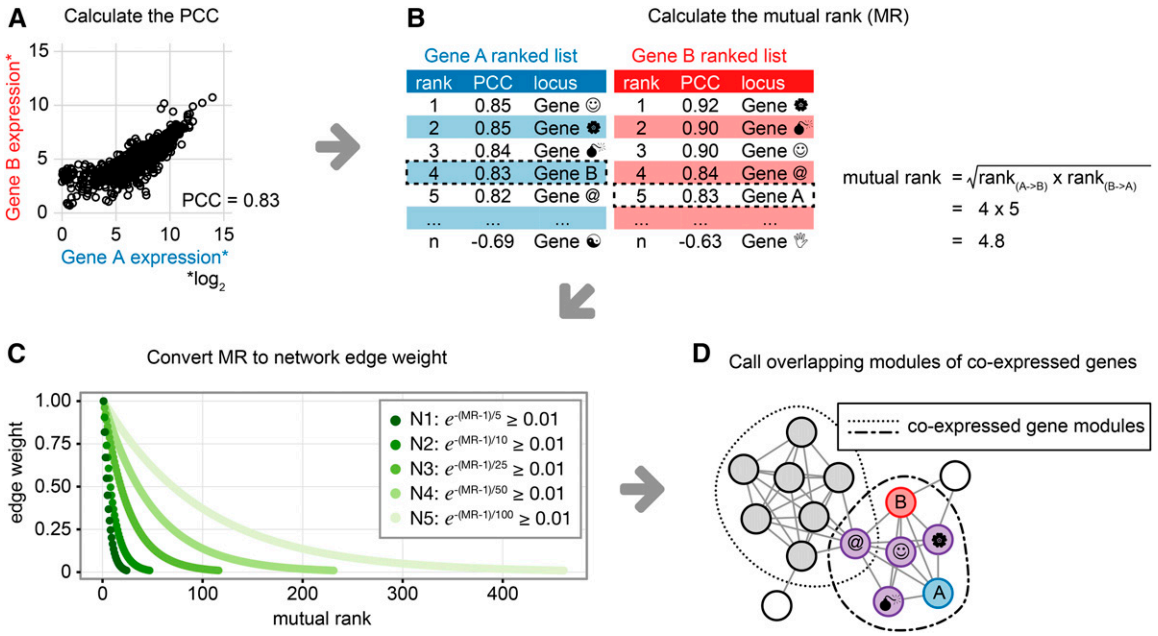

Figure 1. Coexpression Network Pipeline: A Method for Identifying Small, Overlapping Modules of Coexpressed Genes in Global Coexpression Networks.

(A) Calculate the PCC for every gene pair in the genome (e.g., the correlation between genes A and B is 0.83).
(B) Rank correlations and calculate the MR for every gene pair (e.g., the MR of genes A and B is 4.8).
(C) Convert MR to network edge weight using one or more exponential decay functions. Five different rates of decay were evaluated here, resulting in five different networks (N1–N5). Edge weights <0.01 were excluded.
(D) Call overlapping modules of tightly coexpressed genes using ClusterONE. In this example, genes A and B form a module with each other and four additional genes (purple circles). Genes can be assigned to a single module, multiple modules (e.g., Gene@), or no module (white circles).

Averaging across all 10 coexpression data sets, the number of genes assigned to modules ranged from 3251 (13.4% of protein-coding genes) in the N1 networks to 4320 (18.2%) in N5 networks (Supplemental Data Set 2). The average number of modules per network decreased with increasing network size, from 573 modules in the N1 networks to 39 in the N5 networks (Supplemental Data Set 2). Conversely, the average module size (i.e., number of genes within a module) increased with increasing network size (e.g., 7 genes per module in N1 networks, 41 genes per module in N3 networks, and 167 genes per module in N5 networks). Given our goal to recover distinct SM pathways as modules, we focused the remaining analyses on the smaller networks (N1-N3) with average module sizes (<50 genes) consistent with the number of genes typically present in SM pathways.

Coexpressed Gene Modules Recover Known SM Pathways and Predict Hundreds of New SM Gene Associations

To evaluate the correspondence between module genes and genes present in known metabolic pathways, we focused on the 798 genes in 362 Arabidopsis MetaCyc (Caspi et al., 2016) pathways with an experimentally validated metabolic function (Supplemental Data Set 3). Module genes were significantly enriched in many SM-related metabolic functions. Of the 12 higher-order metabolic classes investigated, only the "secondary metabolites" and "cell structures" biosynthesis classes were significantly enriched in module genes ($P < 0.0005$, hypergeometric tests) (Figure 2A). This pattern held true across all networks and data sets investigated (Supplemental Figure 1 and Supplemental Data Set 4). Enrichment of the cell structures biosynthesis class was driven by genes involved in the secondary cell wall (specifically lignin) biosynthesis subclasses ($P < 0.0005$, hypergeometric tests). Enriched subclasses within the secondary metabolites class included those for nitrogen-containing secondary compounds and flavonoid biosynthesis ($P < 0.005$, hypergeometric tests), which contain pathways for glucosinolate and anthocyanin production, respectively. MetaCyc SM pathways that were well recovered as coexpressed modules included those for aliphatic and indolic glucosinolate, camalexin, flavonol, flavonoid, phenylpropanoid, spermidine, and thalianol biosynthesis (Supplemental Data Set 5).

The "amino acids," "carbohydrates," and "cofactors/prosthetic groups/electron carriers biosynthesis" classes were significantly depleted in module genes in some, but not all, networks and data sets ($P < 0.05$, hypergeometric test) (Figure 2A; Supplemental Figure 1). None of the other metabolic classes displayed any significant variation between module and nonmodule genes (Supplemental Figure 1; Supplemental Data Set 4).

To estimate the number of modules that may correspond to SM pathways, we focused on those that contained two or more nonhomologous genes with a significant match to a curated list of Pfam domains that are found commonly in genes from SM pathways (Supplemental Data Set 6); as some of these "SM-like" modules share genes, we collapsed them into nonintersecting

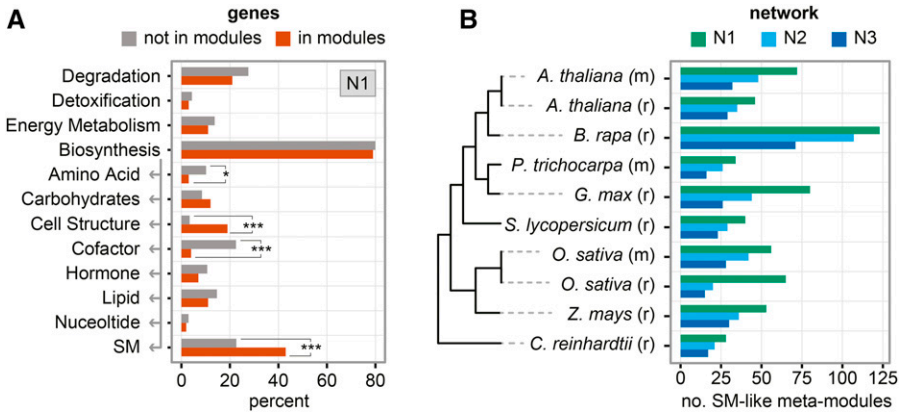

Figure 2. Global Coexpression Network Analysis of Eight Plant Genomes Identifies Coexpressed Modules of Specialized Metabolic Genes.

(A) MetaCyc pathway enrichment analysis of experimentally characterized Arabidopsis genes assigned to modules (orange bars) relative to those that do not form modules (gray bars) in Arabidopsis microarray-based network N1. Gray arrow indicates that the bottom eight pathway categories are children of "biosynthesis" in the MetaCyc hierarchy. Asterisks denote significant enrichment or depletion of MetaCyc categories in module genes: *P ≤ 0.05 and ***P ≤ 0.0005 (Benjamini and Hochberg adjusted P values, hypergeometric test). See Supplemental Figure 1 for enrichment tests in other Arabidopsis networks.
(B) Count of SM-like meta-modules identified in 10 microarray (M) and RNA-seq (R) coexpression data sets from eight Viridiplantae. SM-like modules were collapsed into meta-modules of nonoverlapping gene sets. Networks were constructed using three different rates of exponential decay for converting MR scores to edge weights, where N1 corresponds to smallest network with the steepest rate of decay and, therefore, the fewest edges; conversely, N3 is the largest network with the shallowest rate of decay and the most edges. See Supplemental Figure 2 for meta-module counts for all other networks.

"meta-modules." Dozens of SM-like meta-modules were identified in each species, with the green alga, *C. reinhardtii*, containing the fewest SM-like meta-modules (27 in N1 networks; 17 in N3 networks), and the field mustard, *B. rapa*, containing the most (120 in N1 networks, 71 in N3 networks) (Figure 2B; Supplemental Figure 2 and Supplemental Data Set 2).

Recovery of the Aliphatic Glucosinolate Biosynthesis Pathways in Arabidopsis and *Brassica* from Global Coexpression Data

To illustrate the utility and power of our approach for identifying entire SM pathways, we next focused on examining the correspondence between genes involved in the methionine-derived aliphatic glucosinolate (metGSL) biosynthesis pathway and genes that comprise coexpression modules identified by our analyses (Supplemental Data Set 7). In Arabidopsis, the species with the majority of functional data (Sønderby et al., 2010), coexpression modules recover genes for every biochemical step in this pathway, from methionine chain elongation to side-chain modification of the glucosinolate chemical backbone, as well as a pathway-specific transporter and three transcription factors (Figure 3A). For example, in the smallest network N1, 14/34 enzymatic genes in the metGSL pathway are recovered in a single 17-gene module; only 3/17 genes in this module have not been functionally characterized as involved in metGSL biosynthesis (Figure 3B). Maximum recovery of the metGSL pathway increased to 56.3 and 71.9% in the 22-gene and 43-gene modules recovered from networks N2 and N3, respectively (Supplemental Figure 3 and Supplemental Data Set 8). Although the numbers of genes not known to be involved in metGSL biosynthesis also increased in these modules, several of the genes that are coexpressed with members of the metGSL pathway perform associated biochemical processes (Figure 3A). For example, the two adenosine-5′-phosphosulfate kinase genes, *APK1* and *APK2*, are responsible for activating inorganic sulfate for use in the metGSL pathway, and polymorphisms in these genes alter glucosinolate accumulation (Mugford et al., 2009). Similarly, the cytochrome P450 genes, *CYP79B2* and *CYP79B3*, and the glutathione *S*-transferase gene, *GSTF9*, are involved in the parallel pathway for biosynthesis of glucosinolates from tryptophan instead of methionine (MetaCyc PWY-601) (Sønderby et al., 2010).

Notably, some genes implicated in metGSL biosynthesis were never recovered in coexpressed modules, including *GGP1*, which encodes a class I glutamine amidotransferase-like protein. Microarray-based coexpression data weakly associate *GGP1* with metGSL biosynthesis in Arabidopsis, and *GGP1* has been shown to increase glucosinolate production when heterologously expressed in *Nicotiana benthamiana* (Geu-Flores et al., 2009). However, our metGSL-containing modules across all RNA-seq-based networks showed that a different class I glutamine amidotransferase-like gene, *DJ1F*, is more highly coexpressed with metGSL biosynthetic genes (Figure 3). Importantly, *DJ1F* is not represented on the Arabidopsis Affymetrix GeneChip, explaining why *GGP1* and not this gene was identified as the most correlated one in earlier analyses. However, the postulated role of both *DJ1F* and *GGP1* in metGSL biosynthesis remains to be confirmed in planta.

The remaining genes in the metGSL pathway that were never recovered in coexpressed modules all encode secondary enzymes responsible for terminal modifications to the backbone glucosinolate product (Kliebenstein and Osbourn, 2012). One of these, *AOP2*, encoding a 2-oxoglutarate-dependent dioxygenase, has been pseudogenized in the Arabidopsis (ecotype Columbia) reference genome (Kliebenstein et al., 2001). The high level of natural variation present in these terminal metabolic branches is

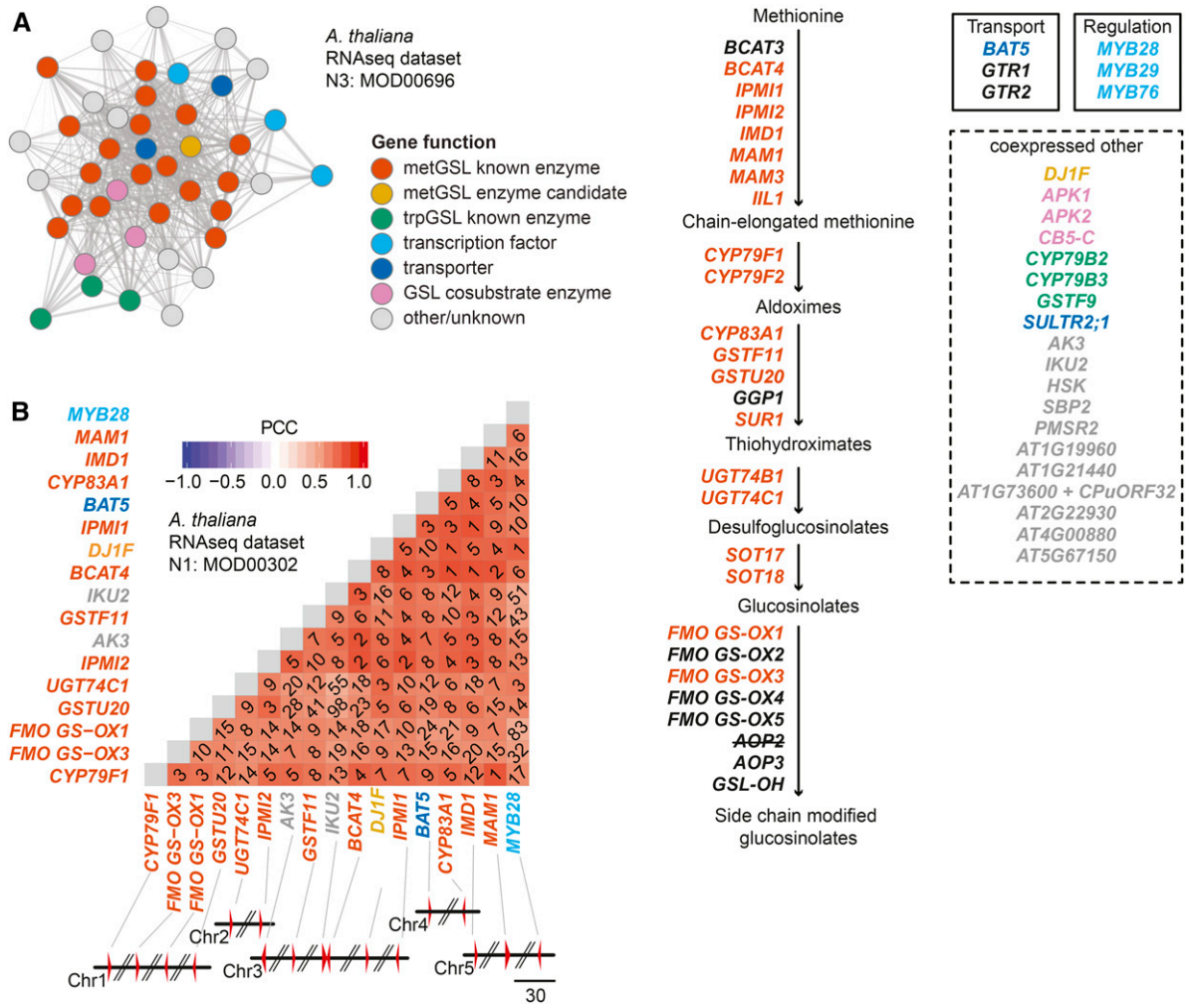

Figure 3. Coexpression Modules Efficiently Recover the Majority of Genes for metGSL Biosynthesis in Arabidopsis.

(A) Network map of an example coexpression module involved in metGSL biosynthesis. Nodes in the map represent genes, and edges connecting two genes represent the weight (transformed MR score) for the association. Network maps were drawn using a Fruchterman-Reingold force-directed layout using the igraph R package (http://igraph.org). The diagram of the metGSL biosynthesis pathway is depicted at right. Other coexpressed genes not known to be involved in metGSL biosynthesis are shown in the dashed box. Nodes and gene names are colored according to their known or putative function. MetGSL genes not recovered in modules are colored black. Genes not present in the coexpression data set are crossed out.

(B) Heat map depicting the correlation of co-expression of a second example coexpression module involved in metGSL biosynthesis. Diagonal numbers within the heat map indicate MR score. Gene names are colored as in part **(A)**. Module genes are depicted as red triangles in the accompanying chromosome segments (parallel lines indicate the genes are not physically colocated; scale bar is in kilobase pairs). Note: Data from the RNA-seq-based networks are shown because two metGSL genes (*SOT17* and *SOT18*) are not present in the microarray data set. Microarray-based networks performed similarly (Supplemental Data Set 8).

responsible for the diverse glucosinolates present in different ecotypes (Kerwin et al., 2015; Brachi et al., 2015) but likely also makes it more challenging to connect them to the rest of the metGSL pathway using global coexpression data (Supplemental Figure 4).

Brassica species also produce aliphatic glucosinolates, but a whole-genome triplication event subsequent to their divergence from Arabidopsis (Town et al., 2006) has complicated identification of functional metGSL genes in these species. To gain insight into the metGSL pathway in *B. rapa*, we cross-referenced our coexpression modules with 59 candidate metGSL genes identified based on orthology to Arabidopsis metGSL genes (Wang et al., 2011). As in Arabidopsis, modules recovered every biochemical step of the *B. rapa* metGSL pathway as well as pathway-specific transporters and transcription factors (Figure 4; Supplemental Data Sets 7 and 8). Also as in Arabidopsis, *DJ1F* rather than *GGP1* is coexpressed with other metGSL genes, providing further evidence that the DJ1F enzyme may be the more likely candidate for the γ-glutamyl peptidase activity in glucosinolate biosynthesis (Sønderby et al., 2010). Furthermore, as

several enzymes are encoded by multiple gene copies in *B. rapa*, we harnessed the power of our module analysis to identify which of these copies was coexpressed with other metGSL genes and, therefore, most likely to be functionally involved in the pathway. For example, out of the six methylthioalkylmalate synthase (*MAM*) gene copies in *B. rapa*, only *Bra029355* and *Bra013007* were recovered in metGSL modules (Figure 4; Supplemental Figure 5). Module data also suggest that the glutathione *S*-transferase class tau (GSTU) activity is one step of the core pathway that may differ between the two species. Specifically, in Arabidopsis, *GSTU20* is thought to encode the enzyme catalyzing this reaction, and this gene was recovered in metGSL modules in our analysis (Figure 3A). However, this association was not recovered in *B. rapa*. Instead, three paralogous GSTUs (*Bra003647*, *Bra026679*, and *Bra026680*), corresponding to the Arabidopsis *GSTU23* and *GSTU25* genes, respectively, formed modules with metGSL genes, making these genes good candidates for investigation of GSTU activity in *B. rapa* (Figure 4; Supplemental Figure 6).

Modules Recover Functionally Characterized BGCs and Identify Associated, Unclustered Genes

We next investigated whether our approach also recovered BGCs by examining whether our coexpression modules recovered the six functionally characterized BGCs in these eight plant genomes (Supplemental Data Set 9). All six BGCs were recovered in our module analysis (Supplemental Data Set 8). Specifically, coexpression modules recovered all genes comprising the BGCs involved in the production of the triterpenoids marneral (Field et al., 2011) (3/3 genes; Figure 5A) and thaliaol (Field and Osbourn, 2008) (4/4 genes; Figure 5B) in Arabidopsis and the diterpenoid momilactone (Shimura et al., 2007) (5/5 genes; Figure 5C) in rice. Modules recovered 7/9 genes in the phytocassane (Swaminathan et al., 2009) diterpene cluster in rice; the *OsKSL5* and *CYP71Z6* genes forming a terpene synthase-cytochrome p450 pair of genes were strongly coexpressed with each other but not with the rest of the pathway (Figure 5D). The two triterpenoid BGCs in Arabidopsis were typically combined into the same coexpression module (Figure 6A; Supplemental Figure 7); the same pattern was observed for the two diterpenoid BGCs in rice (Figure 6B; Supplemental Figure 8). Genes within these BGCs were also strongly coexpressed with additional genes located outside the BGC boundaries, including one putative transcription factor and several putative transporters (Figure 6; Supplemental Figures 7 and 8).

Seven of eight genes in the partially clustered pathway for production of the steroidal alkaloid α-tomatine in tomato (Itkin et al., 2013) were recovered by our coexpression analysis (Figure 5E). Only the glucosyltransferase gene, *GAME2*, encoding the last enzymatic reaction in the proposed α-tomatine pathway, showed a conspicuously different expression profile, consistent with previous reports (Itkin et al., 2013; Cárdenas et al., 2016). Several glucosyltransferase genes paralogous to *GAME2* were strongly coexpressed with the rest of the genes in this pathway (Figure 6C; Supplemental Figure 9), but whether or not these genes participate in α-tomatine biosynthesis is yet to be determined. Additional genes strongly coexpressed with the rest of the α-tomatine pathway include, among others, one putative transcription factor

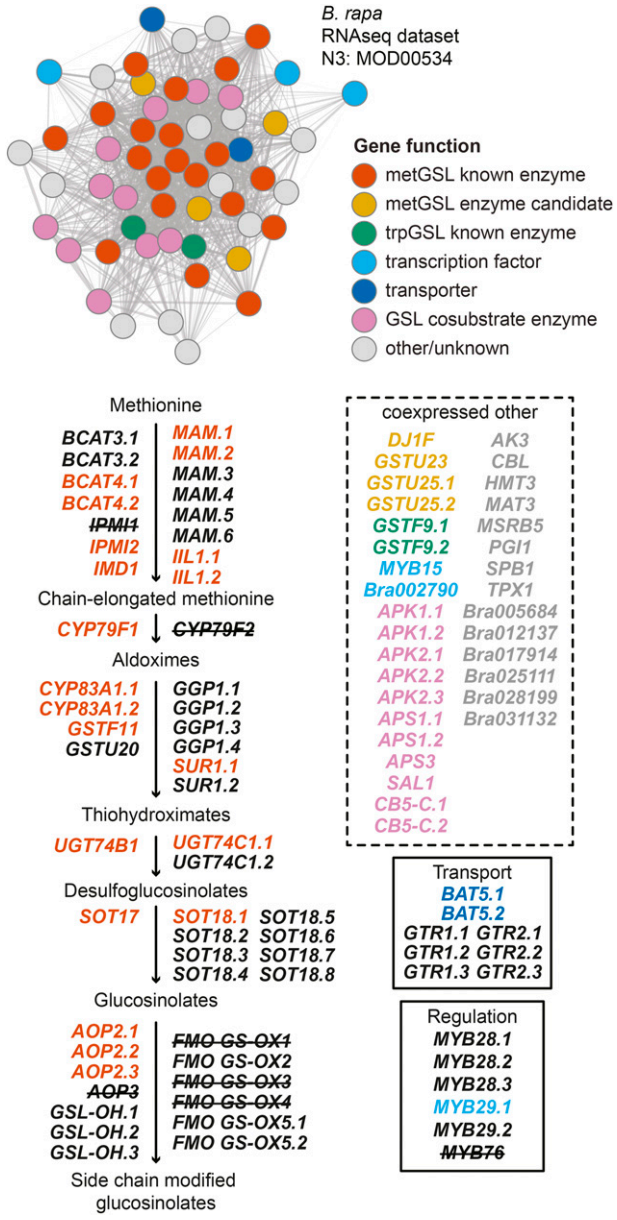

Figure 4. Coexpression Modules Predict Functional metGSL Biosynthesis Genes in *B. rapa*.

Network map of an example coexpression module involved in metGSL biosynthesis. The network map was constructed as described in Figure 3. The diagram of the metGSL biosynthesis pathway is depicted, below, with all predicted orthologs to known metGSL genes in Arabidopsis as listed on brassicadb.org. Other coexpressed genes not known to be involved in metGSL biosynthesis are shown in the dashed box. Nodes and gene names are colored according to their known or putative function. MetGSL orthologs not recovered in modules are colored black. Arabidopsis metGSL genes with no known ortholog in *B. rapa* are crossed out. See Supplemental Data Set 7 for associated NCBI and Ensembl gene IDs.

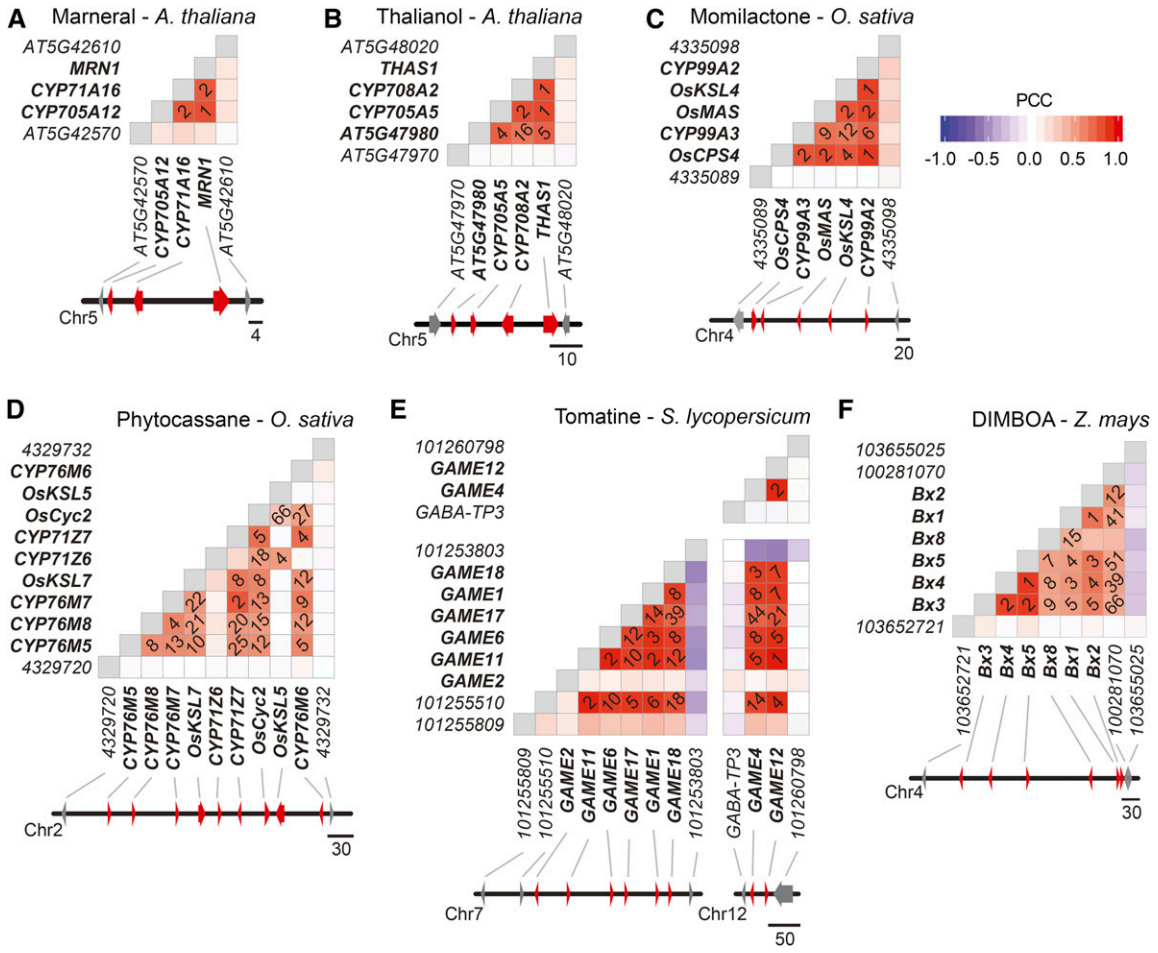

Figure 5. Coexpression Patterns of Six Functionally Characterized BGCs in Plants.

Heat maps depict the correlation of coexpression of six BGCs for the production of marneral **(A)**, thalianol **(B)**, momilactone **(C)**, phytocassane **(D)**, tomatine **(E)**, and DIMBOA **(F)**. Diagonal numbers indicate MR scores; squares are blank if MR ≥ 100. BGC genes are bolded in the heat map and colored red in the accompanying chromosome segments. Scale bars are in kilobase pairs. Genes are represented by their gene symbols, when available, or their NCBI gene IDs. Note: Gene 100281070 in the DIMBOA cluster corresponds to an uncharacterized glucosyltransferase, GRMZM2G085854.

and several possible metabolite transporters (Figure 6C; Supplemental Figure 9) as well as a cellulose synthase-like gene located adjacent to the BGC (Figure 5E).

Lastly, five of the six genes in the benzoxazinoid 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) (Frey et al., 1997) cluster in maize formed coexpression modules in our analysis (Figure 5F). Specifically, the first five genes in the DIMBOA pathway (*Bx1-Bx5*), responsible for the biosynthesis of the precursor 2,4-dihydroxy-1, 4-benzoxazin-3-one (DIBOA), formed modules with each other but not with the final gene in the BGC, *Bx8* (Figure 7).

Similar to the modifying genes of the metGSL pathway in Arabidopsis, terminal *Bx* genes appear to have unique gene expression signatures distinct from the core pathway. For example, DIBOA is modified to DIMBOA by the action of two additional unclustered genes (*Bx6* and *Bx7*) (Jonczyk et al., 2008), neither of which was assigned to modules with core genes or each other. Toxic DIBOA/DIMBOA is transformed into the stable glucoside, DIBOA-Glc/DIMBOA-Glc, by glucosyltransferases (*Bx8* and *Bx9*), which were likewise not assigned to modules in our analysis. However, a gene adjacent to the DIMBOA BGC, encoding an uncharacterized glucosyltransferase (GT; *GRMZM2G085854*) with 27% amino acid identity to *Bx8*, does belong to the same module as the core *Bx* genes in network N3 (Figure 7), but the MR scores of this gene to core *Bx* genes are noticeably weaker than those between the core *Bx* genes (Figure 5F). Additional *Bx* genes (*Bx10-Bx14*), which are not part of the BGC and are responsible for the biosynthesis of modified benzoxazinoid compounds (e.g., HDMBOA-Glc and DIM_2BOA-Glc) (Meihls et al., 2013; Handrick et al., 2016), were also not assigned to modules in our analysis (Figure 7). This pattern is similar to that observed with the terminal reactions of the metGSL biosynthesis pathway.

Bx1 is thought to represent the first committed step in benzoxazinoid biosynthesis, encoding an indole-3-glycerolphosphate lyase that converts indole-3-glycerolphosphate to indole. However, in our module analysis, an additional gene coexpressed with the core *Bx* genes is an indole-3-glycerolphosphate synthase

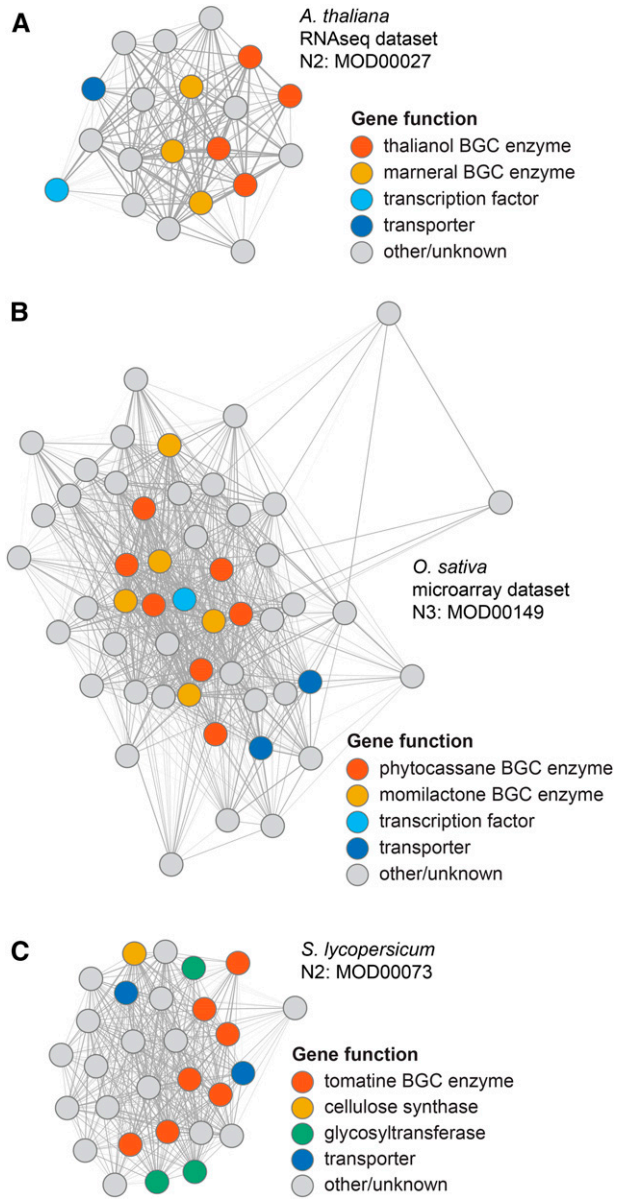

Figure 6. Network Maps of Coexpression Modules That Overlap with Known Plant BGCs.

Network maps were constructed as described in Figure 3 and depict modules involved in thalianol and marneral triterpenoid biosynthesis in Arabidopsis **(A)**, phytocassane and momilactone biosynthesis in rice **(B)**, and tomatine biosynthesis in tomato **(C)**. Gene nodes are colored according to their known or putative function. Note: The cellulose synthase gene (NCBI gene ID: 101255510) in **(C)** is located adjacent to the tomatine BGC (see Figure 5E).

gene (*IGPS*; *GRMZM2G106950*), which catalyzes the reaction directly upstream of *Bx1* (Figure 7). Two additional genes encoding indole-3-glycerolphosphate synthases are present in maize (*GRMZM2G169516* and *GRMZM2G145870*), but neither was strongly coexpressed with those in the benzoxazinoid pathway. Similarly, the two additional paralogs to *Bx1* in maize (*Tsa1; GRMZM5G841619* and *Igl1;GRMZM2G046191*, responsible for the production of tryptophan and volatile indole, respectively) formed independent coexpression modules, consistent with their distinct metabolic and ecological roles (Figure 7) (Frey et al., 2000, 2009). The inclusion of an unlinked IGPS gene in the benzoxazinoid coexpression modules suggests that the first committed step in the biosynthesis pathway may start one reaction earlier than previously predicted based on the DIMBOA BGC gene content.

To test whether *GT* and *IGPS* are likely to be involved in benzoxazinoid biosynthesis, we looked up their gene expression levels in response to two different types of insect herbivory (aphid and caterpillar), ecological conditions under which benzoxazinoid biosynthesis genes are typically induced (Tzin et al., 2017, 2015). *GT* showed gene expression responses similar to *Bx8* and *Bx9*, being induced within the first few hours after the introduction of insect herbivores (Figure 8A). Although the median fold change of expression relative to controls is small (<5) for all three glucosyltransferases, this result is consistent with a putative role of *GT* in creating stable benzoxazinoid glucosides along with *Bx8* and *Bx9*. *IGPS* was also significantly induced in response to insect herbivory, mostly notably in the caterpillar feeding experiment in which *IGPS* expression increased over 50-fold during a 24-h period (Figure 8B). In contrast, the two other indole-3-glycerolphosphate synthase genes showed little to no response to herbivory, consistent with this *IGPS* encoding a specialized enzyme involved in benzoxazinoid biosynthesis or volatile indole, which is also induced by caterpillar herbivory (Erb et al., 2015).

Bioinformatically Predicted BGCs in Plants Do Not Form Coexpression Modules and Are Typically Not Coexpressed

To examine whether putative BGCs (i.e., predicted based on physical clustering and with no known associated products) show evidence of coregulation in response to specific environmental conditions, we investigated whether they were also recovered in our coexpression network analysis. We found that two different sets of putative BGCs showed little to no coexpression (Figure 9; Supplemental Figure 10). Specifically, both the 137 Enzyme Commission (EC)-based BGCs predicted by Chae et al. (2014) and the 51 BGCs predicted by the antibiotics and secondary metabolism analysis shell (antiSMASH) (Weber et al., 2015) had median MR scores of 9670 and 10,890, respectively. Furthermore, the EC-based BGCs' distribution of coexpression was similar to that of the control distribution of neighboring genes ($P = 0.187$, Wilcoxon rank sum test), whereas the coexpression of antiSMASH BGCs was significantly lower than that of the control ($P = 0.027$) (Figure 9A; Supplemental Data Set 10). In contrast, the six validated BGCs had a median MR score of 17.4 and were significantly more coexpressed than the control ($P = 3.20 \times 10^{-4}$) (Figure 9A; Supplemental Data Set 10). Similarly, the 13 terpene synthase-cytochrome P450 (TS-CYP) pairs identified by Boutanaev et al. (2015) were variably coexpressed with a median MR score of 45. Although two of the 13 TS-CYP pairs were negatively correlated in their expression, the TS-CYP distribution was still significantly better than the control ($P = 2.73 \times 10^{-4}$) (Figure 9A; Supplemental Data Set 10).

Not surprisingly, given the lack of coexpression, putative BGCs, by and large, did not form coexpression modules, with only 7/188 putative BGCs overlapping by three genes or more with

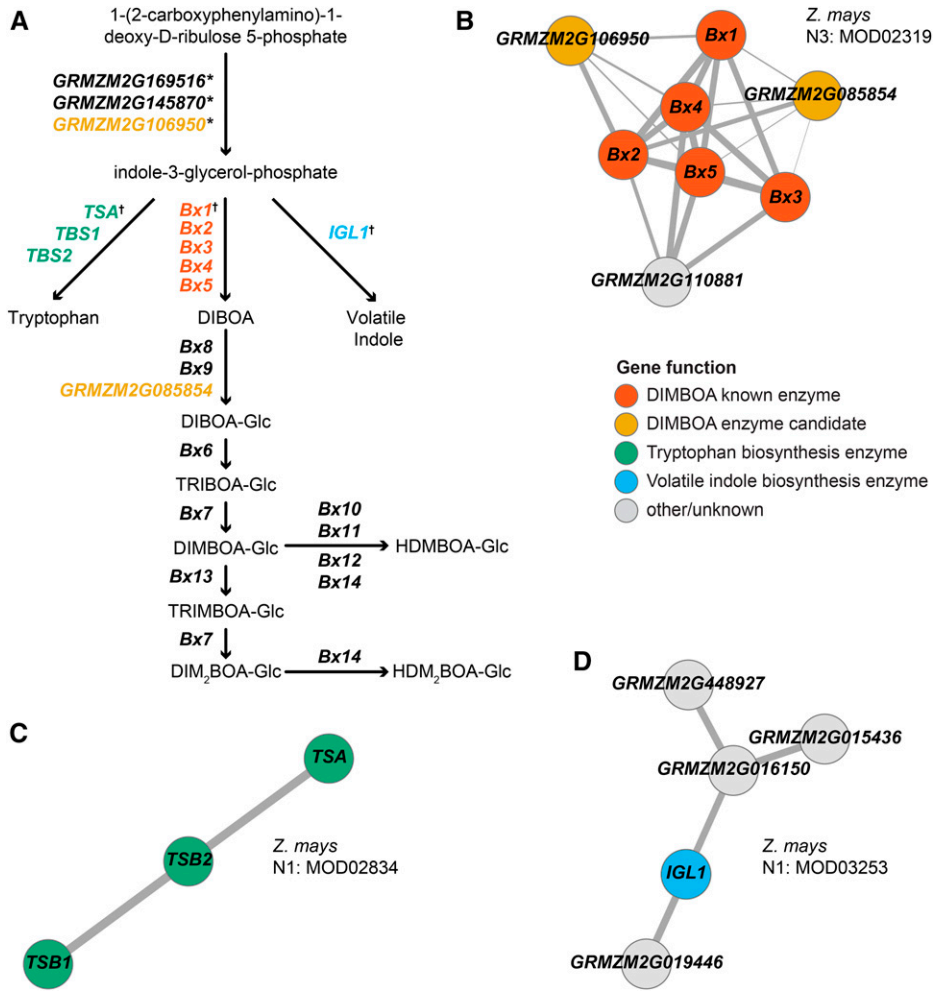

Figure 7. Pathway Diagram and Network Map of Benzoxazinoid Biosynthesis in Maize.

Diagram of benzoxazinoid biosynthesis and associated pathways in maize **(A)**. Of the three indole-3-glycerol phosphate synthases (*), only GRMZM2G106950 is significantly coexpressed with benzoxazinoid biosynthesis genes **(B)**. Similarly, each of the indole-3-glycerol phosphate lyases (†) are assigned to non-overlapping modules for the production of benzoxazinoids **(B)**, tryptophan **(C)**, and volatile indole **(D)**. Network maps were constructed as described in Figure 3. Nodes and gene names are colored according to their known or putative function. Benzoxazinoid genes not recovered in modules are colored black.

coexpression modules (Figure 9B). For example, three genes in a 14-gene EC-based BGC (containing a TS-CYP pair identified by Boutanaev et al. involved in the production of arabidiol; Sohrabi et al., 2015) were coexpressed in a module (Supplemental Figure 10). Moreover, 78/188 putative BGCs overlapped with coexpression modules by only one gene, indicating that the genes within these BGCs were more strongly coexpressed with genes outside their cluster than with those inside (Figure 9B; Supplemental Data Set 8).

An example of the poor association between coexpression modules and putative BGCs comes from the antiSMASH-predicted BGC30. Only 2/6 genes in BGC30 showed strong pairwise coexpression: a TS-CYP pair also identified by Boutanaev et al. and labeled PAIR6 (Figure 9C). The terpene synthase (*AT5G44630*) of PAIR6 is known to be involved in the production of sesquiterpenoid flower volatiles (Tholl et al., 2005). This functional annotation is supported by our module analysis, which assigned PAIR6 to a coexpression module consisting of 46 physically unlinked genes that are significantly enriched for gene ontology terms associated with flower development (Figure 9D; Supplemental Data Set 11). A second example comes from the EC-mapped BGC130. None of the genes in this BGC were strongly coexpressed with each other (Supplemental Figure 9). Instead, one gene in the BGC, *GSTU20*, is a known participant in metGSL biosynthesis, an association that is recovered by coexpression modules in our analysis (Figure 3; Supplemental Data Set 8).

DISCUSSION

An enormous number of novel plant SMs await discovery and characterization (Wurtzel and Kutchan, 2016). Yet, due to their

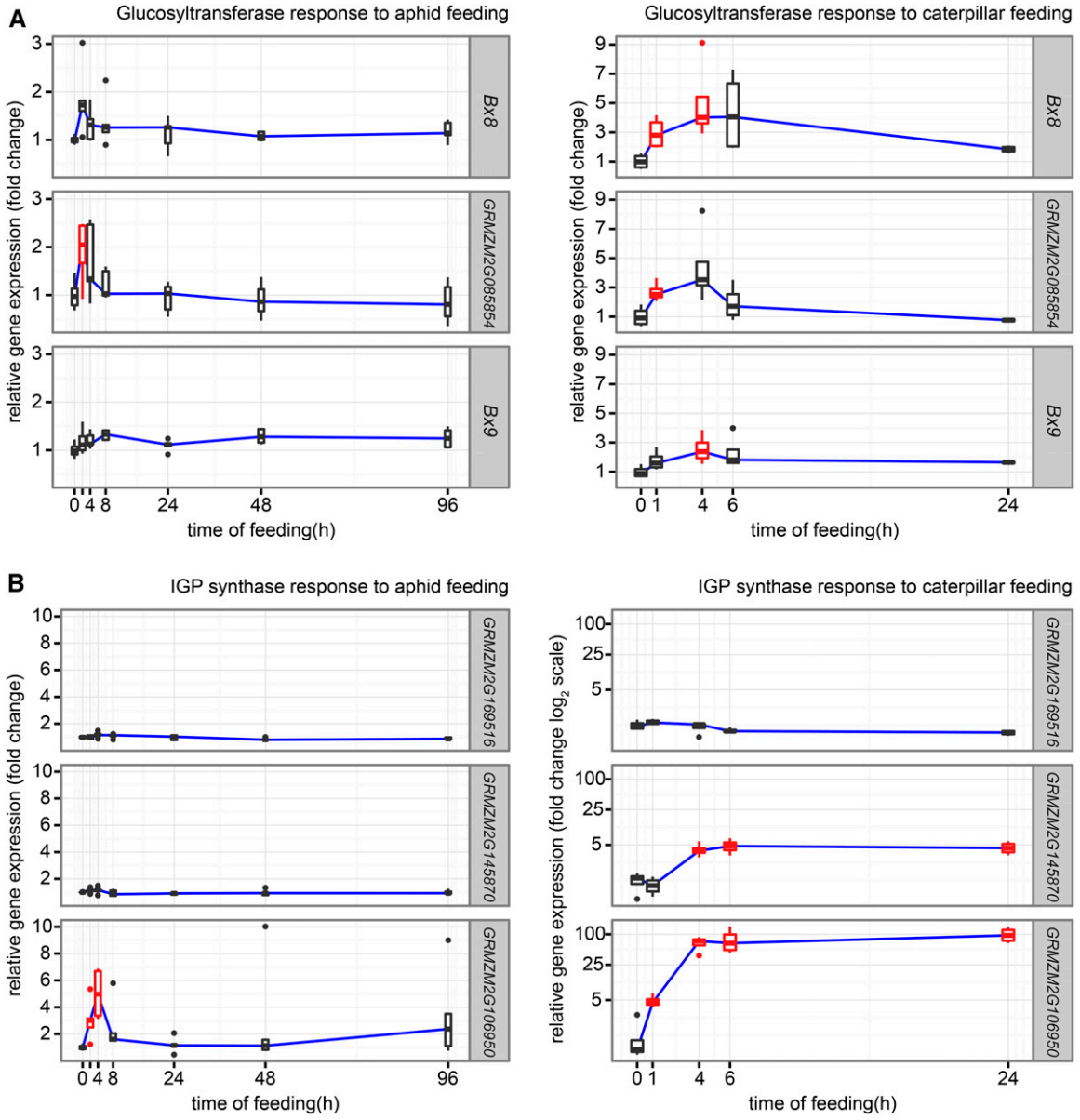

Figure 8. Gene Expression Response to Insect Feeding in Maize.

(A) Glucosyltransferase response to insect feeding.
(B) IGP synthase response to insect feeding. Box plots depict fold change relative to uninfested (0 h) controls ($n = 5$). Red box plots are significantly different from control ($P < 0.05$, Student's t test; median fold change > 2).

rapid evolution and narrow taxonomic distribution (Pichersky and Lewinsohn, 2011; Chae et al., 2014; Mukherjee et al., 2015), SM pathways and genes are often unknown, slowing the pace of discovery. Gene coexpression and chromosomal proximity are two omics-level traits that can be harnessed for high-throughput prediction of SM pathways and genes (Wurtzel and Kutchan, 2016), but their general utility remained unknown. By examining 10 global coexpression data sets—each a meta-analysis of 172 to 15,275 transcriptome experiments—across eight plant model organisms, we found that gene coexpression was powerful in identifying known SM pathways, irrespective of the location of their genes in the genome, as well as in predicting novel SM gene associations. Below, we discuss why gene proximity may not be a reliable method of SM pathway identification in plant genomes as well as enumerate the advantages and caveats of our coexpression network-based approach.

It is well established that genes in SM pathways are spatially and temporally regulated in response to diverse ecological conditions; arguably, this shared regulatory program is one of the defining characteristics uniting genes belonging to SM pathways (Tohge

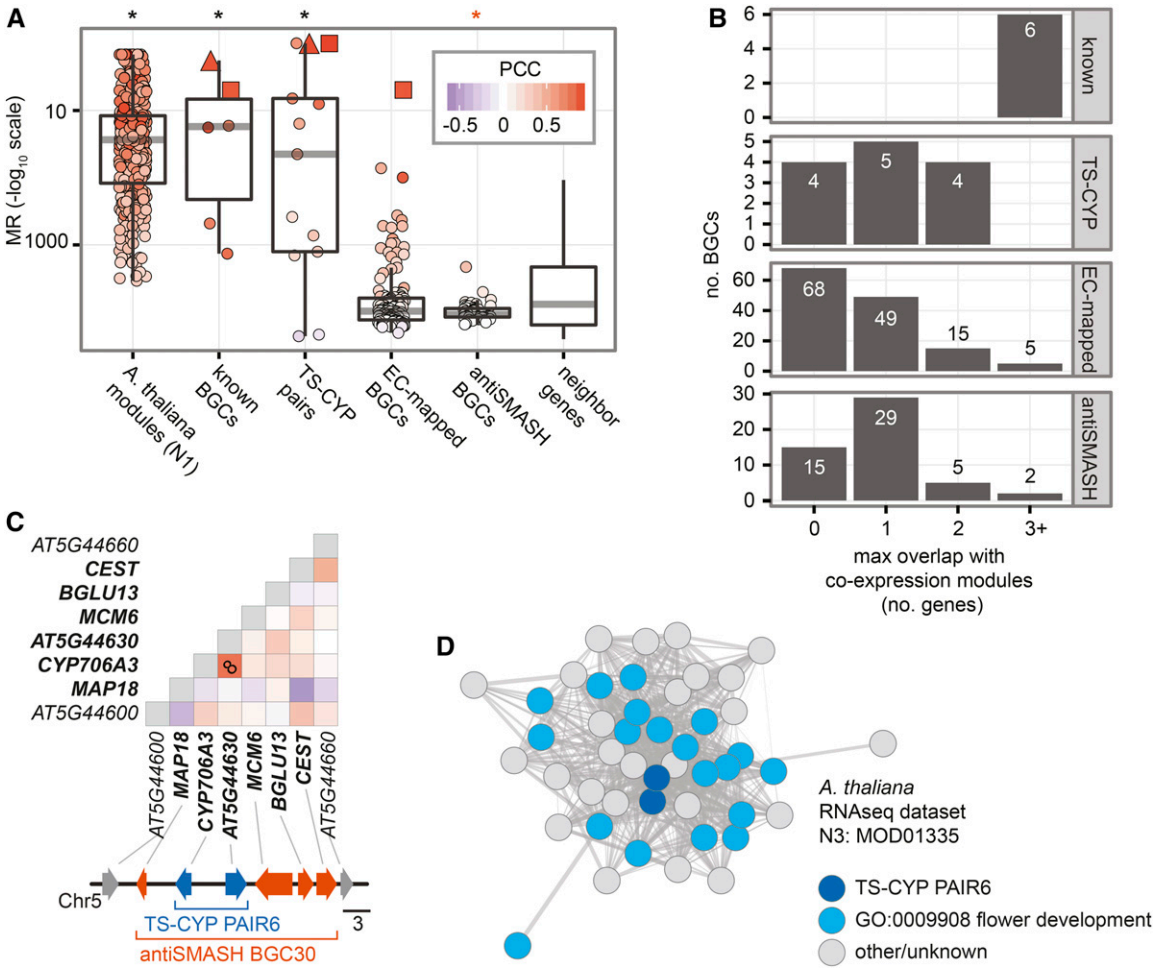

Figure 9. The Genes Comprising the Majority of Bioinformatically Predicted BGCs Are Not Coexpressed.

(A) Comparison of average coexpression of modules versus characterized and putative BGCs. The bottom and top of each box plot correspond to the first and third quartiles (the 25th and 75th percentiles), respectively. The lower whisker extends from the box bottom to the lowest value within 1.5 * IQR (interquartile range, defined as the distance between the first and third quartiles) of the first quartile. The upper whisker extends from the box top to the highest value that is within 1.5 * IQR of the third quartile. Red squares and triangles indicate BGCs or gene pairs that correspond to the all or part of the thalianol and marneral BGCs, respectively. Asterisks denote significant deviation from the control distribution of neighboring genes; *$P \leq 0.05$ (Wilcoxon rank sum tests).
(B) From top to bottom, histogram of maximum overlap between coexpression modules and known (characterized) BGCs, TS-CYP gene pairs, EC-mapped BGCs, and antiSMASH BGCs.
(C) Heat map depicting the correlation of coexpression for a eight-gene region of chromosome five in Arabidopsis containing an example antiSMASH BGC (BGC30) and TS-CYP gene pair (PAIR6). Diagonal number indicates MR score; squares are blank if MR $\geq$ 100. Heat map scale is the same as in **(A)**. BGC genes are bolded in the heat map and colored red in the accompanying chromosome segments (TS-CYP pair is colored dark blue). Scale bars are in kilobase pairs.
(D) Network map of a module that maximally overlaps with BGC30. Overlapping genes (TS-CYP PAIR6) are colored dark blue.

and Fernie, 2012; Hartmann, 2007; Grotewold, 2005). Furthermore, numerous gene expression studies of the genes participating in diverse SM pathways, including BGCs, from diverse organisms show that SM pathway genes typically share similar gene expression patterns (i.e., they are coexpressed). Simply put, gene coexpression can be predictive of membership in a given SM pathway. The question then is whether one can employ genome-wide or global gene expression data to predict SM pathways in a high-throughput fashion. The results of our analyses suggest that this is the case; modules in global coexpression networks constructed from genome-wide expression studies across myriads of different conditions in Arabidopsis were significantly enriched in genes associated with diverse SM-related metabolic functions (Figure 2A). Moreover, modules recovered many experimentally validated SM pathways in these plants (Supplemental Data Sets 5 and 8), including the six known to form BGCs (Figure 5).

It is also well established that gene arrangement in plant genomes is not random (Hurst et al., 2004). For example, as much as 60% of metabolic pathways in Arabidopsis (as measured by

KEGG [Kyoto Encyclopedia of Genes and Genomes] analysis) show statistically significant higher levels of physical proximity in the genome than expected by chance (Lee and Sonnhammer, 2003). The most extreme version of this "closer than expected" gene arrangement is the growing list BGCs involved in plant SM biosynthesis (Nützmann et al., 2016). While the statistical significance of this pattern is nondebatable, the degree to which gene arrangement is predictive of genes' participation in the same pathway is not immediately obvious. For example, the genes of many known plant SM pathways (Sønderby et al., 2010; Winkel-Shirley, 2001) do not form BGCs, while other pathways consist of a combination of clustered and unclustered genes (Itkin et al., 2013; Handrick et al., 2016; Zhou et al., 2016). Complicating matters further, SM pathways may form a BGC in some species but not others (Sue et al., 2011). Given that the majority of known plant SM pathways do not form BGCs, it is perhaps not surprising that nearly all putative plant SM BGCs, which were predicted based solely on gene proximity, were not coexpressed (Figure 9).

We interpret this absence of coexpression as evidence that most of these putative BGCs likely do not correspond with actual SM pathways and that gene proximity is insufficient to be used as the primary input for predicting SM pathways in plant genomes. Admittedly, the strength of this argument rests on whether the global coexpression networks that we have constructed accurately capture the spatial and temporal regulation of BGCs in response to the diverse ecological conditions plants experience, which is at least partially dependent on the number and types of the conditions sampled (Ballouz et al., 2015). For example, genes in a BGC or pathway that are never expressed or are not variably expressed across conditions would not be correlated with each other in our analysis. Although this is a valid concern, the hundreds to thousands of conditions (Aoki et al., 2016b) used to construct each coexpression data set (Supplemental Data Set 1), as well as the recovery of many known SM pathways from these organisms (Supplemental Data Sets 5 and 8), suggest that its effect is unlikely to influence our major findings. Going forward, increased resolution of BGCs and SM pathways in coexpression networks will require the inclusion of data from more tissues, time points, and environmental conditions during which SM genes and pathways are likely to vary in their regulation, for example, different types of insect herbivory (Handrick et al., 2016; Tzin et al., 2015, 2017; Ralph et al., 2006) and complex field conditions (Richards et al., 2012).

Another caveat associated with predicting SM pathways from global coexpression networks is that SM pathways whose expression profiles are highly similar would be predicted to comprise a single pathway. This will likely be a more common occurrence, and examples of this behavior are present in our results. Specifically, the two triterpenoid BGCs in Arabidopsis were almost always combined in the same coexpression module, regardless of the network investigated, and the same was true for the two diterpenoid BGCs in rice (Figure 6). Although predicting individual SM pathways is obviously ideal, the lumping of multiple pathways into one may in some cases reveal novel biology. For example, such a pattern could also be indicative of crosstalk between SM pathways or BGCs or that multiple SM pathways are employed in response to the same set of environmental conditions.

The final caveat is that our approach will not be as powerful in cases where some of the genes in the pathway are not under the same regulatory program as the others (Uygun et al., 2016). For example, we noted that the genes encoding terminal modification enzymes, such as the genes for side-chain modification of glucosinolates (Supplemental Figure 4) or the UDP-glucosyltransferases in tomato (*GAME2*) and maize (*Bx8-Bx14*), had expression profiles that were quite different from those of core pathway genes; thus, they were often not recovered in the same modules as their corresponding core SM pathway genes. It is possible that additional sampling of appropriate expression conditions could allow for recovery of these terminal metabolic branches in coexpression modules that include the rest of the pathway. However, the terminal SM genes and products can be under balancing or diversifying selection (Kerwin et al., 2015); moreover, the core and terminal steps in an SM pathway may take place in different tissues (Hartmann and Ober, 2000). In cases like these, the terminal metabolic branches and core SM pathway may be identified as distinct coexpression modules in global coexpression networks no matter how many conditions are sampled.

In summary, our results indicate that generating and constructing global gene coexpression networks is a powerful and promising approach to the challenge of high-throughput prediction and study of plant SM pathways. Global gene coexpression networks can straightforwardly be constructed for any species, model or non-model, as long as the organism's transcriptome can be sampled under a range of conditions. In principle, this would not require a genome sequence, only a high quality de novo transcriptome assembly. Future use of global coexpression networks could include identification of new genes associated with known SM pathways (e.g., glucosinolate and benzoxazinoid pathways). Furthermore, uncharacterized coexpression modules could be cross-referenced with other high-throughput data types (e.g., proteomics, metabolomics) to identify new SM pathways. We believe that combining high-throughput transcriptomics across ecological conditions with network biology will transform our understanding of the genetic basis and architecture of plant natural products and usher in a new era of exploration of their chemodiversity.

METHODS

Coexpression Network Analysis

Genome annotations and protein sequences were downloaded from NCBI RefSeq and JGI Genome Portal databases (Supplemental Data Set 1). Condition-independent gene coexpression values, measured using PCC and MR, across the eight plant species were downloaded from the ATTED-II (Aoki et al., 2016b), ALCOdb (Aoki et al., 2016a), and data sets with <50% coverage of the target genome were excluded (Supplemental Data Set 1). The MR score for two example genes A and B is given by the formula:

$$MR_{(AB)} = \sqrt{Rank_{(A\rightarrow B)} \times Rank_{(B\rightarrow A)}}$$

where $Rank_{(A\rightarrow B)}$ is the rank of gene B in a PCC-ordered list of gene A against all other genes in the microarray or RNA-seq meta-analysis; similarly, $Rank_{(B\rightarrow A)}$ is the rank of gene A in a PCC-ordered list of gene B against all other genes, with smaller MR scores indicating stronger coexpression between gene pairs (Obayashi and Kinoshita, 2009). MR scores

were converted to network edge weights using five different rates of exponential decay (Figure 1). Any edge with PCC <0.3 or edge weight <0.01 was excluded.

Comparison of MR- and PCC-based networks showed that the MR-based networks were more comparable between species and data sets. For example, PCC-based networks were more sensitive (variable) to differences in the number of experimental samples and genome coverage between data sets in the two species that had microarray- and RNA-seq-based data sets (*Arabidopsis thaliana* and rice [*Oryza sativa*]). In contrast, the MR-based networks were more robust to data set differences (Supplemental Figure 2), in agreement with the original description of the MR metric by Obayashi and Kinoshita (2009). Moreover, MR-based networks were remarkably consistent with respect to the number of genes they contained; in contrast, PCC-based networks sometimes varied by orders of magnitude in the number of genes included (Supplemental Figure 2). Finally, MR-based networks consistently included nearly all genes in a given data set, regardless of the MR threshold stringency employed; that was not the case with PCC-based networks (Supplemental Figure 2 and Supplemental Data Set 2). For these reasons, we chose to focus the investigation on the MR-based networks.

Modules of tightly coexpressed genes were detected using ClusterONE using default parameters (Nepusz et al., 2012). Modules with ClusterONE P value > 0.1 were excluded. Modules were considered "SM like" if they contained two or more nonhomologous genes with a significant match to a curated list of Pfam domains present in experimentally verified (evidence = EV-EXP) genes assigned to MetaCyc (Caspi et al., 2016) secondary biosynthesis pathways (hmmsearch using default inclusion thresholds; Eddy, 2011) (Supplemental Data Set 6). SM-like modules were then binned into meta-modules of nonoverlapping gene sets. Coexpression modules identified in this analysis are included in Supplemental File 1.

Bioinformatically predicted BGCs were obtained from the published literature (Boutanaev et al., 2015; Chae et al., 2014) and by running the Arabidopsis reference genome (TAIR10; each protein-coding gene was represented by its longest transcript) through antiSMASH v3.0.4 (Weber et al., 2015) with the –clusterblast–subclusterblast–smcogs options enabled. Average coexpression of each gene set (module or BGC) was calculated as the average MR score across all gene pairs in the set.

All statistical analyses were performed in R, including dhyper (hypergeometric), wilcox.test (Wilcoxon rank sum), and p.adjust (Benjamini and Hochberg adjusted P value) from the stats package. Network maps were drawn using a Fruchterman-Reingold force-directed layout using the igraph R package (http://igraph.org).

Phylogenetic Analysis

Transcripts of Brassicaceae *MAM* and *GSTU* genes were downloaded from EnsemblPlants (Kersey et al., 2016). Sequences were aligned and masked using the GUIDANCE2 server (Sela et al., 2015) using the codon setting and the MAFFT multiple sequence alignment algorithm (Katoh and Standley, 2013); residues with guidance scores <0.9 were masked. The gene phylogenies were inferred using maximum likelihood as implemented in RAxML version 8.0.25 (Stamatakis, 2014) using rapid bootstrapping (1000 replications) and a GTRGAMMAIX substitution model, which was the best model as indicated by the Bayesian Information Criterion in IQ-TREE version 1.3.8 (Nguyen et al., 2015). The phylogenies were midpoint rooted, and branches with <50% bootstrap support were collapsed using TREECOLLAPSECL version 4.0 (http://emmahodcroft.com/TreeCollapseCL.html, last accessed October 24, 2016). The alignments and Newick trees can be found in Supplemental Files 2 and 3.

Insect Herbivory Experiments

Normalized expression levels of *GT* and *IGPS* genes from a maize (*Zea mays*) B73 inbred line were taken from two earlier investigations of maize leaf aphid (*Rhopalosiphum maidis*) and caterpillar (*Spodoptera exigua*) feeding on maize (Tzin et al., 2015, 2017).

Accession Numbers

GenBank GeneID/TAIR/MaizeGDB identifiers for genes referenced in this article can be found in Supplemental Data Sets 7 and 9. PCCs and MRs used in this work are available for download at ATTED-II http://atted.jp/download.shtml. Coexpression modules identified in this analysis are included in Supplemental File 1.

Supplemental Data

Supplemental Figure 1. MetaCyc pathway enrichment analysis of experimentally characterized genes in Arabidopsis.

Supplemental Figure 2. Comparison of mutual rank-based and Pearson's correlation-based networks.

Supplemental Figure 3. Overlapping coexpressed modules recover the pathway for metGSL biosynthesis in Arabidopsis.

Supplemental Figure 4. Comparison of degree of gene coexpression in core versus terminal modification genes in metGSL biosynthesis.

Supplemental Figure 5. Maximum likelihood phylogeny of Brassicaceae MAM and IPMS sequences.

Supplemental Figure 6. Maximum likelihood phylogeny of Brassicaceae GSTU sequences.

Supplemental Figure 7. Network maps of coexpression modules involved in thalianol and marneral triterpenoid biosynthesis in Arabidopsis.

Supplemental Figure 8. Network map of coexpression module involved in momilactone and phytocassane diterpenoid biosynthesis in rice.

Supplemental Figure 9. Network maps of coexpression modules involved in tomatine biosynthesis in tomato.

Supplemental Figure 10. Coexpression pattern of seven putative BGCs in plants.

Supplemental Data Set 1. Downloaded data sets.

Supplemental Data Set 2. Descriptive statistics for coexpression networks.

Supplemental Data Set 3. Arabidopsis genes assigned to MetaCyc pathways and pathway ontologies.

Supplemental Data Set 4. Test for enrichment/depletion of MetaCyc pathway categories and classes in module genes.

Supplemental Data Set 5. Recovery of MetaCyc pathways in coexpression modules.

Supplemental Data Set 6. List of Pfam domains found in SM pathways in MetaCyc.

Supplemental Data Set 7. metGSL biosynthesis genes in Arabidopsis and *B. rapa*.

Supplemental Data Set 8. Recovery of metGSL pathways, characterized BGCs, and putative BGCs in coexpression modules.

Supplemental Data Set 9. List of functionally characterized BGCs in plants with coexpression data on ATTED-II.

Supplemental Data Set 10. Average coexpression of gene modules, characterized BGCs, and putative BGCs.

Supplemental Data Set 11. GO enrichment test of a 46-gene Arabidopsis module involved in flower development.

Supplemental File 1. Coexpression modules.

Supplemental File 2. Brassicaceae MAM gene family alignment and tree.

Supplemental File 3. Brassicaceae GSTU gene family alignment and tree.

ACKNOWLEDGMENTS

We thank members of the Rokas lab and the National Science Foundation's Plant Genome Research Program for helpful discussions. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This material is based upon work supported by the National Science Foundation (http://www.nsf.gov) under Grants IOS-1401682 to J.H.W., DEB-1442113 to A.R., and IOS-1339237 to G.J.

AUTHOR CONTRIBUTIONS

J.H.W., V.T., G.J., D.J.K., and A.R. conceived and designed the experiments. J.H.W., A.T.B., and V.T. performed the experiments. J.H.W., A.T.B., V.T., G.J., D.J.K., and A.R. analyzed the data. J.H.W., V.T., G.J., and A.R. contributed reagents/materials/analysis tools. J.H.W. and A.R. wrote the paper. All authors read, commented on, and approved the manuscript.

Received January 6, 2017; revised March 12, 2017; accepted April 9, 2017; published April 13, 2017.

REFERENCES

Andersen, M.R., Nielsen, J.B., Klitgaard, A., Petersen, L.M., Zachariasen, M., Hansen, T.J., Blicher, L.H., Gotfredsen, C.H., Larsen, T.O., Nielsen, K.F., and Mortensen, U.H. (2013). Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proc. Natl. Acad. Sci. USA **110:** E99–E107.

Aoki, Y., Okamura, Y., Ohta, H., Kinoshita, K., and Obayashi, T. (2016a). ALCOdb: Gene coexpression database for microalgae. Plant Cell Physiol. **57:** e3.

Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., and Obayashi, T. (2016b). ATTED-II in 2016: A plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol. **57:** e5.

Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics **31:** 2123–2130.

Boachon, B., et al. (2015). CYP76C1 (cytochrome P450)-mediated linalool metabolism and the formation of volatile and soluble linalool oxides in *Arabidopsis* flowers: a strategy for defense against floral antagonists. Plant Cell **27:** 2972–2990.

Boutanaev, A.M., Moses, T., Zi, J., Nelson, D.R., Mugford, S.T., Peters, R.J., and Osbourn, A. (2015). Investigation of terpene diversification across multiple sequenced plant genomes. Proc. Natl. Acad. Sci. USA **112:** E81–E88.

Brachi, B., Meyer, C.G., Villoutreix, R., Platt, A., Morton, T.C., Roux, F., and Bergelson, J. (2015). Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana.* Proc. Natl. Acad. Sci. USA **112:** 4032–4037.

Bradshaw, R.E., Slot, J.C., Moore, G.G., Chettri, P., de Wit, P.J.G. M., Ehrlich, K.C., Ganley, A.R.D., Olson, M.A., Rokas, A., Carbone, I., and Cox, M.P. (2013). Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. New Phytol. **198:** 525–535.

Caspi, R., et al. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **44** (D1): D471–D480.

Castillo, D.A., Kolesnikova, M.D., and Matsuda, S.P.T. (2013). An effective strategy for exploring unknown metabolic pathways by genome mining. J. Am. Chem. Soc. **135:** 5885–5894.

Cárdenas, P.D., et al. (2016). GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. Nat. Commun. **7:** 10654.

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S.Y. (2014). Genomic signatures of specialized metabolism in plants. Science **344:** 510–513.

Cimermancic, P., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell **158:** 412–421.

D'Auria, J.C., and Gershenzon, J. (2005). The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. Curr. Opin. Plant Biol. **8:** 308–316.

De Luca, V., Salim, V., Atsumi, S.M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. Science **336:** 1658–1661.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLOS Comput. Biol. **7:** e1002195.

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

Eisen, J.A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. **8:** 163–167.

Erb, M., Veyrat, N., Robert, C.A.M., Xu, H., Frey, M., Ton, J., and Turlings, T.C.J. (2015). Indole is an essential herbivore-induced volatile priming signal in maize. Nat. Commun. **6:** 6273.

Field, B., and Osbourn, A.E. (2008). Metabolic diversification–independent assembly of operon-like gene clusters in different plants. Science **320:** 543–547.

Field, B., Fiston-Lavier, A.-S., Kemen, A., Geisler, K., Quesneville, H., and Osbourn, A.E. (2011). Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc. Natl. Acad. Sci. USA **108:** 16116–16121.

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grün, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R.B., Briggs, S.P., Simcox, K., and Gierl, A. (1997). Analysis of a chemical plant defense mechanism in grasses. Science **277:** 696–699.

Frey, M., Schullehner, K., Dick, R., Fiesselmann, A., and Gierl, A. (2009). Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. Phytochemistry **70:** 1645–1651.

Frey, M., Stettner, C., Pare, P.W., Schmelz, E.A., Tumlinson, J.H., and Gierl, A. (2000). An herbivore elicitor activates the gene for indole emission in maize. Proc. Natl. Acad. Sci. USA **97:** 14801–14806.

Fridman, E., and Pichersky, E. (2005). Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. Curr. Opin. Plant Biol. **8:** 242–248.

Geu-Flores, F., Nielsen, M.T., Nafisi, M., Møldrup, M.E., Olsen, C.-E., Motawia, M.S., and Halkier, B.A. (2009). Glucosinolate engineering identifies a gamma-glutamyl peptidase. Nat. Chem. Biol. **5:** 575–577.

Gibbons, J.G., Beauvais, A., Beau, R., McGary, K.L., Latgé, J.-P., and Rokas, A. (2012a). Global transcriptome changes underlying colony growth in the opportunistic human pathogen *Aspergillus fumigatus.* Eukaryot. Cell **11:** 68–78.

Gibbons, J.G., Salichos, L., Slot, J.C., Rinker, D.C., McGary, K.L., King, J.G., Klich, M.A., Tabb, D.L., McDonald, W.H., and Rokas, A. (2012b). The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae.* Curr. Biol. **22:** 1403–1409.

Grotewold, E. (2005). Plant metabolic diversity: a regulatory perspective. Trends Plant Sci. **10:** 57–62.

Guo, J., et al. (2016). Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. New Phytol. **210:** 525–534.

Hadjithomas, M., et al. (2015). IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio **6:** e00932.

Handrick, V., et al. (2016). Biosynthesis of 8-o-methylated benzoxazinoid defense compounds in maize. Plant Cell **28:** 1682–1700.

Hartmann, T. (2007). From waste products to ecochemicals: fifty years research of plant secondary metabolism. Phytochemistry **68:** 2831–2846.

Hartmann, T., and Ober, D. (2000). Biosynthesis and metabolism of pyrrolizidine alkaloids in plants and specialized insect herbivores. In Biosynthesis: Aromatic Polyketides, Isoprenoids, Alkaloids, F.J. Leeper and J.C. Vederas, eds (Berlin, Heidelberg, Germany: Springer), pp. 207–243.

Hirai, M.Y., et al. (2007). Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. USA **104:** 6478–6483.

Hiss, M., et al. (2014). Large-scale gene expression profiling data for the model moss *Physcomitrella patens* aid understanding of developmental progression, culture and stress conditions. Plant J. **79:** 530–539.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J.F., Zhu, J.-K., Cushman, J.C., Gollery, M., and Girke, T. (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol. **147:** 41–57.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004). The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5:** 299–310.

Itkin, M., et al. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science **341:** 175–179.

Itkin, M., et al. (2016). The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii.* Proc. Natl. Acad. Sci. USA **113:** E7619–E7628.

Jonczyk, R., Schmidt, H., Osterrieder, A., Fiesselmann, A., Schullehner, K., Haslbeck, M., Sicker, D., Hofmann, D., Yalpani, N., Simmons, C., Frey, M., and Gierl, A. (2008). Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of *Bx6* and *Bx7.* Plant Physiol. **146:** 1053–1063.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30:** 772–780.

Kersey, P.J., et al. (2016). Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. **44:** D574–D580.

Kerwin, R., et al. (2015). Natural genetic variation in *Arabidopsis thaliana* defense metabolism genes modulates field fitness. eLife **4:** e05604.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet. Biol. **47:** 736–741.

Kliebenstein, D.J., and Osbourn, A. (2012). Making new molecules - evolution of pathways for novel metabolites in plants. Curr. Opin. Plant Biol. **15:** 415–423.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J., and Mitchell-Olds, T. (2001). Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis.* Plant Cell **13:** 681–693.

Lau, W., and Sattely, E.S. (2015). Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. Science **349:** 1224–1228.

Lawler, K., Hammond-Kosack, K., Brazma, A., and Coulson, R.M.R. (2013). Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus *Fusarium graminearum.* BMC Syst. Biol. **7:** 52.

Lee, J.M., and Sonnhammer, E.L.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. **13:** 875–882.

Lind, A.L., Smith, T.D., Saterlee, T., Calvo, A.M., and Rokas, A. (2016). Regulation of secondary metabolism by the Velvet complex is temperature-responsive in *Aspergillus.* G3 (Bethesda) **6:** 4023–4033.

Lo, H.-C., Entwistle, R., Guo, C.-J., Ahuja, M., Szewczyk, E., Hung, J.-H., Chiang, Y.-M., Oakley, B.R., and Wang, C.C.C. (2012). Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in *Aspergillus nidulans.* J. Am. Chem. Soc. **134:** 4709–4720.

Lodeiro, S., Xiong, Q., Wilson, W.K., Kolesnikova, M.D., Onak, C.S., and Matsuda, S.P.T. (2007). An oxidosqualene cyclase makes numerous products by diverse mechanisms: a challenge to prevailing concepts of triterpene biosynthesis. J. Am. Chem. Soc. **129:** 11213–11222.

Maeda, H., Yoo, H., and Dudareva, N. (2011). Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. Nat. Chem. Biol. **7:** 19–21.

Mao, L., Van Hemert, J.L., Dash, S., and Dickerson, J.A. (2009). *Arabidopsis* gene co-expression network and its functional modules. BMC Bioinformatics **10:** 346.

McChesney, J.D., Venkataraman, S.K., and Henri, J.T. (2007). Plant natural products: back to the future or into extinction? Phytochemistry **68:** 2015–2022.

Medema, M.H., and Osbourn, A. (2016). Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Nat. Prod. Rep. **33:** 951–962.

Meihls, L.N., Handrick, V., Glauser, G., Barbier, H., Kaur, H., Haribal, M.M., Lipka, A.E., Gershenzon, J., Buckler, E.S., Erb, M., Köllner, T.G., and Jander, G. (2013). Natural variation in maize aphid resistance is associated with 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one glucoside methyltransferase activity. Plant Cell **25:** 2341–2355.

Mentzen, W.I., and Wurtele, E.S. (2008). Regulon organization of *Arabidopsis.* BMC Plant Biol. **8:** 99.

Mugford, S.G., et al. (2009). Disruption of adenosine-5′-phosphosulfate kinase in *Arabidopsis* reduces levels of sulfated secondary metabolites. Plant Cell **21:** 910–927.

Mukherjee, D., Mukherjee, A., and Ghosh, T.C. (2015). Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in *Arabidopsis thaliana.* Genome Biol. Evol. **8:** 17–28.

Naoumkina, M.A., Modolo, L.V., Huhman, D.V., Urbanczyk-Wochniak, E., Tang, Y., Sumner, L.W., and Dixon, R.A. (2010). Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in *Medicago truncatula.* Plant Cell **22:** 850–866.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods **9:** 471–472.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32:** 268–274.

Nützmann, H.-W., Huang, A., and Osbourn, A. (2016). Plant metabolic clusters - from genetics to genomics. New Phytol. **211:** 771–789.

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. DNA Res. **16:** 249–260.

Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet. **26:** 449–457.

Pichersky, E., and Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. Annu. Rev. Plant Biol. **62:** 549–566.

Rajniak, J., Barco, B., Clay, N.K., and Sattely, E.S. (2015). A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. Nature 525: 376–379.

Ralph, S.G., et al. (2006). Conifer defence against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. Plant Cell Environ. **29:** 1545–1570.

Raskin, I., et al. (2002). Plants and human health in the twenty-first century. Trends Biotechnol. **20:** 522–531.

Richards, C.L., Rosas, U., Banta, J., Bhambhra, N., and Purugganan, M.D. (2012). Genome-wide patterns of *Arabidopsis* gene expression in nature. PLoS Genet. **8:** e1002662.

Sanchez, J.F., Entwistle, R., Hung, J.-H., Yaegashi, J., Jain, S., Chiang, Y.-M., Wang, C.C.C., and Oakley, B.R. (2011). Genome-based deletion analysis reveals the prenyl xanthone biosynthesis pathway in *Aspergillus nidulans.* J. Am. Chem. Soc. **133:** 4010–4017.

Sawada, Y., Toyooka, K., Kuwahara, A., Sakata, A., Nagano, M., Saito, K., and Hirai, M.Y. (2009). *Arabidopsis* bile acid:sodium symporter family protein 5 is involved in methionine-derived glucosinolate biosynthesis. Plant Cell Physiol. **50:** 1579–1586.

Schlapfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., Kahn, D., and Rhee, S.Y. (2017). Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. Plant Physiol. 173: 2071–2059.

Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. **43:** W7–W14.

Shimura, K., et al. (2007). Identification of a biosynthetic gene cluster in rice for momilactones. J. Biol. Chem. **282:** 34013–34018.

Sohrabi, R., Huh, J.-H., Badieyan, S., Rakotondraibe, L.H., Kliebenstein, D.J., Sobrado, P., and Tholl, D. (2015). In planta variation of volatile biosynthesis: an alternative biosynthetic route to the formation of the pathogen-induced volatile homoterpene DMNT via triterpene degradation in *Arabidopsis* roots. Plant Cell **27:** 874–890.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:** 1312–1313.

Sue, M., Nakamura, C., and Nomura, T. (2011). Dispersed benzoxazinone gene cluster: molecular characterization and chromosomal localization of glucosyltransferase and glucosidase genes in wheat and rye. Plant Physiol. **157:** 985–997.

Swaminathan, S., Morrone, D., Wang, Q., Fulton, D.B., and Peters, R.J. (2009). CYP76M7 is an *ent*-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. Plant Cell **21:** 3315–3325.

Sønderby, I.E., Geu-Flores, F., and Halkier, B.A. (2010). Biosynthesis of glucosinolates–gene discovery and beyond. Trends Plant Sci. **15:** 283–290.

Tholl, D., Chen, F., Petri, J., Gershenzon, J., and Pichersky, E. (2005). Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from *Arabidopsis* flowers. Plant J. **42:** 757–771.

Tohge, T., and Fernie, A.R. (2012). Co-expression and co-responses: within and beyond transcription. Front. Plant Sci. **3:** 248.

Town, C.D., et al. (2006). Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell **18:** 1348–1359.

Tzin, V., et al. (2015). Dynamic maize responses to aphid feeding are revealed by a time series of transcriptomic and metabolomic assays. Plant Physiol. **169:** 1727–1743.

Tzin, V., Hojo, Y., Strickler, S.R., Bartsch, L.J., Archer, C.M., Ahern, K.R., Christensen, S.A., Galis, I., Mueller, L.A., and Jander, G. (2017). *Spodoptera exigua* caterpillar feeding induces rapid defense responses in maize leaves. bioRxiv, http://dx.doi.org/10.10.1101/108076.

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

Uygun, S., Peng, C., Lehti-Shiu, M.D., Last, R.L., and Shiu, S.-H. (2016). Utility and limitations of using gene expression data to identify functional associations. PLOS Comput. Biol. **12:** e1005244.

Wang, H., Wu, J., Sun, S., Liu, B., Cheng, F., Sun, R., and Wang, X. (2011). Glucosinolate biosynthetic genes in *Brassica rapa.* Gene **487:** 135–142.

Weber, T., et al. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. **43:** W237–W243.

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiol. **126:** 485–493.

Wurtzel, E.T., and Kutchan, T.M. (2016). Plant metabolism, the diverse chemistry set of the future. Science **353:** 1232–1236.

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

Yu, J., Fedorova, N.D., Montalbano, B.G., Bhatnagar, D., Cleveland, T.E., Bennett, J.W., and Nierman, W.C. (2011). Tight control of mycotoxin biosynthesis gene expression in *Aspergillus flavus* by temperature as revealed by RNA-seq. FEMS Microbiol. Lett. **322:** 145–149.

Zhou, Y., et al. (2016). Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. Nat. Plants **2:** 16183.

Zimmermann, P., Laule, O., Schmitz, J., Hruz, T., Bleuler, S., and Gruissem, W. (2008). Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. Mol. Plant **1:** 851–857.