A Global Co-expression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants

Jennifer H. Wisecaver¹, Alexander T. Borowsky¹, Vered Tzin², Georg Jander³, Daniel J. Kliebenstein⁴, and Antonis Rokas¹,*

¹Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA
²French Associates Institute for Agriculture and Biotechnology of Drylands, Jacob Blaustein Institute for Desert Research, Ben Gurion University, Sede-Boqer Campus 84990, Israel
³Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY 14853, USA
⁴Department of Plant Sciences, University of California-Davis, One Shields Ave, Davis, CA 95616, USA
*Corresponding Author: antonis.rokas@vanderbilt.edu

Short title: Co-expression of plant specialized pathways

One-sentence summary: Global gene co-expression network analysis is a powerful approach for identifying known and novel plant-specialized metabolic pathways, irrespective of the genomic location of the constitutive genes.

The author(s) 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).

ABSTRACT
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 co-regulation, we hypothesized that genes within a SM pathway would form tight associations (modules) with each other in co-expression networks, facilitating their identification. To evaluate this hypothesis, we used 10 global co-expression datasets, each a meta-analysis of hundreds to thousands of experiments, across eight plant species to identify hundreds of co-expressed gene modules per dataset. In support of our hypothesis, 15.3–52.6% 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 co-expressed 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 co-expression is a rich, largely untapped resource for discovering the genetic basis and architecture of plant natural products.

©2017 American Society of Plant Biologists. All Rights Reserved
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).

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 to 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., 2016). 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 co-expression data, as a proxy for co-regulation, 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). Further, given the availability of data from hundreds to thousands of individual gene expression experiments, integrative global co-expression 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 co-expression 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 co-expression 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 co-locate in the genome, forming biosynthetic gene clusters (BGCs) (Osbourn, 2010). As expected of SM pathways, genes within these microbial BGCs are co-regulated and display strong signatures of co-expression, 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 co-expression 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 co-expressed, independently of being organized into BGCs, in line with their ecological roles that typically require strong temporal and spatial co-regulation (Tohge and Fernie, 2012; Hartmann, 2007; Grotewold, 2005). To test our hypothesis, we developed a gene co-expression network-based approach for plant SM pathway discovery (Figure 1). Using data from 10 meta-analyses of global co-expression that collectively contain 21,876 microarray or RNA-Seq experiments across eight plant species, we identified dozens to hundreds of modules of co-expressed 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 co-expression patterns than their non-clustered 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 co-expression...
data, when in abundance, are very powerful in the high-throughput identification of plant SM pathways.

RESULTS

Network analysis identifies small, overlapping modules of co-expressed genes in global co-expression networks

Given that SM pathway genes are often co-regulated 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 co-expression networks. To identify modules of co-expressed SM genes, we accessed three microarray- and seven RNAseq-based co-expression datasets 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 dataset 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 dataset, from 172 in the C. reinhardtii RNAseq dataset (Aoki et al., 2016a) to 15,275 in the A. thaliana microarray-based set (Aoki et al., 2016b). Pairwise measurements of gene co-expression were specified as Mutual Ranks (Obayashi and Kinoshita, 2009) (MRs; 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 dataset, we constructed five MR-based networks, each using a different co-expression 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 co-expressed 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 non-linear, containing multiple branch points and alternative end products (e.g., terpenoid biosynthesis pathways; Guo et al., 2016; Lodeiro et al., 2007). Averaging across all 10 co-expression datasets, the number of genes assigned to modules ranged from 3,251 (13.4% of protein-coding genes) in the N1 networks to 4,320 (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.

**Co-expressed 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 *A. thaliana* 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 datasets 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 co-expressed 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 datasets ($P < 0.05$, hypergeometric test) (Figure 2A and Supplemental Figure 1). None of the other metabolic classes displayed any significant variation between module and non-module genes (Supplemental Figure 1 and 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 non-homologous 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 non-intersecting “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 co-expression 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 co-expression modules identified by our analyses (Supplemental Data Set 7). In A. thaliana, the species with the majority of functional data (Sønderby et al., 2010), co-expression 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 co-expressed 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 co-expressed modules, including GGP1, which encodes a class I glutamine amidotransferase-like protein. Microarray-based co-expression data weakly associate GGP1 with metGSL biosynthesis in A. thaliana, 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 RNAseq-based networks showed that a different class I glutamine amidotransferase-like gene, DJ1F, is more highly co-expressed with metGSL biosynthetic genes (Figure 3). Importantly, DJ1F is not represented on the A. thaliana 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 co-expressed 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
*A. thaliana* (ecotype: Columbia) reference genome (Kliebenstein et al., 2001). The high
level of natural variation present in these terminal metabolic branches is 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 co-expression data (Supplemental Figure 4).

Brassicas also produce aliphatic glucosinolates, but a whole genome triplication
event subsequent to their divergence from *A. thaliana* (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 co-expression modules
with 59 candidate metGSL genes identified based on orthology to *A. thaliana* metGSL
genes (Wang et al., 2011). As in *A. thaliana*, modules recovered every biochemical step
of the *B. rapa* metGSL pathway as well as pathway-specific transporters and
transcription factors (Figure 4, Supplemental Data Set 7, and Supplemental Data Set 8).

Also as in *A. thaliana*, DJ1F rather than GGP1 is co-expressed 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 co-
expressed 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 and 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 *A. thaliana*, 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
*A. thaliana* 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 and 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 co-expression 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, co-expression 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 *A. thaliana* and the diterpenoid momilactone (Shimura et al., 2007) (5 / 5 genes; Figure 5C) in *O. sativa*. Modules recovered 7 / 9 genes in the phytocassane (Swaminathan et al., 2009) diterpene cluster in *O. sativa* (rice); the *OsKSL5* and *CYP71Z6* genes forming a terpene synthase-cytochrome p450 pair of genes were strongly co-expressed with each other but not with the rest of the pathway (Figure 5D).

The two triterpenoid BGCs in *A. thaliana* were typically combined into the same co-expression module (Figure 6A and Supplemental Figure 7); the same pattern was observed for the two diterpenoid BGCs in *O. sativa* (Figure 6B and Supplemental Figure 8). Genes within these BGCs were also strongly co-expressed with additional genes located outside the BGC boundaries, including one putative transcription factor and several putative transporters (Figure 6, Supplemental Figure 7 and Supplemental Figure 8).

Seven of eight genes in the partially clustered pathway for production of the steroidal alkaloid α-tomatine in *S. lycopersicum* (Itkin et al., 2013) (tomato) were recovered by our co-expression 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 co-expressed with the rest of the genes in this pathway (Figure 6C).
and Supplemental Figure 9), but whether or not these genes participate in α-tomatine biosynthesis is yet to be determined. Additional genes strongly co-expressed with the rest of the α-tomatine pathway include, among others, one putative transcription factor and several possible metabolite transporters (Figure 6C and 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 Z. mays formed co-expression 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 A. thaliana, 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, \( \text{DIM}_2\text{BOA-Glc} \)) (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 (IGL) that converts indole-3-glycerolphosphate to indole. However, in our module analysis, an additional gene co-expressed with the core Bx genes is an indole-3-glycerolphosphate synthase gene (IGPS; GRMZM2G106950), which catalyzes the reaction directly upstream of Bx1 (Figure 7). Two additional genes encoding indole-3-glycerolphosphate synthases are present in Z. mays (GRMZM2G169516 and GRMZM2G145870), but neither was strongly co-expressed with those in the benzoxazinoid pathway. Similarly, the two additional paralogs to Bx1 in Z. mays (Tsa1;GRMZM5G841619 and IgI1;GRMZM2G046191, responsible for the production of tryptophan and volatile indole, respectively) formed independent co-expression 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 co-expression 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-hour 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 co-expression modules and are typically not co-expressed

To examine whether putative BGCs (i.e., predicted based on physical clustering and with no known associated products) show evidence of co-regulation in response to specific environmental conditions, we investigated whether they were also recovered in our co-expression network analysis. We found that two different sets of putative BGCs showed little to no co-expression (Figure 9 and 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 9,670 and 10,890, respectively. Furthermore, the EC-based BGCs’ distribution of co-expression was similar to that of the control distribution of neighboring genes (P = 0.187, Wilcoxon rank sum test), whereas the co-expression of antiSMASH BGCs was significantly lower than that of the control (P = 0.027) (Figure 9A and Supplemental Data Set 10). In contrast, the six validated BGCs had a median MR score of 17.4 and were significantly more co-expressed than the control (P = 3.20 × 10^{-4}) (Figure 9A and Supplemental Data Set 10). Similarly, the 13 terpene synthase-cytochrome P450 (TS-CYP) pairs identified by Boutanaev et al. (2015) were variably co-expressed 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 × 10^{-4}) (Figure 9A and Supplemental Data Set 10).

Not surprisingly, given the lack of co-expression, putative BGCs, by and large, did not form co-expression modules, with only 7 / 188 putative BGCs overlapping by three genes or more with co-expression 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 co-expressed in a module (Supplemental Figure 10). Moreover, 78 / 188 putative BGCs overlapped with co-expression modules by only one gene, indicating that the genes within these BGCs
were more strongly co-expressed with genes outside their cluster than with those inside (Figure 9B and Supplemental Data Set 8).

An example of the poor association between co-expression modules and putative BGCs comes from the antiSMASH-predicted BGC30. Only 2 / 6 genes in BGC30 showed strong pairwise co-expression: 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 co-expression module consisting of 46 physically unlinked genes that are significantly enriched for gene ontology terms associated with flower development (Figure 9D and Supplemental Data Set 11). A second example comes from the EC-mapped BGC130. None of the genes in this BGC were strongly co-expressed 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 co-expression modules in our analysis (Figure 3 and Supplemental Data Set 8).

DISCUSSION

An enormous number of novel plant SMs await discovery and characterization (Wurtzel and Kutchan, 2016). Yet, due to their rapid evolution and narrow taxonomic distribution (Pichersky and Lewinsohn, 2011; Chae et al., 2014; Mukherjee et al., 2016), SM pathways and genes are often unknown, slowing the pace of discovery. Gene co-expression 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 co-expression datasets—each a meta-analysis of 172 to 15,275 transcriptome experiments—across eight plant model organisms, we found that gene co-expression 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 co-expression 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 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 co-expressed). Simply put, gene co-expression 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 co-expression networks constructed from genome-wide expression studies across myriads of different conditions in *A. thaliana* 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 Set 5 and Supplemental Data Set 8), including the six known to form BGCs (Figures 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 *A. thaliana* (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 non-debatable, 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 co-expressed (Figure 9).

We interpret this absence of co-expression 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 co-expression 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 co-expression dataset (Supplemental Data Set 1), as well as the recovery of many known SM pathways from these organisms (Supplemental Data Set 5 and Supplemental Data Set 8), suggest that its effect is unlikely to influence our major findings. Going forward, increased resolution of BGCs and SM pathways in co-expression 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; Ralph et al., 2006; Tzin et al., 2017) and complex field conditions (Richards et al., 2012).

Another caveat associated with predicting SM pathways from global co-expression 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 \textit{A. thaliana} were almost always combined in the same co-expression module, regardless of the network investigated, and the same was true for the two diterpenoid BGCs in \textit{O. sativa} (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 *S. lycopersicum* (*GAME2*) and *Z. mays* (*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 co-expression 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 co-expression modules in global co-expression networks no matter how many conditions are sampled.

In summary, our results indicate that generating and constructing global gene co-expression networks is a powerful and promising approach to the challenge of high-throughput prediction and study of plant SM pathways. Global gene co-expression 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 co-expression networks could include identification of new genes associated with known SM pathways (e.g., glucosinolate and benzoxazinoid pathways). Furthermore, uncharacterized co-expression 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

Co-expression network analysis

Genome annotations and protein sequences were downloaded from NCBI RefSeq and JGI Genome Portal databases (Supplemental Data Set 1). Condition-independent gene co-expression values, measured using Pearson’s correlation coefficient (PCC) and mutual rank (MR), across the eight plant species were downloaded from the ATTED-II (Aoki et al., 2016b), ALCOdb (Aoki et al., 2016a), and datasets with less than 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,

\[ MR_{AB} = \sqrt{\text{Rank}_{A \rightarrow B} \times \text{Rank}_{B \rightarrow A}} \]

where \( \text{Rank}_{A \rightarrow B} \) is the rank of gene B in a PCC-ordered list of gene A against all other genes in the microarray or RNAseq meta-analysis; similarly, \( \text{Rank}_{B \rightarrow A} \) is the rank of gene A in a PCC-ordered list of gene B against all other genes, with smaller MR scores indicating stronger co-expression between gene pairs (Obayashi and Kinoshita, 2009). MR scores were converted to network edge weights using 5 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 datasets. For example, PCC-based networks were more sensitive (variable) to differences in the number of experimental samples and genome coverage between datasets in the two species that had microarray- and RNAseq-based datasets (\textit{A. thaliana} and \textit{O. sativa}). In contrast, the MR-based networks were more robust to dataset differences (Supplemental Figure 2), in agreement with the original description of the MR metric by Obayashi and Kinoshita (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 dataset, 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 co-expressed 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 2 or more non-homologous 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 (hmmssearch using default inclusion thresholds; Eddy, 2011) (Supplemental Data Set 6). SM-like modules were then binned into meta-modules of non-overlapping gene sets. Co-expression 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 *A. thaliana* 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 co-expression 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), 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 less than 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 $Z. \text{mays}$ B73 inbred line were taken from two earlier investigations of corn 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 Set 7 and Supplemental Data Set 9.
PCCs and MRs used in this work are available for download at ATTED-II http://atted.jp/download.shtml.
Co-expression 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 $A. \text{thaliana}$.
Supplemental Figure 2. Comparison of Mutual Rank-based and Pearson’s correlation-based networks.
Supplemental Figure 3. Overlapping co-expressed modules recover the pathway for metGSL biosynthesis in $A. \text{thaliana}$.
Supplemental Figure 4. Comparison of degree of gene co-expression 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 co-expression modules involved in thalianol and marneral triterpenoid biosynthesis in A. thaliana.

Supplemental Figure 8. Network map of co-expression module involved in momilactone and phytocassane diterpenoid biosynthesis in O. sativa.

Supplemental Figure 9. Network maps of co-expression modules involved in tomatine biosynthesis in S. lycopersicum.

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

Supplemental Data Set 1. Downloaded datasets.

Supplemental Data Set 2. Descriptive statistics for co-expression networks.

Supplemental Data Set 3. A. thaliana 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 co-expression modules.

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

Supplemental Data Set 7. metGSL biosynthesis genes in A. thaliana and B. rapa.

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

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

Supplemental Data Set 10. Average co-expression of gene modules, characterized BGCs, and putative BGCs.

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


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

Acknowledgements
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 JHW, DEB-1442113 to AR, and IOS-1339237 to GJ.

Author Contributions
Conceived and designed the experiments: JHW VT GJ DJK AR. Performed the experiments: JHW ATB VT. Analyzed the data: JHW ATB VT GJ DJK AR. Contributed reagents/materials/analysis tools: JHW VT GJ AR. Wrote the paper: JHW AR. All authors read, commented on, and approved the manuscript.

References


dent assembly of operon-like gene clusters in different plants. Science

320: 543–547.


Khalid, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and


Ralph, S.G., Yueh, H., and Friedmann, M. (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


**FIGURE LEGENDS**

**Figure 1.** Co-expression network pipeline. A method for identifying small, overlapping modules of co-expressed genes in global co-expression networks. A. Calculate the Pearson’s correlation of co-expression (PCC) for every gene pair in the genome (e.g., the correlation between Gene A and B is 0.83). B. Rank correlations and calculate the mutual rank (MR) for every gene pair (e.g., the MR of Gene 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 co-expressed genes using ClusterONE. In this example, Gene 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).

**Figure 2.** Global co-expression network analysis of eight plant genomes identifies co-expressed modules of specialized metabolic genes. A) MetaCyc pathway enrichment analysis of experimentally characterized *A. thaliana* genes assigned to modules (orange bars) relative to those that do not form modules (grey bars) in *A. thaliana* microarray-based network N1. Grey 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, ***P ≤ 0.0005 (Benjamini & Hochberg adjusted P-values, hypergeometric test). See Supplemental Figure 1 for enrichment tests in other *A. thaliana* networks. B) Count
SM-like meta-modules identified in 10 microarray (m) and RNAseq (r) co-expression datasets from eight Viridiplantae. SM-like modules were collapsed into meta-modules of non-overlapping 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.

Figure 3. Co-expression modules efficiently recover the majority of genes for metGSL biosynthesis in A. thaliana. A) Network map of an example co-expression 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, right. Other co-expressed 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 co-expression dataset are crossed out. B) Heatmap depicting the correlation of co-expression of a second example co-expression module involved in metGSL biosynthesis. Diagonal numbers within the heatmap 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 co-located; scale bar is in kilobase pairs). Note: data from the RNAseq-based networks are shown because two metGSL genes (SOT17 and SOT18) are not present in the microarray dataset. Microarray-based networks performed similarly (Supplemental Data Set 8).

Figure 4. Co-expression modules predict functional metGSL biosynthesis genes in B. rapa. Network map of an example co-expression 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 *A. thaliana* as listed on brassicadb.org. Other co-expressed
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. *A. thaliana* 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. Co-expression patterns of six functionally characterized BGCs in plants.**
Heatmaps depict the correlation of co-expression of six BGCs for the production of A) marneral, B) thalianol, C) momilactone, D) phytocassane, E) tomatine, and F) DIMBOA. Diagonal numbers indicate MR scores; squares are blank if MR ≥ 100. BGC genes are bolded in the heatmap 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.

**Figure 6. Network maps of co-expression 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 *A. thaliana* (A), phytocassane and momilactone biosynthesis in *O. sativa* (B), and tomatine biosynthesis in *S. lycopersicum* (C). Gene nodes are colored according to their known or putative function. Note: the cellulose synthase gene (NCBI gene ID: 101255510) in subpanel C is located adjacent to the tomatine BGC (see Figure 5E).

**Figure 7. Pathway diagram and network map of benzoxazinoid biosynthesis in Z. mays.** A) Diagram of benzoxazinoid biosynthesis and associated pathways in *Z. mays*. Of the three indole-3-glycerol phosphate synthases*, only GRMZM2G106950 is significantly co-expressed 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.

**Figure 8. Gene expression response to insect feeding in Z. mays. A)**
Glucosyltransferase response to insect feeding. B) IGP synthase response to insect feeding. Boxplots depict fold change relative to uninfested (0 h) controls (n=5). Red boxplots are significantly different from control (P < 0.05, Student’s t test; median fold change > 2).

**Figure 9. The genes comprising the majority of bioinformatically predicted BGCs are not co-expressed.** A) Comparison of average co-expression 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 (Inter-Quartile 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 ≤ 0.05 (Wilcoxon rank sum tests). B) From top to bottom, histogram of maximum overlap between co-expression modules and known (characterized) BGCs, TS-CYP gene pairs, EC-mapped BGCs, and antiSMASH BGCs. C) Heatmap depicting the correlation of co-expression for a eight-gene region of chromosome five in A. thaliana containing an example antiSMASH BGC (BGC30) and TS-CYP gene pair (PAIR6). Diagonal number indicates MR score; squares are blank if MR ≥ 100. Heatmap scale is the same as in part a. BGC genes are bolded in the heatmap 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.
Figure 1. Co-expression network pipeline. A method for identifying small, overlapping modules of co-expressed genes in global co-expression networks. A. Calculate the Pearson’s correlation (PCC) of co-expression for every gene pair in the genome (e.g., the correlation between Gene A and B is 0.83). B. Rank correlations and calculate the mutual rank (MR) for every gene pair (e.g., the MR of Gene 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 co-expressed genes using ClusterONE. In this example, Gene 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).
Figure 2. Global co-expression network analysis of eight plant genomes identifies co-expressed modules of specialized metabolic genes. A) MetaCyc pathway enrichment analysis of experimentally characterized *A. thaliana* genes assigned to modules (orange bars) relative to those that do not form modules (grey bars) in *A. thaliana* microarray-based network N1. Grey 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, ***P ≤ 0.0005 (Benjamini & Hochberg adjusted P-values, hypergeometric test). See Figure S1 for enrichment tests in other *A. thaliana* networks. B) Count of SM-like meta-modules identified in 10 microarray (m) and RNAseq (r) co-expression datasets from eight Viridiplantae. SM-like modules were collapsed into meta-modules of non-overlapping 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 Figure S2 for meta-module counts for all other networks.
Figure 3. Co-expression modules efficiently recover the majority of genes for metGSL biosynthesis in *A. thaliana*.  
A) Network map of an example co-expression 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, right. Other co-expressed 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 co-expression dataset are crossed out. B) Heatmap depicting the correlation of co-expression of a second example co-expression module involved in metGSL biosynthesis. Diagonal numbers within the heatmap 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 co-located; scale bar is in kilobase pairs). Note: data from the RNAseq-based networks are shown because two metGSL genes (*SOT17* and *SOT18*) are not present in the microarray dataset. Microarray-based networks performed similarly (Table S8).
Figure 4. Co-expression modules predict functional metGSL biosynthesis genes in *B. rapa*. Network map of an example co-expression 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 *A. thaliana* as listed on brassicadb.org. Other co-expressed 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. *A. thaliana* metGSL genes with no known ortholog in *B. rapa* are crossed out. See Table S7 for associated NCBI and Ensembl gene IDs.
Figure 5. Co-expression pattern of six functionally characterized BGCs in plants. Heatmaps depict the correlation of co-expression of six BGCs for the production of A) marneral, B) thalianol, C) momilactone, D) phytocassane, E) tomatine, and F) DIMBOA. Diagonal numbers indicate MR scores; squares are blank if MR ≥ 100. BGC genes are bolded in the heatmap 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.
Figure 6. Network maps of co-expression 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 *A. thaliana* (A), phytocassane and momilactone biosynthesis in *O. sativa* (B), and tomatine biosynthesis in *S. lycopersicum* (C). Gene nodes are colored according to their known or putative function. Note: the cellulose synthase gene (NCBI gene ID: 101255510) in subpanel C is located adjacent to the tomatine BGC (see Figure 5E).
Figure 7. Pathway diagram and network map of benzoxazinoid biosynthesis in *Z. mays*. A) Diagram of benzoxazinoid biosynthesis and associated pathways in *Z. mays*. Of the three indole-3-glycerol phosphate synthases*, only GRMZM2G106950 is significantly co-expressed 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.
Figure 8. Gene expression response to insect feeding in *Z. mays*. A) Glucosyltransferase response to insect feeding. B) IGP synthase response to insect feeding. Boxplots depict fold change relative to uninfested (0h) controls (n=5). Red boxplots are significantly different from control (P < 0.05, Student’s t test; median fold change > 2).
Figure 9. The genes comprising the majority of bioinformatically predicted BGCs are not co-expressed. A) Comparison of average co-expression 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 (Inter-Quartile 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 ≤ 0.05 (Wilcoxon rank sum tests). B) From top to bottom, histogram of maximum overlap between co-expression modules and known (characterized) BGCs, TS-CYP gene pairs, EC-mapped BGCs, and antiSMASH BGCs. C) Heatmap depicting the correlation of co-expression for an eight-gene region of chromosome five in A. thaliana containing an example antiSMASH BGC (BGC30) and TS-CYP gene pair (PAIR6). Diagonal number indicates MR score; squares are blank if MR ≥ 100. Heatmap scale is the same as in part a. BGC genes are bolded in the heatmap 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.
A Global Co-expression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants
Jennifer H. Wisecaver, Alexander T. Borowsky, Vered Tzin, Georg Jander, Daniel J. Kliebenstein and Antonis Rokas

Plant Cell; originally published online April 13, 2017;
DOI 10.1105/tpc.17.00009

This information is current as of April 13, 2017