A global co-expression network approach for connecting genes to specialized metabolic pathways in plants

Jennifer H. Wisecaver[1], Alexander T. Borowsky[1], Vered Tzin[2], Georg Jander[3], Daniel J. Kliebenstein[4], and Antonis Rokas[1]

[1]*Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA*

[2]*French Associates Institute for Agriculture and Biotechnology of Drylands, Jacob Blaustein Institute for Desert Research, Ben Gurion University, Sede-Boqer Campus 84990, Israel*

[3]*Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY 14853, USA*

[4]*Department of Plant Sciences, University of California-Davis, One Shields Ave, Davis, CA 95616, USA*

Corresponding author: Antonis Rokas (antonis.rokas@vanderbilt.edu, Tel: +1-615-936-3892, Fax: +1-615-343-6707)

Keywords: secondary metabolism, plant natural product, chemodiversity, glucosinolates, benzoxazinoids, terpenoids, steroidal alkaloids

Abstract

Plants produce a tremendous diversity of specialized metabolites (SMs) to interact with and manage their environment. A major challenge hindering efforts to tap this seemingly boundless source of pharmacopeia is the identification of SM pathways and their constituent genes. Given the well-established observation that the genes comprising a SM pathway are 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, facilitating their identification. To evaluate this hypothesis, we used 10 global co-expression datasets—each a meta-analysis of hundreds to thousands of expression experiments—across eight plant model organisms to identify hundreds of modules of co-expressed genes for each species. In support of our hypothesis, 15.3-52.6% of modules contained two or more known SM biosynthetic genes (e.g., cytochrome P450s, terpene synthases, and chalcone synthases), and module genes were enriched in SM functions (e.g., glucoside and flavonoid biosynthesis). Moreover, modules recovered many experimentally validated SM pathways in these plants, including all six known to form biosynthetic gene clusters (BGCs). In contrast, genes predicted based on physical proximity on a chromosome to form plant BGCs were no more co-expressed than the null distribution for neighboring genes. These results not only suggest that most predicted plant BGCs do not represent genuine SM pathways but also argue that BGCs are unlikely to be a hallmark of plant specialized metabolism. We submit that global gene co-expression is a rich, but largely untapped, data source for discovering the genetic basis and architecture of plant natural products, which can be applied even without knowledge of the genome sequence.

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[1]. 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[2]. Yet the genes responsible for the production and regulation of most SMs across the kingdom Plantae are unknown, which ultimately limits their potential utility in agricultural, pharmaceutical, and biotechnological applications[3,4].

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[5]. There are two explanations for why this is so; first, SM pathways are highly variable in the number and functions of genes they contain[1,6]. Second, consistent with their involvement in the production of ecologically specialized bioactive molecules, SM genes exhibit narrow taxonomic distributions, are fast evolving both in terms of sequence divergence and rate of gene family diversification, and display extensive functional divergence[7-9]. The consequence of this lack of evolutionary and functional conservation

is that traditional sequence homology metrics for inferring gene function[10] 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[1,11,12]. 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[13-21]. 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[22-24]. 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[25-27].

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)[28]. 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[29-33]. As the proximity of genes on chromosomes is far easier to measure than their co-expression

92 across multiple experimental conditions, bioinformatic algorithms strongly rely this
93 "clustering" of genes to predict SM pathways in microbial genomes[34-36]. Thus, thousands
94 of microbial BGCs have been predicted and hundreds validated (i.e., connected to known
95 products), suggesting that gene proximity is informative for SM pathway identification,
96 at least in these organisms[37]. Nevertheless, the number of SM pathways in bacteria and
97 fungi that do not (or only partially) form BGCs is unknown[38-40].
98
99 In plants, most characterized SM pathways (e.g., glucosinolate biosynthesis) are
100 not clustered, and their genes are distributed across the genome[41]. More recently,
101 however, nearly two dozen BGCs responsible for the production of SM defensive
102 compounds have been identified and functionally characterized from 15 plant species[42],
103 raising the possibility that gene proximity could also be used for predicting plant SM
104 pathways[43]. To this end, computational searches based on gene clustering similar to those
105 developed for fungal and bacterial genomes postulate the existence of dozens to hundreds
106 more BGCs across a wide variety of plant genomes[8,44,45]. However, the vast majority of
107 these putative plant BGCs has not been functionally validated, and the fraction of plant
108 SM pathways that form BGCs is unclear.
109
110 We hypothesized that plant SM pathways are co-expressed, independently of
111 being organized into BGCs, in line with their ecological roles that require strong temporal
112 and spatial co-regulation[1,11,12]. To test our hypothesis, we developed a gene co-expression
113 network-based approach for plant SM pathway discovery (Figure S1) using data from 10
114 meta-analyses of global co-expression that collectively contain 21,876 microarray or

RNA-Seq experiments across eight plant species. Doing so, 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[25] and ALCOdb[46] for eight Viridiplantae species (*Arabidopsis thaliana*, *Brassica rapa*, *Chlamydomonas reinhardtii*, *Glycine max*, *Oryza sativa* Japonica group, *Populus trichocarpa*, *Solanum lycopersicum*, and *Zea mays*; Table S1). 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[46] to 15,275 in the *A. thaliana* microarray-based set[25]. Pairwise measurements of gene co-expression were specified as Mutual Ranks[47] (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). 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. 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[48], which allowed genes to belong to multiple modules. 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[49,50]). 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 (Table S2). 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 (Table S2). 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[51] pathways with an experimentally validated metabolic function (Table S3). 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 1a). This pattern held true across all networks and datasets investigated (Figure S2 and Table S4). 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 (Table S5).

The AMINO ACIDS, CARBOHYDRATES, and COFACTORS/PROSTHETIC GROUPS/ELECTRON CARRIERS biosynthesis classes were significantly depleted in module

184 genes in some, but not all, networks and datasets ($P < 0.05$, hypergeometric test) (Figure
185 1a and Figure S2). None of the other metabolic classes displayed any significant variation
186 between module and non-module genes (Figure S2 and Table S4).
187
188 To estimate the number of modules that may correspond to SM pathways, we
189 focused on those that contained two or more non-homologous genes with a significant
190 match to a curated list of PFAM domains that are found commonly in genes from SM
191 pathways (Table S6); as some of these "SM-like" modules share genes, we collapsed
192 them into non-intersecting "meta-modules". Dozens of SM-like meta-modules were
193 identified in each species, with the green alga, *C. reinhardtii*, containing the fewest SM-
194 like meta-modules (27 in N1 networks, 17 in N3 networks), and the field mustard, *B.*
195 *rapa*, containing the most (120 in N1 networks, 71 in N3 networks) (Figure 1b and Table
196 S2).
197
198 ***Recovery of the aliphatic glucosinolate biosynthesis pathways in* Arabidopsis *and***
199 **Brassica *from global co-expression data.*** To illustrate the utility and power of our
200 approach for identifying entire SM pathways, we next focused on examining the
201 correspondence between genes involved in the methionine-derived aliphatic glucosinolate
202 (metGSL) biosynthesis pathway and genes that comprise co-expression modules
203 identified by our analyses (Table S7). In *A. thaliana*, the species with the majority of
204 functional data[52], co-expression modules recover genes for every biochemical step in this
205 pathway, from methionine chain elongation to side-chain modification of the
206 glucosinolate chemical backbone, as well as a pathway-specific transporter and three

transcription factors (Figure 2a). 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 2b). 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 (Figure S3 and Table S8). 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 2a). 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[53]. 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)[52].

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*[54]. 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-

230 expressed with metGSL biosynthetic genes (Figure 2). Importantly, *DJ1F* is not
231 represented on the *A. thaliana* Affymetrix GeneChip, explaining why *GGP1* and not this
232 gene was identified as the most correlated one in earlier analyses. However, the
233 postulated role of both *DJ1F* and *GGP1* in metGSL biosynthesis remains to be confirmed
234 *in planta*.
235
236 The remaining genes in the metGSL pathway that were never recovered in co-
237 expressed modules all encode secondary enzymes responsible for terminal modifications
238 to the backbone glucosinolate product[41]. One of these, *AOP2*, encoding a 2-oxoglutarate-
239 dependent dioxygenase, has been pseudogenized in the *A. thaliana* (ecotype: Columbia)
240 reference genome[55]. The high level of natural variation present in these terminal
241 metabolic branches is responsible for the diverse glucosinolates present in different
242 ecotypes[56,57] but likely also makes it more challenging to connect them to the rest of the
243 metGSL pathway using global co-expression data (Figure S4).
244
245 Brassicas also produce aliphatic glucosinolates, but a whole genome triplication
246 event subsequent to their divergence from *A. thaliana*[58] has complicated identification of
247 functional metGSL genes in these species. To gain insight into the metGSL pathway in *B.*
248 *rapa*, we cross-referenced our co-expression modules with 59 candidate metGSL genes
249 identified based on orthology to *A. thaliana* metGSL genes[59]. As in *A. thaliana*, modules
250 recovered every biochemical step of the *B. rapa* metGSL pathway as well as pathway-
251 specific transporters and transcription factors (Figure 3, Table S7, and Table S8). Also as
252 in *A. thaliana*, *DJ1F* rather than *GGP1* is co-expressed with other metGSL genes,

253 providing further evidence that the DJ1F enzyme may be the more likely candidate for
254 the γ-glutamyl peptidase activity in glucosinolate biosynthesis[52]. Furthermore, as several
255 enzymes are encoded by multiple gene copies in *B. rapa*, we harnessed the power of our
256 module analysis to identify which of these copies was co-expressed with other metGSL
257 genes and therefore most likely to be functionally involved in the pathway. For example,
258 out of the six *MAM* gene copies in *B. rapa*, only *Bra029355* and *Bra013007* were
259 recovered in metGSL modules (Figure 3 and Figure S5). Module data also suggest that
260 the glutathione S-transferase class tau (GSTU) activity is one step of the core pathway
261 that may differ between the two species. Specifically, in *A. thaliana*, *GSTU20* is thought
262 to encode this reaction, and this gene was recovered in metGSL modules in our analysis
263 (Figure 2a). However, this association was not recovered in *B. rapa.* Instead, three
264 paralogous GSTUs (*Bra003647*, *Bra026679*, and *Bra026680*), corresponding to the *A.*
265 *thaliana GSTU23* and *GSTU25* genes, respectively, formed modules with metGSL genes,
266 making these genes good candidates for investigation of GSTU activity in *B. rapa*
267 (Figure 3 and Figure S6).
268
269 ***Modules recover functionally characterized BGCs and identify associated, unclustered***
270 ***genes.*** We next investigated whether our approach also recovered BGCs by examining
271 whether our co-expression modules recovered the six functionally characterized BGCs in
272 these eight plant genomes (Table S9). All six BGCs were recovered in our module
273 analysis (Table S8). Specifically, co-expression modules recovered all genes comprising
274 the BGCs involved in the production of the triterpenoids marneral[60] (3 / 3 genes; Figure
275 4a) and thaliaol[61] (4 / 4 genes; Figure 4b) in *A. thaliana* and the diterpenoid

momilactone[62] (5 / 5 genes; Figure 4c) in *O. sativa*. Modules recovered 7 / 9 genes in the phytocassane[63] 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 4d). The two triterpenoid BGCs in *A. thaliana* were typically combined into the same co-expression module; the same pattern was observed for the two diterpenoid BGCs in *O. sativa* (Figure S7 and Figure S8). 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 S7 and Figure S8).

Seven of eight genes in the partially clustered pathway for production of the steroidal alkaloid α-tomatine in *S. lycopersicum*[64] (tomato) were recovered by our co-expression analysis (Figure 4e). 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[64,65]. Several glucosyltransferase genes paralogous to *GAME2* were strongly co-expressed with the rest of the genes in this pathway (Figure S9), 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, several possible metabolite transporters, and a cellulose synthase-like gene located adjacent to the BGC (Figure 4e and Figure S9).

298 Lastly, five of the six genes in the benzoxazinoid 2,4-dihydroxy-7-methoxy-1,4-
299 benzoxazin-3-one (DIMBOA)[66] cluster in *Z. mays* formed co-expression modules in our
300 analysis (Figure 4f). Specifically, the first five genes in the DIMBOA pathway (*Bx1-*
301 *Bx5*), responsible for the biosynthesis of the precursor 2,4-dihydroxy-1, 4-benzoxazin-3-
302 one (DIBOA), formed modules with each other but not with the final gene in the BGC,
303 *Bx8* (Figure S10).
304
305 Similar to the modifying genes of the metGSL pathway in *A. thaliana*, terminal
306 *Bx* genes appear to have unique gene expression signatures distinct from the core
307 pathway. For example, DIBOA is modified to DIMBOA by the action of two additional
308 unclustered genes (*Bx6* and *Bx7*)[67], neither of which was assigned to modules with core
309 genes or each other. Toxic DIBOA/DIMBOA is transformed into stable glucoside,
310 DIBOA-Glc/DIMBOA-Glc, by glucosyltransferases (*Bx8* and *Bx9*), which were likewise
311 not assigned to modules in our analysis. However, a gene adjacent to the DIMBOA BGC,
312 encoding an uncharacterized glucosyltransferase (GT; *GRMZM2G085854*) with 27%
313 amino acid identity to *Bx8*, does belong to the same module as the core *Bx* genes in
314 network N3 (Figure S10), but the MR scores of this gene to core *Bx* genes are noticeably
315 weaker than those between the core *Bx* genes (Figure 4f). Additional *Bx* genes (*Bx10-*
316 *Bx14*), which are not part of the BGC and are responsible for the biosynthesis of modified
317 benzoxazinoid compounds (e.g., HDMBOA-Glc, DIM_2BOA-Glc)[68,69], were also not
318 assigned to modules in our analysis (Figure S10); this pattern is similar to that observed
319 with the terminal reactions of the metGSL biosynthesis pathway.
320

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 (IGPS; *GRMZM2G106950*), which catalyzes the reaction directly upstream of *Bx1* (Figure S10). 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* (*TSA* and *IGL*, responsible for the production of tryptophan and volatile indole, respectively) formed independent co-expression modules, consistent with their distinct metabolic and ecological roles (Figure S10)[70,71]. 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 measured their gene expression responses to two different types of insect herbivory (aphid and caterpillar), ecological conditions under which benzoxazinoid biosynthesis genes are typically induced[72,73]. *GT* showed gene expression responses similar to *Bx8* and *Bx9*, being induced within the first few hours after the introduction of insect herbivores (Figure S11). 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 S11). 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[74].

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 S12). Specifically, both the 137 Enzyme Commission (EC)-based BGCs predicted by Chae et al.[8] and the 51 BGCs predicted by the antibiotics and secondary metabolism analysis shell (antiSMASH)[34] 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 5a and Table S10). In contrast, the six validated BGCs had a median MR score of 17.4 and were significantly more co-expressed than the control ($P = 3.20 \times 10^{-4}$) (Figure 5a and Table S10). Similarly, the 13 terpene synthase-cytochrome P450 (TS-CYP) pairs identified by Boutanaev et al.[44] 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 ($P = 2.73 \times 10^{-4}$) (Figure 5a and Table S10).

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. In contrast, 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 5b and Table S8).

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 5c). The terpene synthase (AT5G44630) of PAIR6 is known to be involved in the production of sesquiterpenoid flower volatiles[75]. 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 5d and Table S11). A second example comes from the EC-mapped BGC130. None of the genes in this BGC were strongly co-expressed with each other (Figure S12). 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 2 and Table S8).

Discussion

An enormous number of novel plant SMs awaits discovery and characterization[4]. Yet, due to their rapid evolution and narrow taxonomic distribution[7-9], 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[4], 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[1,11,12]. 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) [21,32,33,64,65,76,77]. 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

413 gene expression data to predict SM pathways in a high-throughput fashion. The results of
414 our analyses suggest that this is the case; modules in global co-expression networks
415 constructed from genome-wide expression studies across myriads of different conditions
416 in *A. thaliana* were significantly enriched in genes associated with diverse SM-related
417 metabolic functions (Figure 1a). Moreover, modules recovered many experimentally
418 validated SM pathways in these plants (Table S5 and Table S8), including the six known
419 to form BGCs (Figures 4).
420
421 It is also well established that gene arrangement in plant genomes is not random[78].
422 For example, as much as 60% of metabolic pathways in *A. thaliana* (as measured by
423 KEGG) show statistically significant higher levels of physical proximity in the genome
424 than expected by chance[79]. The most extreme version of this "closer than expected" gene
425 arrangement is the growing list BGCs involved in plant SM biosynthesis[42]. While the
426 statistical significance of this pattern is non-debatable, the degree to which gene
427 arrangement is predictive of genes' participation in the same pathway is not immediately
428 obvious. For example, the genes of many known plant SM pathways[52,80] do not form
429 BGCs, while other pathways consist of a combination of clustered and unclustered
430 genes[64,69,81]. Complicating matters further, SM pathways may form a BGC in some
431 species but not others[82]. Given that the majority of known plant SM pathways does not
432 form BGCs, it is perhaps not surprising that nearly all putative plant SM BGCs, which
433 were predicted based solely on gene proximity, were not co-expressed (Figure 5).
434

435 We interpret this absence of co-expression as evidence that most of these putative
436 BGCs likely do not correspond with actual SM pathways and that gene proximity is
437 insufficient to be used as the primary input for predicting SM pathways in plant genomes.
438 Admittedly, the strength of this argument rests on whether the global co-expression
439 networks that we have constructed accurately capture the spatial and temporal regulation
440 of BGCs in response to the diverse ecological conditions plants experience, which is at
441 least partially dependent on the number and types of the conditions sampled[83]. For
442 example, genes in a BGC or pathway that are never expressed or are not variably
443 expressed across conditions would not be correlated with each other in our analysis.
444 Although this is a valid concern, the hundreds to thousands of conditions[25] used to
445 construct each co-expression dataset (Table S1) as well as the recovery of many known
446 SM pathways from these organisms (Table S5 and Table S8), suggest that its effect is
447 unlikely to influence our major findings. Going forward, increased resolution of BGCs
448 and SM pathways in co-expression networks will require the inclusion of data from more
449 tissues, time points, and environmental conditions during which SM genes and pathways
450 are likely to vary in their regulation, for example different types of insect herbivory[69,72,84]
451 and complex field conditions[85].
452
453 Another caveat associated with predicting SM pathways from global co-
454 expression networks is that SM pathways whose expression profiles are highly similar
455 would be predicted to comprise a single pathway. This will likely be a more common
456 occurrence, and examples of this behavior are present in our results. Specifically, the two
457 triterpenoid BGCs in *A. thaliana* were almost always combined in the same co-

458 expression module, regardless of the network investigated (Figure S7); the same was true
459 for the two diterpenoid BGCs in *O. sativa* (Figure S8). Although predicting individual
460 SM pathways is obviously ideal, the lumping of multiple pathways into one may in some
461 cases reveal novel biology. For example, such a pattern could also be indicative of
462 crosstalk between SM pathways or BGCs, or that multiple SM pathways are employed in
463 response to the same set of environmental conditions.
464
465 The final caveat is that our approach will not be as powerful in cases where some
466 of the genes in the pathway are not under the same regulatory program as the others. For
467 example, we noted that the genes encoding terminal modification enzymes, such as the
468 genes for side-chain modification of glucosinolates (Figure S4) or the UDP-
469 glucosyltransferases in *S. lycopersicum* (*GAME2*) and *Z. mays* (*Bx8-Bx14*), had
470 expression profiles that were quite different from those of core pathway genes; thus, they
471 were often not recovered in the same modules as their corresponding core SM pathway
472 genes. It is possible that additional sampling of appropriate expression conditions could
473 allow for recovery of these terminal metabolic branches in co-expression modules that
474 include the rest of the pathway. However, the terminal SM genes and products can be
475 under balancing or diversifying selection[56]; moreover, the core and terminal steps in an
476 SM pathway may take place in different tissues[86]. In cases like these, the terminal
477 metabolic branches and core SM pathway may be identified as distinct co-expression
478 modules in global co-expression networks no matter how many conditions are sampled.
479

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 plant, 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. Furthermore, global gene co-expression networks could be used in conjunction with other high-throughput data types (e.g., proteomics, metabolomics). 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.

Materials and Methods

Genome annotations, protein sequences, and 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[25], ALCOdb[46], NCBI RefSeq and JGI Genome Portal databases (Table S1). ATTED-II co-expression datasets with less than 50% coverage of the target genome were excluded. 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 RNAseq 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

502 indicating stronger co-expression between gene pairs[47]. MR scores were converted to
503 network edge weights using 5 different rates of exponential decay (Figure S1). Any edge
504 with PCC < 0.3 or edge weight < 0.01 was excluded.
505
506 Comparison of MR- and PCC-based networks, showed that the MR-based
507 networks were more comparable between species and datasets. For example, PCC-based
508 networks were more sensitive (variable) to differences in the number of experimental
509 samples and genome coverage between datasets in the two species that had microarray-
510 and RNAseq-based datasets (*A. thaliana* and *O. sativa*). In contrast, the MR-based
511 networks were more robust to dataset differences (Figure S13), in agreement with the
512 original description of the MR metric by Obayashi and Kinoshita[47]. Moreover, MR-based
513 networks were remarkably consistent with respect to the number of genes they contained;
514 in contrast, PCC-based networks sometimes varied by orders of magnitude in the number
515 of genes included (Figure S13), Finally, MR-based networks consistently included nearly
516 all genes in a given dataset, regardless of the MR threshold stringency employed; that
517 was not the case with PCC-based networks (Figure S13 and Table S2). For these reasons,
518 we chose to focus the investigation on the MR-based networks.
519
520 Modules of tightly co-expressed genes were detected using ClusterONE using
521 default parameters[48]. Modules with ClusterONE *P* value > 0.1 were excluded. Modules
522 were considered 'SM-like' if they contained 2 or more non-homologous genes with a
523 significant match to a curated list of PFAM domains present in experimentally verified
524 (Evidence = 'EV-EXP') genes assigned to MetaCyc[51] SECONDARY BIOSYNTHESIS

pathways (hmmsearch[87] using default inclusion thresholds; Table S6). SM-like modules were then binned into meta-modules of non-overlapping gene sets.

Bioinformatically-predicted BGCs were obtained from the published literature[8,44] and by running the *A. thaliana* reference genome (TAIR10; each protein-coding gene was represented by its longest transcript) through antiSMASH v3.0.4[34] 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).

Data Availability

All co-expression modules identified in our analysis are included in the supplemental files online (Dataset S1).

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

548 Science Foundation (http://www.nsf.gov) under Grants IOS-1401682 to JHW, DEB-
549 1442113 to AR, and IOS-1339237 to GJ.

550

551 ## Author Contributions

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

References

1. Hartmann, T. From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* **68,** 2831–2846 (2007).
2. Raskin, I. *et al.* Plants and human health in the twenty-first century. *Trends in Biotechnology* **20,** 522–531 (2002).
3. McChesney, J. D., Venkataraman, S. K. & Henri, J. T. Plant natural products: back to the future or into extinction? *Phytochemistry* **68,** 2015–2022 (2007).
4. Wurtzel, E. T. & Kutchan, T. M. Plant metabolism, the diverse chemistry set of the future. *Science* **353,** 1232–1236 (2016).
5. De Luca, V., Salim, V., Atsumi, S. M. & Yu, F. Mining the biodiversity of plants: a revolution in the making. *Science* **336,** 1658–1661 (2012).
6. D'Auria, J. C. & Gershenzon, J. The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr Opin Plant Biol* **8,** 308–316 (2005).
7. Pichersky, E. & Lewinsohn, E. Convergent evolution in plant specialized metabolism. *Annu Rev Plant Biol* **62,** 549–566 (2011).
8. Chae, L., Kim, T., Nilo-Poyanco, R. & Rhee, S. Y. Genomic signatures of specialized metabolism in plants. *Science* **344,** 510–513 (2014).
9. Mukherjee, D., Mukherjee, A. & Ghosh, T. C. Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in *Arabidopsis thaliana*. *Genome Biol. Evol.* **8,** 17–28 (2016).
10. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8,** 163–167 (1998).
11. Tohge, T. & Fernie, A. R. Co-expression and co-responses: within and beyond transcription. *Frontiers in Plant Science* **3,** 248 (2012).
12. Grotewold, E. Plant metabolic diversity: a regulatory perspective. *Trends Plant Sci* **10,** 57–62 (2005).
13. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349,** 1224–1228 (2015).
14. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic metabolite in *Arabidopsis* required for inducible pathogen defence. **525,** 376–379 (2015).
15. Yonekura-Sakakibara, K. *et al.* Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. *Plant Cell* **20,** 2160–2176 (2008).
16. Sawada, Y. *et al. Arabidopsis* bile acid:sodium symporter family protein 5 is involved in methionine-derived glucosinolate biosynthesis. *Plant Cell Physiol.* **50,** 1579–1586 (2009).
17. Hirai, M. Y. *et al.* Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 6478–6483 (2007).
18. Maeda, H., Yoo, H. & Dudareva, N. Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. *Nat. Chem. Biol.* **7,** 19–21 (2011).
19. Naoumkina, M. A. *et al.* Genomic and coexpression analyses predict multiple genes involved in triterpene saponin biosynthesis in *Medicago truncatula*. *Plant Cell* **22,** 850–866 (2010).
20. Fridman, E. & Pichersky, E. Metabolomics, genomics, proteomics, and the

identification of enzymes and their substrates and products. *Curr Opin Plant Biol* **8,** 242–248 (2005).

21. Itkin, M. *et al.* The biosynthetic pathway of the nonsugar, high-intensity sweetener mogroside V from *Siraitia grosvenorii*. *Proc. Natl. Acad. Sci. U.S.A.* **113,** E7619–E7628 (2016).
22. Horan, K. *et al.* Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* **147,** 41–57 (2008).
23. Mentzen, W. I. & Wurtele, E. S. Regulon organization of *Arabidopsis*. *BMC Plant Biol* **8,** 99 (2008).
24. Mao, L., Van Hemert, J. L., Dash, S. & Dickerson, J. A. *Arabidopsis* gene co-expression network and its functional modules. *BMC Bioinformatics* **10,** 346 (2009).
25. Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K. & Obayashi, T. ATTED-II in 2016: A plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.* **57,** e5 (2016).
26. Hiss, M. *et al.* 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 (2014).
27. Zimmermann, P. *et al.* Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. *Mol Plant* **1,** 851–857 (2008).
28. Osbourn, A. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.* **26,** 449–457 (2010).
29. Yu, J. *et al.* Tight control of mycotoxin biosynthesis gene expression in *Aspergillus flavus* by temperature as revealed by RNA-Seq. *Fems Microbiol Lett* **322,** 145–149 (2011).
30. Lawler, K., Hammond-Kosack, K., Brazma, A. & Coulson, R. M. R. Genomic clustering and co-regulation of transcriptional networks in the pathogenic fungus *Fusarium graminearum*. *BMC Syst Biol* **7,** 52 (2013).
31. Gibbons, J. G. *et al.* Global transcriptome changes underlying colony growth in the opportunistic human pathogen *Aspergillus fumigatus*. *Eukaryot. Cell* **11,** 68–78 (2012).
32. Gibbons, J. G. *et al.* The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*. *Curr. Biol.* **22,** 1403–1409 (2012).
33. Lind, A. L., Smith, T. D., Saterlee, T., Calvo, A. M. & Rokas, A. Regulation of secondary metabolism by the Velvet complex is temperature-responsive in *Aspergillus*. *G3 (Bethesda)* **6,** 4023–4033 (2016).
34. Weber, T. *et al.* antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43,** W237–W243 (2015).
35. Khaldi, N. *et al.* SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* **47,** 736–741 (2010).
36. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158,** 412–421 (2014).
37. Hadjithomas, M. *et al.* IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* **6,** e00932–15

(2015).
38. Bradshaw, R. E. *et al.* Fragmentation of an aflatoxin-like gene cluster in a forest pathogen. *New Phytol* **198,** 525–535 (2013).
39. Sanchez, J. F. *et al.* Genome-based deletion analysis reveals the prenyl xanthone biosynthesis pathway in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **133,** 4010–4017 (2011).
40. Lo, H.-C. *et al.* Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **134,** 4709–4720 (2012).
41. Kliebenstein, D. J. & Osbourn, A. Making new molecules - evolution of pathways for novel metabolites in plants. *Curr Opin Plant Biol* **15,** 415–423 (2012).
42. Nützmann, H.-W., Huang, A. & Osbourn, A. Plant metabolic clusters - from genetics to genomics. *New Phytol* (2016). doi:10.1111/nph.13981
43. Medema, M. H. & Osbourn, A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat Prod Rep* **33,** 951–962 (2016).
44. Boutanaev, A. M. *et al.* Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E81–E88 (2015).
45. Castillo, D. A., Kolesnikova, M. D. & Matsuda, S. P. T. An effective strategy for exploring unknown metabolic pathways by genome mining. *J. Am. Chem. Soc.* **135,** 5885–5894 (2013).
46. Aoki, Y., Okamura, Y., Ohta, H., Kinoshita, K. & Obayashi, T. ALCOdb: Gene coexpression database for microalgae. *Plant Cell Physiol.* **57,** e3–e3 (2016).
47. Obayashi, T. & Kinoshita, K. Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res* **16,** 249–260 (2009).
48. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* **9,** 471–472 (2012).
49. Guo, J. *et al.* Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. *New Phytol* **210,** 525–534 (2016).
50. Lodeiro, S. *et al.* An oxidosqualene cyclase makes numerous products by diverse mechanisms: A challenge to prevailing concepts of triterpene biosynthesis. *J. Am. Chem. Soc.* **129,** 11213–11222 (2007).
51. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–D480 (2016).
52. Sønderby, I. E., Geu-Flores, F. & Halkier, B. A. Biosynthesis of glucosinolates – gene discovery and beyond. *Trends Plant Sci* **15,** 283–290 (2010).
53. Mugford, S. G. *et al.* Disruption of adenosine-5'-phosphosulfate kinase in *Arabidopsis* reduces levels of sulfated secondary metabolites. *Plant Cell* **21,** 910–927 (2009).
54. Geu-Flores, F. *et al.* Glucosinolate engineering identifies a gamma-glutamyl peptidase. *Nat. Chem. Biol.* **5,** 575–577 (2009).
55. Kliebenstein, D. J., Lambrix, V. M., Reichelt, M., Gershenzon, J. & Mitchell-Olds, T. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in

Arabidopsis. *Plant Cell* **13,** 681–693 (2001).

56. Kerwin, R. *et al.* Natural genetic variation in *Arabidopsis thaliana* defense metabolism genes modulates field fitness. *Elife* **4,** e05604 (2015).
57. Brachi, B. *et al.* Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 4032–4037 (2015).
58. Town, C. D. *et al.* Comparative genomics of *Brassica oleracea* and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell* **18,** 1348–1359 (2006).
59. Wang, H. *et al.* Glucosinolate biosynthetic genes in *Brassica rapa*. *Gene* **487,** 135–142 (2011).
60. Field, B. *et al.* Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 16116–16121 (2011).
61. Field, B. & Osbourn, A. E. Metabolic diversification—independent assembly of operon-like gene clusters in different plants. *Science* **320,** 543–547 (2008).
62. Shimura, K. *et al.* Identification of a biosynthetic gene cluster in rice for momilactones. *J Biol Chem* **282,** 34013–34018 (2007).
63. Swaminathan, S., Morrone, D., Wang, Q., Fulton, D. B. & Peters, R. J. CYP76M7 is an *ent*-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21,** 3315–3325 (2009).
64. Itkin, M. *et al.* Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341,** 175–179 (2013).
65. Cárdenas, P. D. *et al.* GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nat. Commun.* **7,** 10654 (2016).
66. Frey, M. *et al.* Analysis of a chemical plant defense mechanism in grasses. *Science* **277,** 696–699 (1997).
67. Jonczyk, R. *et al.* Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of *Bx6* and *Bx7*. *Plant Physiol* **146,** 1053–1063 (2008).
68. Meihls, L. N. *et al.* 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 (2013).
69. Handrick, V. *et al.* Biosynthesis of 8-o-methylated benzoxazinoid defense compounds in maize. *Plant Cell* **28,** 1682–1700 (2016).
70. Frey, M. *et al.* An herbivore elicitor activates the gene for indole emission in maize. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 14801–14806 (2000).
71. Frey, M., Schullehner, K., Dick, R., Fiesselmann, A. & Gierl, A. Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. *Phytochemistry* **70,** 1645–1651 (2009).
72. Tzin, V. *et al.* Dynamic maize responses to aphid feeding are revealed by a time series of transcriptomic and metabolomic assays. *Plant Physiol* **169,** 1727–1743 (2015).
73. Tzin, V. *et al. Spodoptera exigua* caterpillar feeding induces defense responses in maize leaves as revealed by transcriptome and metabolome analysis. manuscript in preparation (2016).

74. Erb, M. *et al.* Indole is an essential herbivore-induced volatile priming signal in maize. *Nat. Commun.* **6,** 6273 (2015).
75. Tholl, D., Chen, F., Petri, J., Gershenzon, J. & Pichersky, E. Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from *Arabidopsis* flowers. *Plant J.* **42,** 757–771 (2005).
76. Darbani, B. *et al.* The biosynthetic gene cluster for the cyanogenic glucoside dhurrin in *Sorghum bicolor* contains its co-expressed vacuolar MATE transporter. *Sci. Rep.* **6,** 37079 (2016).
77. Dhingra, S. *et al.* The fumagillin gene cluster, an example of hundreds of genes under veA control in *Aspergillus fumigatus*. *PLoS ONE* **8,** (2013).
78. Hurst, L., Pal, C. & Lercher, M. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5,** 299–310 (2004).
79. Lee, J. M. & Sonnhammer, E. L. L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* **13,** 875–882 (2003).
80. Winkel-Shirley, B. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol* **126,** 485–493 (2001).
81. Zhou, Y. *et al.* Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat Plants* **2,** 16183 (2016).
82. Sue, M., Nakamura, C. & Nomura, T. Dispersed benzoxazinone gene cluster: molecular characterization and chromosomal localization of glucosyltransferase and glucosidase genes in wheat and rye. *Plant Physiol* **157,** 985–997 (2011).
83. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* **31,** 2123–2130 (2015).
84. Ralph, S. G., Yueh, H. & Friedmann, M. 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 (2006).
85. Richards, C. L., Rosas, U., Banta, J., Bhambhra, N. & Purugganan, M. D. Genome-wide patterns of *Arabidopsis* gene expression in nature. *PLoS Genet.* **8,** e1002662 (2012).
86. Hartmann, T. & Ober, D. Biosynthesis and metabolism of pyrrolizidine alkaloids in plants and specialized insect herbivores in *Biosynthesis: Aromatic Polyketides, Isoprenoids, Alkaloids* **209,** 207–243 (Springer Berlin Heidelberg, 2000).
87. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol* **7,** e1002195 (2011).

Supplemental files

Dataset S1. Co-expression modules (txt).

Figure S1. Co-expression network pipeline (pdf).

Figure S2. MetaCyc pathway enrichment analysis of experimentally characterized genes in *A. thaliana* (pdf).

Figure S3. Overlapping co-expressed modules recover the pathway for metGSL biosynthesis in *A. thaliana* (pdf).

Figure S4. Comparison of degree of gene co-expression in core versus terminal modification genes in metGSL biosynthesis (pdf).

Figure S5. Maximum likelihood phylogeny of Brassicaceae MAM and IPMS sequences (pdf).

Figure S6. Maximum likelihood phylogeny of Brassicaceae GSTU sequences (pdf).

Figure S7. Network maps of co-expression modules involved in thalianol and marneral triterpenoid biosynthesis in *A. thaliana* (pdf).

Figure S8. Network map of co-expression module involved in momilactone and phytocassane diterpenoid biosynthesis in *O. sativa* (pdf).

Figure S9. Network maps of co-expression modules involved in tomatine biosynthesis in *S. lycopersicum*. (pdf).

Figure S10. Pathway diagram and network map of DIMBOA biosynthesis and related pathways in *Z. mays* (pdf).

Figure S11. Gene expression response to insect feeding in *Z. mays* (pdf).

Figure S12. Coexpression pattern of seven putative BGCs in plants (pdf).

Figure S13. Comparison of Mutual Rank-based and Pearson's correlation-based networks

(pdf).

Table S1. Downloaded datasets (xlsx).

Table S2. Descriptive statistics for co-expression networks (xlsx).

Table S3. *A. thaliana* genes assigned to MetaCyc pathways and pathway ontologies (xlsx).

Table S4. Test for enrichment/depletion of MetaCyc pathway categories and classes in module genes (xlsx).

Table S5. Recovery of MetaCyc pathways in co-expression modules (xlsx).

Table S6. List of Pfam domains found in SM pathways in MetaCyc (xlsx).

Table S7. metGSL biosynthesis genes in *A. thaliana* and *B. rapa* (xlsx).

Table S8. Recovery of metGSL pathways, characterized BGCs, and putative BGCs in co-expression modules (xlsx).

Table S9. List of functionally characterized BGCs in plants with co-expression data on ATTED-II (xlsx).

Table S10. Average co-expression of gene modules, characterized BGCs, and putative BGCs (xlsx).

Table S11. GO enrichment test of a 46-gene *A. thaliana* module involved in flower development (xlsx).

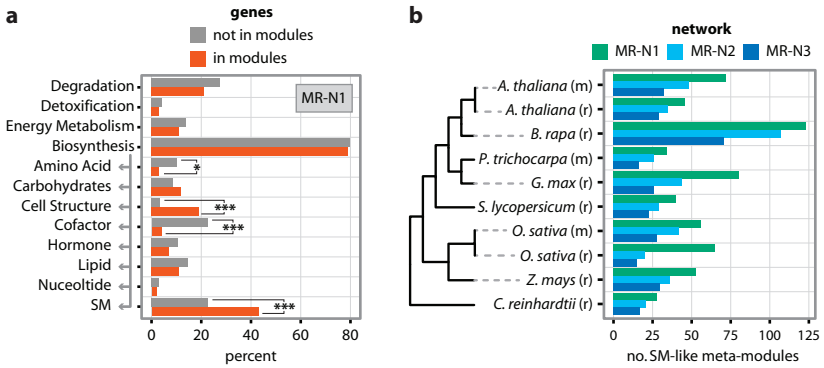

Figure 1. Global co-expression network analysis of eight plant genomes identify co-expressed modules of specialized metabolic genes. a) MetaCyc pathway enrichment analysis of experimentally characterized *A. thaliania* genes assigned to modules (orange bars) relative to those that do not form modules (grey bars) in *A. thaliania* microarray-based network MR-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 S2 for enrichment tests in other *A. thaliania* 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 MR-N1 corresponds to smallest network with the steepest rate of decay and therefore the fewest edges; conversely, MR-N3 is the largest network with the shallowest rate of decay and the most edges.

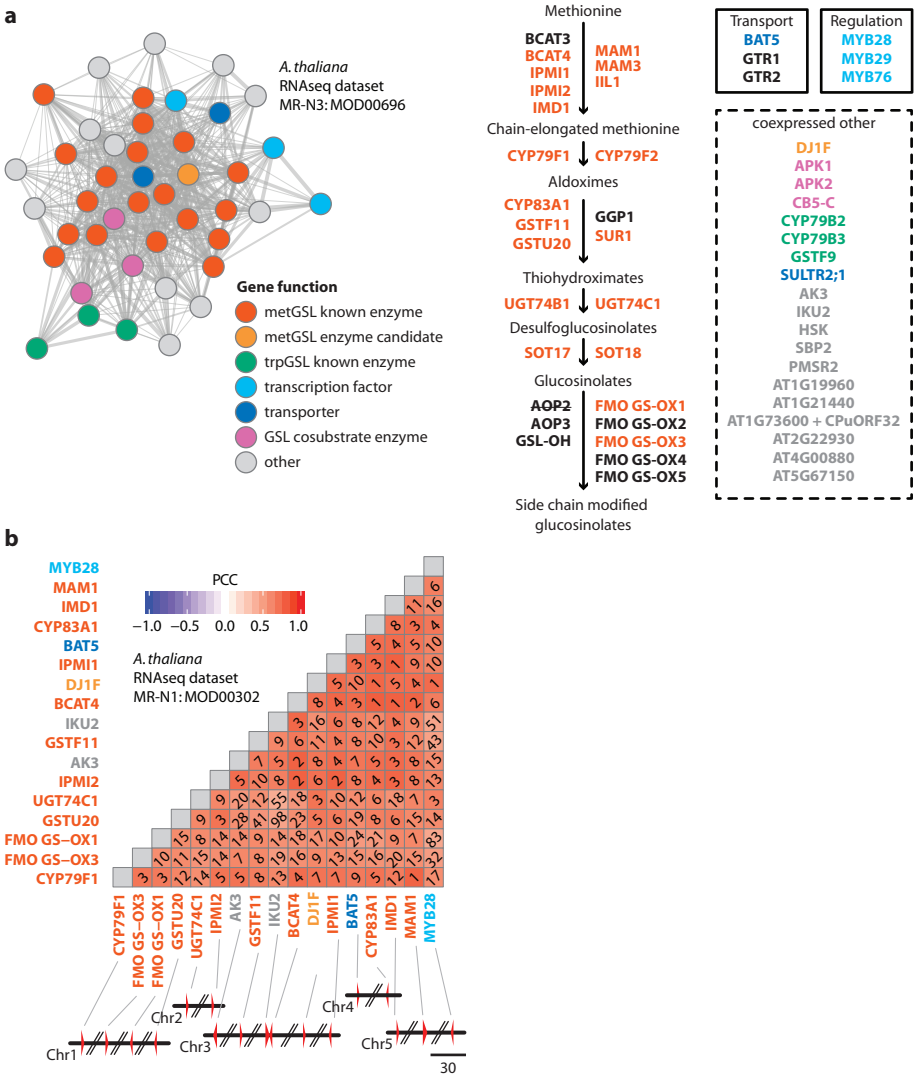

Figure 2. Co-expression modules efficiently recover the majority of genes for metGSL biosynthesis in *A. thaliana*. a) Network map of an example co-expression module involved in metGSL biosynthesis. Nodes in the map represent genes and edges connecting two genes represent the weight (transformed MR score) for the association. 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 lined 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 as two metGSL genes (*SOT17* and *SOT18*) are not present in the microarray dataset. Microarray-based networks performed similarly (Table S8).

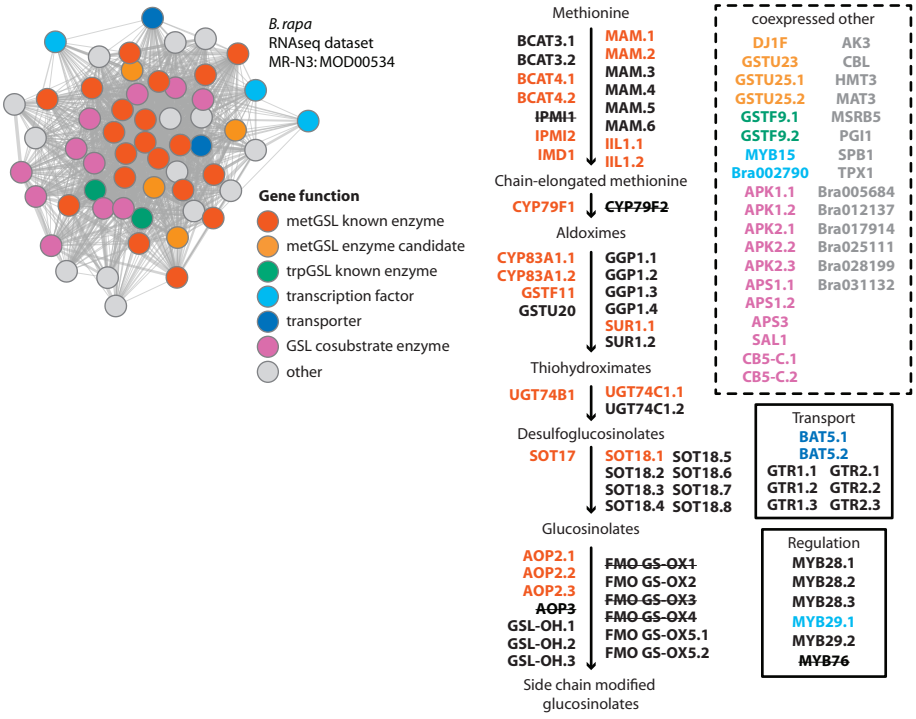

Figure 3. Co-expression modules predict functional metGSL biosynthesis genes in *B. rapa*. Network map of an example co-expression module involved in metGSL biosynthesis. Nodes in the map represent genes and edges connecting two genes represent the weight (transformed MR score) for the association. The diagram of the metGSL biosynthesis pathway is depicted, right, 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 lined out. See Table S7 for associated NCBI and Ensembl gene IDs.

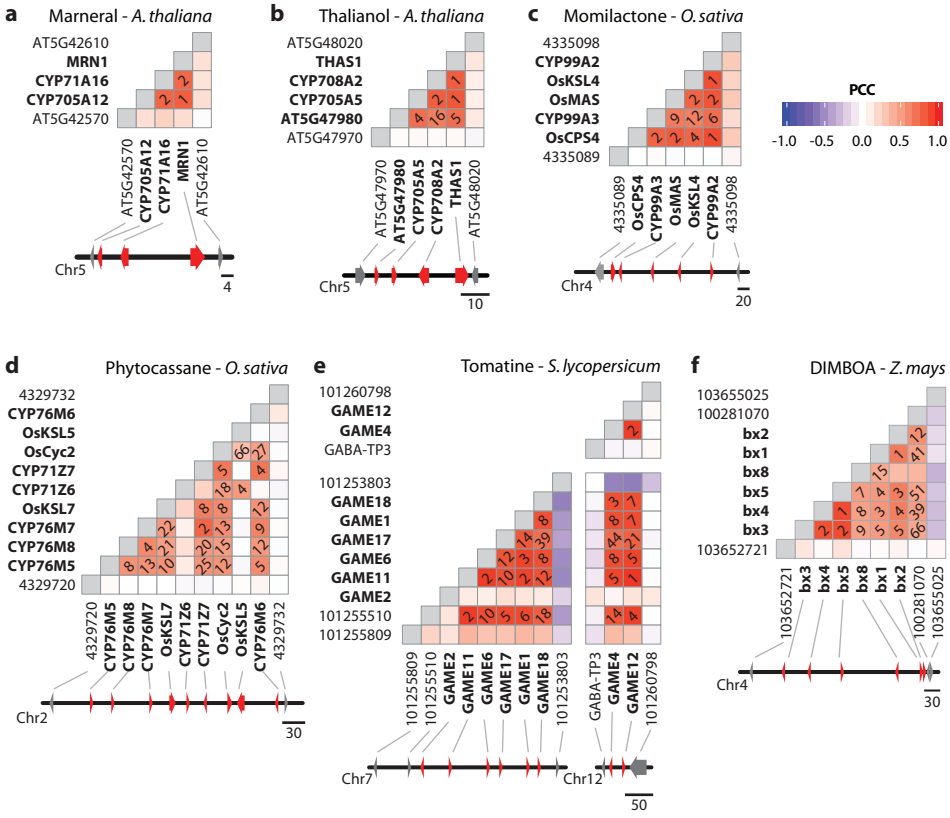

Figure 4. Co-expression pattern of six functionally characterized BGCs in plants. Heatmaps depict the correlation of co-expression of six BGCs for the production of a. marneral, b. thalianol, c. momilactone, d. phytocassane, e. tomatine, and f. DIMBOA. Diagonal numbers indicate MR scores; squares are 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.

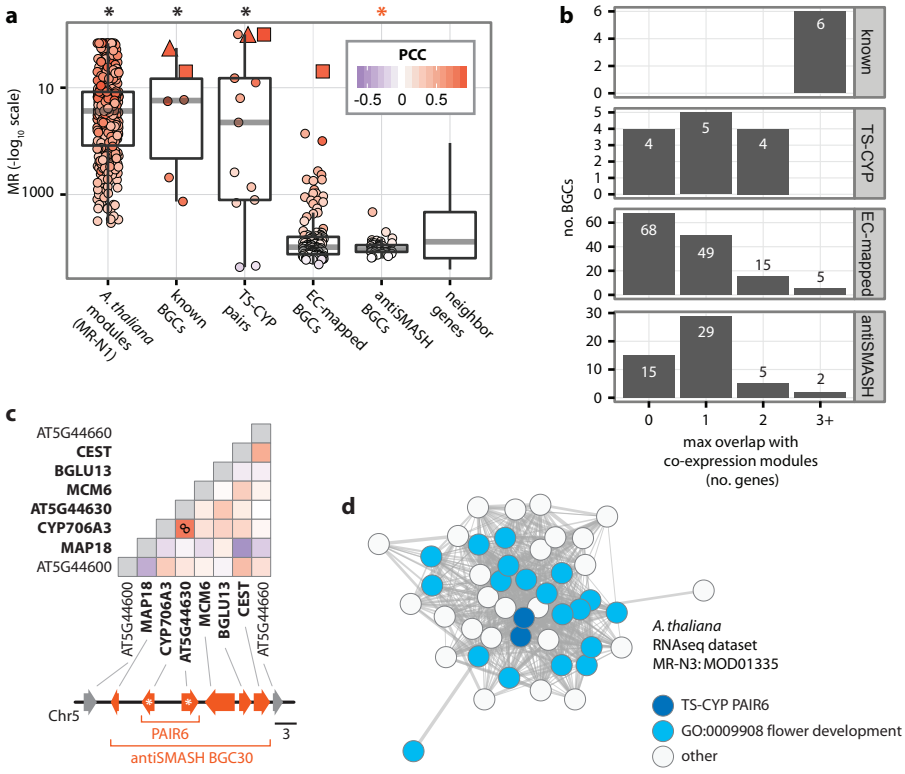

Figure 5. The genes comprising the majority of bioinformatically predicted BGCs are not co-expressed. a) Comparison of average co-expression of modules versus characterized and putative BGCs. The bottom and top of each box plot correspond to the first and third quartiles (the 25th and 75th percentiles), 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; *$P \leq 0.05$ (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 marked with asterisks). 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.