plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters

Satria A. Kautsar*[1,2], Hernando G Suarez Duran*[1], Kai Blin[3], Anne Osbourn[4], Marnix H. Medema[1]

[1] Bioinformatics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[2] Teknik Informatika, Universitas Lampung, Jln. Sumantri Brojonegoro No. 01, Lampung 35141, Indonesia
[3] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
[4] Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, United Kingdom

Abstract

Plants around the globe produce a wide variety of specialized metabolites that play key roles in communication and defense. Recently, evidence has been accumulating that—like in microbes—the genes encoding the biosynthetic pathways towards these metabolites are often densely clustered in specific genomic loci: biosynthetic gene clusters (BGCs). This offers great potential for genome-based discovery of plant natural products. However, effective computational tools to identify and analyze plant BGCs have thus far been lacking. Here, we introduce plantiSMASH, a versatile online analysis platform that automates the identification of candidate plant BGCs, as well as their comparative genomic and transcriptomic analysis. The cluster detection logic, validated on a set of all plant BGCs that have been experimentally characterized thus far, is able to pinpoint many complex metabolic loci across the Plant Kingdom. Additionally, interactively visualized coexpression analysis and comparative cluster-cluster alignment allow users to judge multiple sources of evidence for a candidate BGC to encode a group of enzymes that truly functions jointly in a biosynthetic pathway. Furthermore, plantiSMASH finds coexpression correlations between candidate BGCs and genes elsewhere in the genome. Altogether, this new software provides a comprehensive toolkit for plant geneticists to further explore the nature of gene clustering in plant metabolism. Moreover, spurred by the continuing decrease in costs of plant genome sequencing and assembly, it will soon allow natural product chemists to apply genome mining technologies to the discovery of novel medicinal compounds from a wide range of plant species.

Introduction

Across Planet Earth, bacteria, fungi and plants produce an immense diversity of specialized metabolites, each with their own specific ecological roles in the manifold interorganismal interactions in which they engage. This diverse specialized metabolism is a rich source of natural products that are used widely in medicine, agriculture and manufacturing. In bacteria and fungi, where genes for most specialized metabolic pathways are physically clustered in so-called biosynthetic gene clusters (BGCs), the rapid accumulation of genome sequences has revolutionized the process of natural product discovery: indeed, genome mining has now become a dominant method for the discovery of novel molecules (1–4). In the genome mining process, BGCs are computationally identified in genome sequences and then linked to compounds through functional analysis (e.g., using metabolomic data, chemical structure predictions, mutant libraries, and/or heterologous expression). Many sequence-based aspects of this genome mining procedure are facilitated by the online antiSMASH framework, which was launched in 2010 (5) and has seen continuous development since then (6, 7). The genome mining procedure has two main purposes: 1) finding biosynthetic genes for important known compounds to allow heterologous production through fermentation in industrial strains, and 2) identifying novel natural product chemistry guided by biosynthetic gene cluster diversity. Altogether, this development has appropriately been termed the 'gene cluster revolution' (4).

In recent years, it has become clear that not only microbial, but also plant biosynthetic pathways are frequently chromosomally clustered: after the initial discoveries of the cyclic hydroxamic acid 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) and avenacin gene clusters (8, 9), around thirty plant BGCs have been discovered (10, 11). Together, they encode the production of a wide range of different compounds, including cyclic hydroxamic acids, di- and triterpenes, steroidal and benzylisoquinoline alkaloids, cyanogenic glucosides and polyketides. In the genome of the model plant species *Arabidopsis thaliana* alone, four BGCs have been linked to specific metabolites, and recent analyses based on epigenomic profiling indicate the presence of various additional uncharacterized ones (12).

Various technological developments in eukaryote genome sequencing (13) are finally making complete plant genome sequencing feasible at larger scales: high-quality plant genome sequences for almost 100 species are now already publicly available, and more or less complete genomes can be sequenced for as little as a 10-50k US dollars each. Hence, genome mining may become an important methodology in the study of plant natural products as well, and a realistic opportunity thus presents itself for the plant natural product research community to have a 'gene cluster revolution' of its own. Naturally, a key technology required to realize this is a computational framework specifically designed for the identification and analysis of plant BGCs. Importantly, tools available for bacterial and fungal genome mining do not suffice for plants (14), as 1) plant biosynthetic pathways involve unique enzyme families not found in bacteria and fungi; 2) not all plant biosynthetic pathways are clustered (e.g., anthocyanins (15)), so identification of a biosynthetic gene does not equal identification of a BGC; 3) intergenic distances in plant genomes are larger and much more variable (16–19); 4) plant genomes contain clustered groups of genes (e.g., tandem arrays) whose products do not constitute a pathway; 5) several plant pathways are split across more than one BGC (20, 21).

Here, we introduce antiSMASH for plants (or 'plantiSMASH' in short), which has been designed to tackle each of these challenges. Through a comprehensive library of profile Hidden Markov Models (pHMMs) for enzyme families known to be involved in plant biosynthetic pathways, combined with CD-HIT clustering of predicted protein sequences belonging to the same family, it allows the efficient identification of genomic loci encoding multiple different (sub)families of specialized metabolic enzymes. Moreover, comparative genomic analysis as well as analysis of gene expression patterns within these candidate BGCs allow assessment of each locus for its likelihood to encode genes working together in one pathway. Finally, coexpression analysis between candidate BGCs and with other genes across the genome allows identification of biosynthetic pathways that are encoded on multiple loci. To exploit this new framework, we offer an initial analysis of BGC diversity across the plant kingdom, which showcases the presence of many complex biosynthetic loci in diverse species.

Methods and Implementation

A procedure for the identification of candidate plant biosynthetic gene clusters

The microbial version of antiSMASH (5) predicts BGCs by using HMMer (22) to identify specific (combinations of) signature protein domains that belong to scaffold-generating enzymes specific for a class of biosynthetic pathways. Subsequently, hit genes are used as anchors from which gene clusters are extended upstream and downstream by a specified extension distance.

Although very effective for detecting biosynthetic clusters on bacteria and fungi, this procedure is unfit to detect biosynthetic gene clusters in plants, for the reasons described above. To address these differences, a novel detection strategy was chosen (**Figure 1**): instead of identifying BGCs through the identification of core scaffold-generating genes alone, plantiSMASH identifies them by looking for all genes predicted to encode different types of biosynthetic enzymes, including those required for tailoring of the scaffold.

To determine what constitutes a high-potential candidate BGC, we make use of the recently proposed definition for plant BGCs as 'genomic loci encoding genes for a minimum of three different types of biosynthetic reactions (i.e. genes encoding functionally different (sub)classes of enzymes).' More specifically, with default settings plantiSMASH defines clusters as loci where at least three different enzyme subclasses belonging to at least two different enzyme classes are co-located on the same locus. Enzyme classes are identified using profile Hidden Markov Models (pHMMs) specific for each class; to count the number of subclasses of each enzyme class at a certain locus, the CD-HIT algorithm (23) is employed for sequence-based clustering to identify groups of sequences within an enzyme class with (by default) >50% mutual amino acid sequence identity.

In order to identify all classes of biosynthetic enzymes known to be involved in plant specialized metabolic pathways, we performed a comprehensive literature search of previously characterized plant biosynthetic pathways, which resulted in a list of 62 protein domains (see **SI Table 1**). Most of these protein domains are represented by pHMMs from the Pfam database (24), and custom pHMMs were only generated for enzyme families not (fully) covered by Pfam domains. We consciously refrained from attempting to construct custom pHMMs for all enzyme families known to be involved in plant biosynthetic pathways, as the limited amount of training data available would lead to an overly strict prediction system that would no longer be able to detect biosynthetic novelty; instead, we assume that the broad enzyme families covered by Pfam domains are likely to be biosynthetically involved if multiple enzymes from these different families are encoded together in the same locus. As in the microbial version of antiSMASH, the presence of genes predicted to encode signature enzymes (defined as enzymes that determine the chemical class of the end compound, such as terpene synthases) in a candidate BGC are used to assign a cluster to a biosynthetic class (see **SI Table 2** for cluster rules). However, compared to the microbial version, the biosynthetic classes in 'plantiSMASH' are more of an approximation, since not all signature enzyme families used can be unequivocally used to predict the compound type; e.g., while strictosidine synthase (25) and norcoclaurine synthase (26) are well-characterized members of the Bet v1 enzyme family, it is not clear what proportion of this family have similar Pictet-Spenglerase(-like) catalytic activities.

Another particular challenge for BGC detection in plant genomes is the large variation in gene density that occurs not only between but also within plant genomes (16–19). Replacing the static kilobase distance cut-off of microbial antiSMASH by a fixed cut-off based on the maximum number of genes that lie between each pHMM hit also does not provide a solution, as BGCs would then be allowed to cross large repeat regions or even centromeres. Therefore, we chose an alternative more dynamic cut-off that is a linear function of local gene density (defined as the gene density of the ten genes nearest to a pHMM hit), and applies a multiplier to calculate the cut-off in kb that is optimal for that specific genomic region.

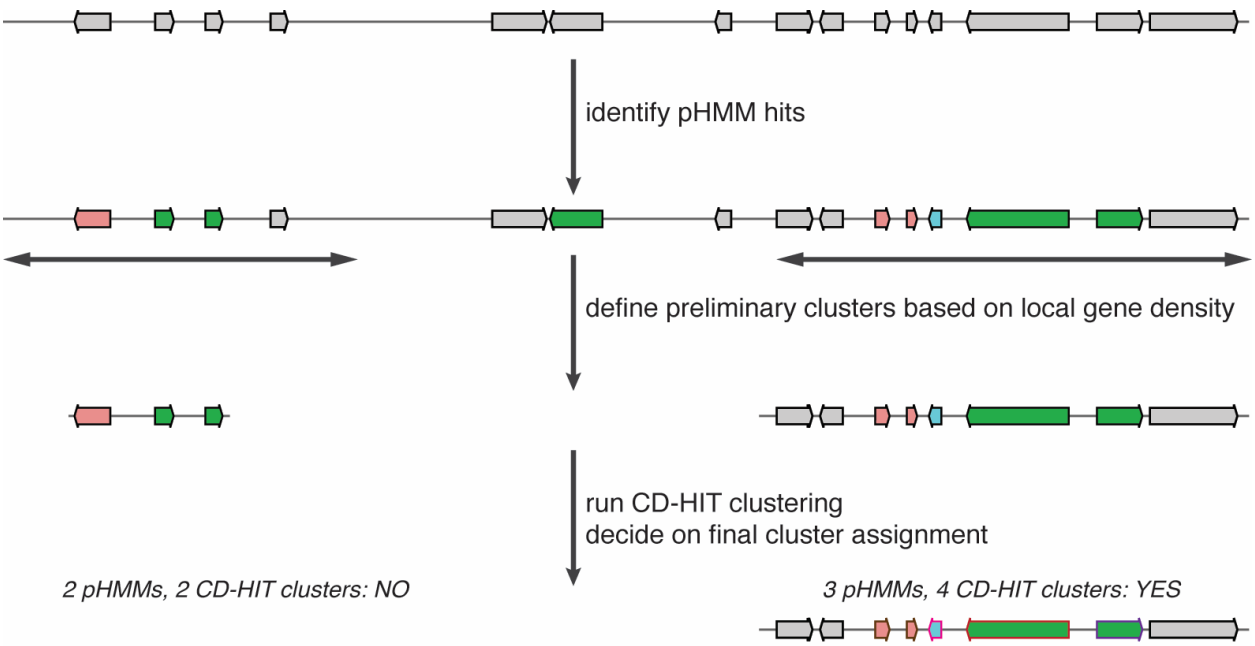


Figure 1: General strategy followed by plantiSMASH for the identification of plant BGCs. First, a library of 62 pHMMs is used to identify genes encoding biosynthetic enzymes across the genome. Subsequently, groups of hit genes that lie close to each other on the genome are combined into clusters; the maximum distance between hit genes is determined based as a function of the local gene density (i.e., the number of genes per kb). By default, a cluster should at least contain hits to two different pHMMs to proceed to the next stage (preliminary cluster assignment). To the evaluate these preliminary clusters on the numbers of enzyme subclasses they encode, , sequence-based clustering is performed on all genes in a preliminary cluster to estimate how many enzyme subclasses are encoded; if a cluster contains a sufficient number of CD-HIT groups (minimally three by default), it is defined as a candidate BGC and displayed on the plantiSMASH output page.

Flexible and user-friendly input and output

To obtain reliable BGC predictions, a high-quality annotation of gene features in a genome is essential. While we do make available the option to run GlimmerHMM (27) on plant genome sequences, performing de novo gene finding on a raw FASTA file is not desirable, given the relatively low accuracy of these procedures. Because, additionally, the GenBank and EMBL input formats previously accepted for antiSMASH are not available for many plant genomes, we now allow users to supply input also in FASTA+GFF3 format, currently the most widely used format for describing plant genome annotations. For this, we implemented a new module based on Biopython's GFF parsing package (http://biopython.org/wiki/GFF_Parsing) capable of combining the CDS features from the sequence input sequence, if any, with those of a file

compliant to the Generic Feature Format Version 3 as defined by The Sequence Ontology in 2003 (https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md). To properly match GFF3 CDS features to their correct sequence, the module demands record names (chromosome/scaffold/contigs) to be identical in both inputs; the only exception being if both inputs only contain one record, in which case the requirement is instead that no feature has coordinates outside the sequence range. This new module allows plantiSMASH to be used with genomes that are only annotated with GFF3 files, such as many of those present in the Joint Genome Institute's Phytozome database (28).

Based on the biosynthetic gene cluster predictions, a rich and interactive HTML output is generated (**Figure 2**), which is largely reminiscent of the output of microbial antiSMASH jobs (5). Additionally, genes in the visualization page for each candidate BGC are colored based on the class of enzymes encoded, and a legend is provided that details the color scheme. On mouse click, panels for each gene provide information on the pHMMs that have hits against it, as well as on the amino acid identity to homologous genes within the same locus as calculated by CD-HIT.

Coexpression analysis identifies pathways within and between gene clusters

As plant scientists are just beginning to understand the phenomenon of metabolic gene clustering in plant genomes, it is currently unknown which proportion of genomic loci that encode multiple contiguous biosynthetic enzyme-encoding genes are bona fide BGCs in the sense that their constituent genes are involved in one specific pathway. One powerful strategy to predict whether genes are involved in the same pathway is the use of coexpression analysis, in which their expression patterns are compared across a wide range of samples. This strategy has proven very effective in the de novo identification of gene sets involved in biosynthetic pathways, even if they are not physically clustered on the chromosome (29).

To allow detailed investigation of whether genes in a cluster show coexpression, we added a dedicated analysis module: CoExpress. This module reads transcriptomic datasets, either in SOFT format (from the NCBI Gene Expression Omnibus) or in comma-separated (CSV) format, and generates powerful visualizations of these data for each candidate BGC. Because combining many datasets into one coexpression analysis may blot out coexpression signals that are very specific to certain biological or chemical treatments (which often highly specifically incite expression of plant specialized metabolic pathways), we designed the module in such a way that it visualizes one transcriptomic dataset at a time. This has the added value that the user can browse through multiple datasets and can individually assess specific samples that are linked to a treatment of interest.

The visualizations of within-cluster coexpression patterns are twofold: First, a hierarchically clustered heatmap visualization, plotted using a modified version of the InCHlib (http://www.openscreen.cz/software/inchlib/home) JavaScript library, offers a direct view of patterns in and relationships between the supplied normalized gene expression values. The dendrogram is generated using a coexpression distance metric with a complete-linkage hierarchical clustering method. In this metric, the Pearson Correlation Coefficient (PCC) is transformed directly into a distance value scaled from 0 to 200 (0 for PCC = 1, or positively correlated, and 200 for PCC = -1, or negatively correlated). In order to make correlations maximally visible, the color scheme is normalized per gene (row) by default; however, the user

can also select for the color scheme to be normalized by sample (column). Second, a gene cluster-specific coexpression network (30) (with a default distance based cutoff of < 50, dynamically adjustable) summarizes the correlations and helps to identify specific groups of genes in the locus that are highly coexpressed: these occur as connected components with high numbers of edges.

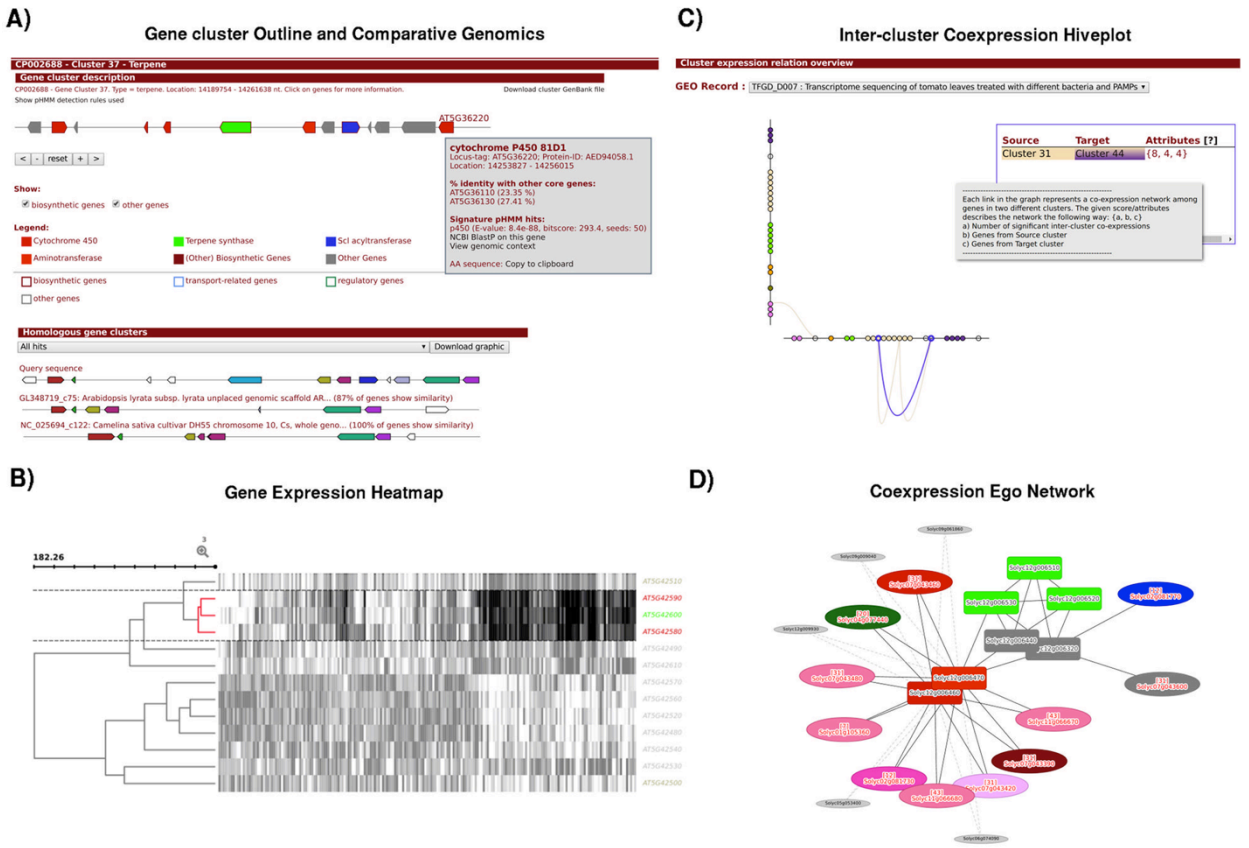


Figure 2: Outputs generated by the plantiSMASH pipeline. The figure illustrates several visualized outputs generated by plantiSMASH, as they appear for various biosynthetic gene clusters of known natural products. A) Visual overview generated for each gene cluster; in this case, the tirucalladienol cluster from *A. thaliana* (31) is shown. Gene annotations and pHMM hit details appear on mouse click. Also, ClusterBlast output showing alignment of homologous genomic loci across other genomes of related species is provided. B) Example of a gene expression heat map, showing coexpression among the core genes of the marneral BGC from *A. thaliana* (32) (and not with the flanking genes). C) Hive plot on the overview page, which highlights pairs of candidate BGCs which show many coexpression correlations between their genes; in this example view, the coexpression links between the two loci encoding alpha-tomatine biosynthesis in *Solanum lycopersicum* (20) are highlighted (clusters 31 & 44). D) Example ego network that summarizes coexpression correlations between members of the alpha-tomatine gene (cluster 44), as well as with genes in other gene clusters (including the other alpha-tomatine biosynthetic locus, cluster 31), and with genes elsewhere on the genome.

Coexpression analysis is not just useful for analysis of functional connections within a candidate BGC, but also allows prediction of functional links with other genomic loci. It is now well-understood that several plant BGCs do not act alone, but rather in concert with another BGC or with individual enzyme-coding genes elsewhere on the genome (11). Therefore, plantiSMASH leverages coexpression data to offer two analyses that identify these trans-genomic

interactions: First, the BGC-specific coexpression network can be extended to display a first-order ego network that incorporates genes elsewhere on the genome that either 1) are members of another candidate BGC and show high gene expression correlation (> 0.9 PCC) with at least one gene in the BGC, or 2) contain a 'biosynthetic' domain (defined as being one of the domains in **SI Table 1**) and show high gene expression correlation with at least two genes in the BGC, at least one of which being a biosynthetic gene itself. Second, interactions between candidate BGCs are summarized in a hive plot, in which pairs of clusters are connected by an edge if the genes of both clusters create at least one subnetwork that satisfies the following criteria: 1) All nodes belong to the same Louvain community (33), as determined by analyzing the full coexpression network of all candidate clusters' genes; 2) All nodes have a transitivity greater than zero; 3) The subnetwork contains at least two genes from each cluster; 4) The subnetwork contains at least one gene per cluster that has a biosynthetic domain; and 5) The subnetwork contains at least three genes with a biosynthetic domain.
All in all, the coexpression analysis of candidate BGCs allows effective prioritization for, e.g., heterologous expression studies. Yet, it should still be kept in mind that loci that do not show high coexpression might still encode genes that are jointly involved in a biosynthetic pathway, e.g., if the transcriptomic samples available do not include any treatments that induce the expression of the pathway, or if expression of the pathway is sequestered either spatially across tissues or in terms of timing.

Comparative genomic analysis shows conservation and diversification

Comparing a candidate BGC with homologous genomic loci in other plant genomes can give important information on its evolutionary conservation or diversification. Whereas strong conservation of clusteredness across larger periods of evolutionary time may point to a selective advantage of clustering for these genes, diversification of BGCs by co-option of other enzyme-coding genes may give clues to finding novel variants of natural products that have been generated through directional pathway evolution. In order to facilitate such comparative analysis on a case-by-case basis, we constructed a plant-specific version of the antiSMASH ClusterBlast module. To do so, we ran plantiSMASH on a collection of all publicly available plant genomes, obtained from NCBI's GenBank, JGI's Phytozome and Kazusa. In order to avoid cases where loci homologous to detected candidate BGCs would not be included in the database by not satisfying the identification criteria, the thresholds for this search were lowered to find all genomic loci with two or more different enzymes, where the CD-HIT cut-off was also set to a generously inclusive level of 0.9. A total of 7,978 genomic loci were thus included in the plant ClusterBlast database. As in the microbial version of antiSMASH, the translated protein sequence of each predicted gene in a candidate BGC is searched against this database using the DIAMOND algorithm (34), and genomic loci are sorted based on the number of hits, conserved synteny and cumulative bit score. To also facilitate direct comparison with known plant BGCs, all plant BGCs with known products for which the sequence was available were added to the MIBiG repository (35), which allows users to find similarities between newly identified and known clusters with the KnownClusterBlast module of antiSMASH.

Precomputed results allow fast access to comprehensive plantiSMASH results

In order to allow users to directly access plantiSMASH results for publicly available plant genomes, runs for 48 high-quality plant genomes were precomputed and made available online at http://plantismash.secondarymetabolites.org. Importantly, publicly available gene expression datasets with sufficient numbers of samples to be suitable for coexpression analysis were loaded into these results. In total, 73 transcriptomic datasets were included for five species: *Arabidopsis thaliana*, *Solanum lycopersicum*, *Oryza sativa*, *Zea mays* and *Glycine max* (**SI Tables 3-5**). Sequences that are not publicly available (as well as available sequences with custom transcriptomic datasets) can be analyzed directly using the plantiSMASH web server at http://plantismash.secondarymetabolites.org. In this way, plantiSMASH results for all kinds of genomes and transcriptomes are optimally available to users.

Results and Discussion

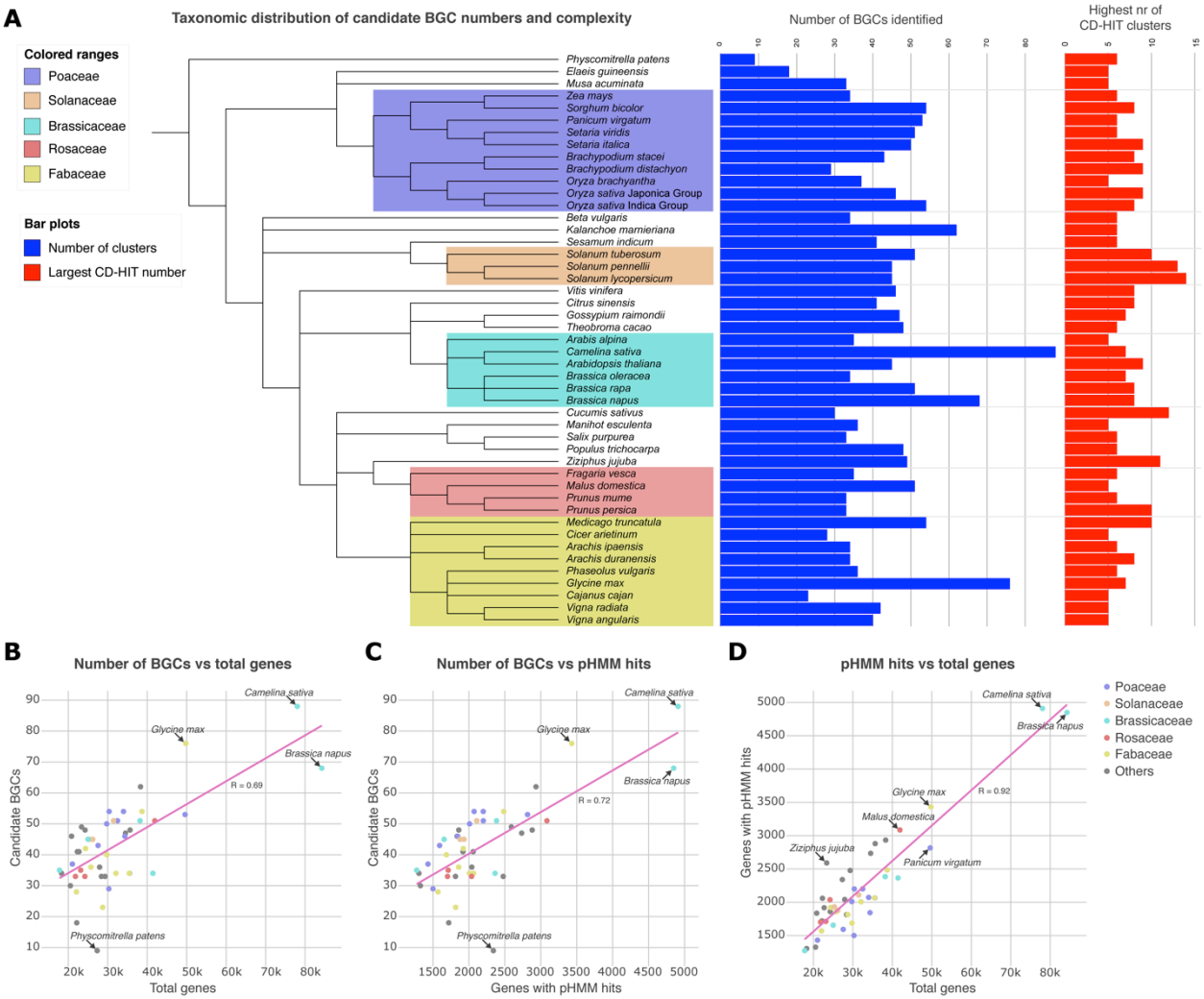


Figure 3: Numbers of candidate BGCs identified across the Plant Kingdom. A) PlantiSMASH BGC predictions plotted onto a phylogenetic tree of plant species for which chromosome-level genome assemblies are available. The blue bars indicate the number of candidate BGCs per genome, the red bars indicate the most complex candidate BGC identified in each species (in terms of the number of unique enzymes encoded, as defined by CD-HIT groups). B) Number of candidate BGCs plotted versus the total number of genes; as expected, more BGCs are found in larger genomes. Outliers represent genomes that have recently undergone whole-genome duplication, and the moss *Physcomitrella patens*, in the genome of which only a very low number of candidate BGCs is found. C) Number of candidate BGCs plotted versus the number of genes with pHMM hits to biosynthetic domains. D) Number of genes with biosynthetic domains plotted against the total number of genes; a linear correspondence is largely observed.

PlantiSMASH successfully detects all experimentally characterized plant biosynthetic gene clusters

Even though only a relatively small set of plant BGCs have been characterized, these ~30 BGCs still present the best objective test case for the BGC detection algorithm. Importantly, they range from complex BGCs with many different enzyme-coding genes, such as the noscapine and cucurbitacin BGCs (21, 36), to relatively simple ones that only encode a couple of enzymes, such as the dhurrin and linamarin/lotaustralin BGCs (37). When plantiSMASH was

run on a multi-GenBank file containing accurately annotated versions of all 19 known BGCs for which sequence information is available, all gene clusters were successfully detected with default settings. When ran on different genome annotation versions available from GenBank or Phytozome, BGCs of low complexity (i.e., with a small number of enzyme-coding genes) were occasionally missed when key genes were missing from the structural annotations, or when many false positive gene assignments were present in the region of interest (affecting the dynamic gene density-based cut-off of plantiSMASH): for example, the linamarin BGC from *Lotus japonicus* was not detected in assembly/annotation version 3.0, while it was detected in the older version 2.5. This highlights the importance of using high-quality genome annotations supported by transcriptomic data when using plantiSMASH to search for BGCs of interest. Alternatively, the stand-alone version of plantiSMASH provides additional cut-off methods (e.g., raw distance-based or gene-count-based) that can be attempted as well to mitigate such issues.

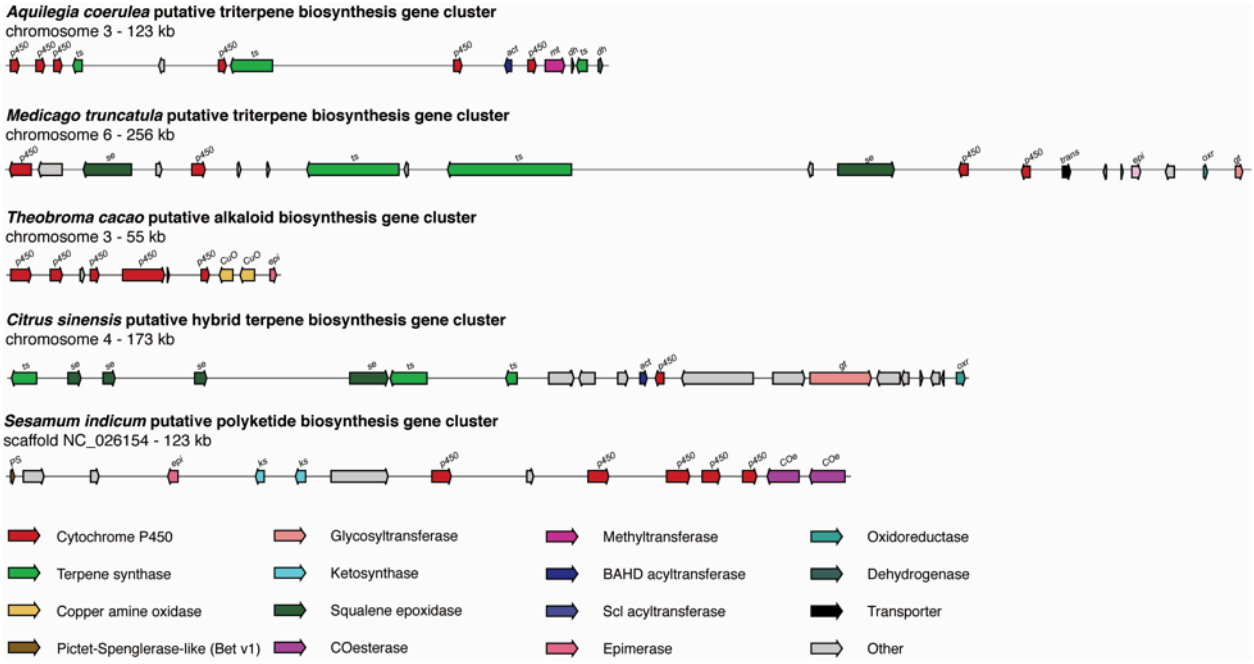


Figure 4: Example candidate BGCs identified by plantiSMASH. Five example candidate BGCs are shown, which cover a diverse range of enzymatic classes. Dozens of candidate BGCs of comparable complexity can be found across the precomputed plantiSMASH results that are available online.

Plant genomes contain large numbers of complex biosynthetic gene clusters

When run on the 43 plant genomes for which chromosome-level assemblies are currently available on either NCBI or Phytozome, plantiSMASH found a wide variety of candidate BGC numbers across plant taxonomy (**Figure 3**). In general, the numbers of candidate BGCs were relatively even between monocots and dicots (while very low in the only moss genome included), while the largest numbers of BGCs were found in dicot genomes. These outliers all corresponded to recent (partial) genome amplification events, such as in the case of *Camelina sativa* (88 candidate BGCs, see Ref. 31), *Brassica napus* (68 candidate BGCs, see Ref. 32), and *Glycine max* (76 candidate BGCs, see Ref. 33). Many of the BGCs in duplicated regions show hallmarks of divergence between the copies.

In many plant genomes, candidate BGCs of high complexity were identified, with as many as seven or eight different enzymatic classes encoded in the same tight genomic region. These constitutions are clearly non-random and make it promising to study candidate BGCs even in the absence of coexpression data. Dozens of such complex BGCs were found, which cover all known as well as putative pathway classes; examples are provided in **Figure 4**.

Coexpression patterns can guide BGC prioritization

To illustrate how coexpression data within plantiSMASH should be interpreted, we subjected the candidate BGCs identified in the genome of *A. thaliana* to a more detailed analysis using an example transcriptomic dataset. For this, we compiled two sets of gene expression datasets, one containing transcriptomic experiments of biological treatments (defense; **SI Table 3**) and one containing experiments of hormone treatments and non-biological stress inductions (**SI Table 4+5**). Together, these datasets comprise transcriptomic measurements of 1047 samples. While, intriguingly, in those datasets, we did not observe a statistically significant distinction between the complement of coexpression Pearson correlation values within all candidate BGCs and those on randomly chosen genomic loci of the same size (the biological induction dataset performed 'best' in the Wilcoxon rank-sum test with P=0.08), individual BGCs did show unusually strong coexpression in these data: Of the four BGCs known to encode entire biosynthetic pathways (for marneral, thalianol, tirucalla and arabidiol/baruol), only one (the thalianol cluster) showed a clearly significant coexpression pattern compared to randomly chosen contiguous groups of genes of the same size elsewhere in the genome (P=3.35e-5, Wilcoxon rank-sum test). Besides this cluster, two other BGCs also showed similarly striking coexpression patterns: the cluster ranging from AT3G57000 to AT3G57060 (P=6.06e-6) and the cluster that ranges from AT4G14050 to AT4G14096 (P=3.34e-3).

There are several explanations for the fact that strong coexpression is observed for some known as well as candidate BGCs but not others. A first explanation is that their coordinated expression is induced by conditions not included in these transcriptomic experiments. After all, previous studies have shown clear coexpression of, e.g., the tirucallol cluster (31). In other words, absence of evidence is not evidence of absence: when genes are not coexpressed in certain data, this should not be interpreted as definitive counterevidence to them working together in a pathway. A second explanation is that some BGCs are expressed at very low levels under standard growth conditions (e.g., the marneral cluster (32)), leading to an unfavorable signal-to-noise ratio. A third explanation is that a number of candidate BGCs probably do not encode entire consistently coexpressed biosynthetic pathways by themselves; evidence for this comes from an analysis of characterized enzyme-coding genes inside these candidate BGCs (**SI Table 6**); e.g., *AT1G24100* and *AT5G57220*, which occur in two different candidate BGCs, are known to be involved in two different branches of glucosinolate biosynthesis (41, 42), a complex multifurcated pathway that shows only partial and fragmented genomic clustering.

Contrary to what might be expected, however, there was no strong correlation (R=0.11, and P=0.44 when fitting linear regression) of coexpression with cluster size (**SI Figure 1**), which suggests that the default plantiSMASH BGC prediction cut-offs are not set too inclusively. Indeed, the highly coexpressed thalianol gene cluster, for example, only comprises a small set of enzyme-coding genes.

All in all, through its flexibility, the CoExpress module in plantiSMASH allows judging each candidate BGC through the lens of multiple transcriptomic datasets (while reducing noise levels); the more datasets are available, the larger the chance is that one of them allows the identification of strong coexpression. Choosing specific datasets (or combinations of datasets) for which, e.g., metabolomic evidence is available, can help users to identify the likelihood that this BGC encodes a specific pathway. In the end, however, coexpression alone often does not provide the final answer on whether a candidate BGC is 'real'; the further development of integrative approaches that combine multiple data types (14) is clearly needed for this in the future.

Identification of plant BGCs paves the way towards genome-based natural product discovery in plants

The highly automated discovery of candidate BGCs by plantiSMASH and the powerful visualizations of coexpression data that allow their prioritization present a key technological step in the route towards high-throughput genome mining of plant natural products. As plant genome sequencing and assembly technologies continue to improve at a rapid pace, it is likely that high-quality plant genomes for thousands of species will soon be available; hence, 'clustered' biosynthetic pathways present low-hanging fruits for the discovery of novel molecules. Empowered by synthetic biology tools and powerful heterologous expression systems in yeast and tobacco (43–47), this will likely make it possible to scale up plant natural product discovery tremendously.

Continued development of the antiSMASH/plantiSMASH framework in the future is needed to further accelerate this process: e.g., the development of (machine-learning) algorithms that predict substrate specificities of key enzymes like terpene synthases, and the systematic construction of pHMMs for automated subclassification of complex enzyme families such as cytochrome P450s and glycosyltransferases, will allow more powerful predictions of the natural product structural diversity encoded in diverse BGCs. Additionally, detailed evolutionary genomic analysis of the phenomenon of gene clustering, including BGC birth, death and change processes, will further our understanding of how BGCs facilitate natural product diversification during evolution. As more plant BGCs are experimentally characterized, the algorithms will co-evolve with the knowledge gained, and more detailed class-specific cluster detection rules could be designed; moreover, it will become clearer what does and what does not constitute a bona fide BGC. Finally, when scientists further unravel the complexities of tissue-specific and differentially timed gene expression of plant biosynthetic pathways, we will learn more on how best to leverage coexpression data for biosynthetic pathway prediction.

Thus, a more comprehensive understanding of the remarkable successes of evolution to generate an immense diversity of powerful bioactive molecules will hopefully make it possible for biological engineers to mimick nature's strategies and deliver many useful new molecules for use in agricultural, cosmetic, dietary and clinical applications.

Acknowledgements

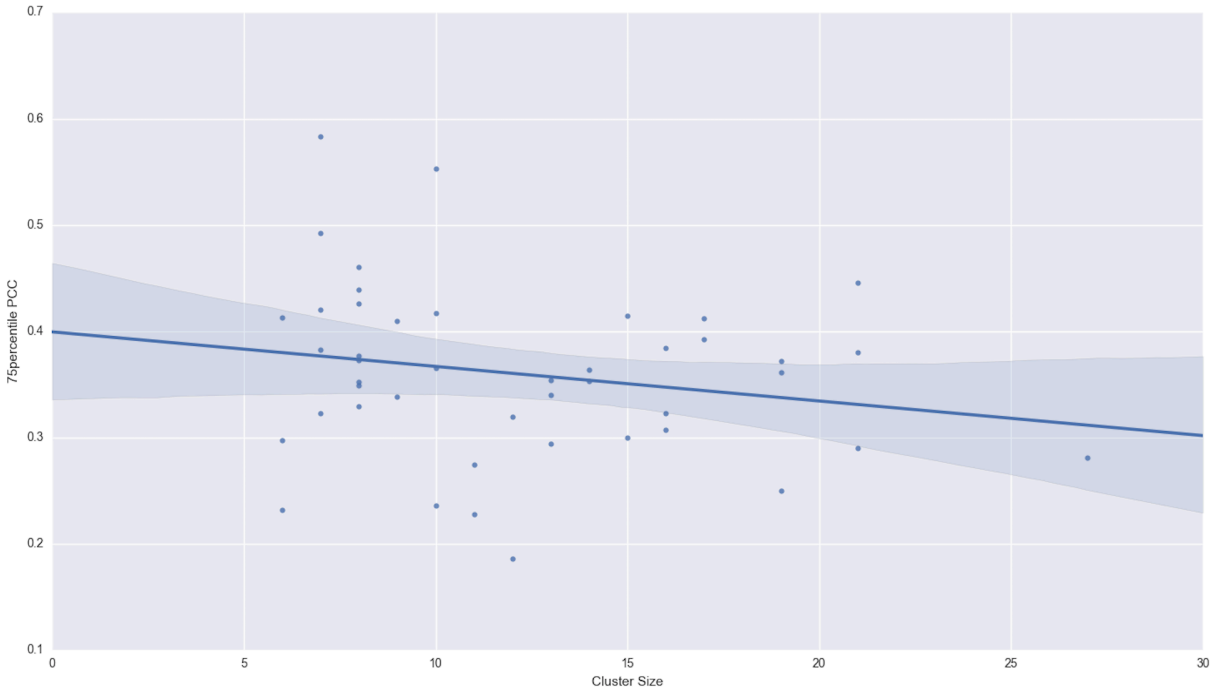
We thank Linh Nguyen for providing SI Table 6. S.A.K. is supported by the Graduate School for Experimental Plant Sciences (EPS). K.B. is supported by a grant from the Novo Nordisk Foundation. A.O. is supported by the UK Biotechnological and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Grant 'Understanding and Exploiting Plant and Microbial Metabolism' (BB/J004561/1), the John Innes Foundation, the joint Engineering and Physical Sciences Research Council/ BBSRC-funded OpenPlant Synthetic Biology Research Centre grant BB/L014130/1 and a National Institutes of Health Genome to Natural Products Network award U101GM110699. M.H.M. is supported by VENI grant 863.15.002 from The Netherlands Organization for Scientific Research (NWO), and by the Genome to Natural Products Network.

References

1. Rutledge,P.J. and Challis,G.L. (2015) Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.*, **13**, 509–523.
2. Medema,M.H. and Fischbach,M.A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.
3. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.*, **33**, 988–1005.
4. Jensen,P.R. (2016) Natural Products and the Gene Cluster Revolution. *Trends Microbiol.*, 10.1016/j.tim.2016.07.006.
5. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences . *Nucleic Acids Res.*, **39**, W339-W346.
6. Blin,K., Medema,M.H., Kazempour,D., Fischbach,M.A., Breitling,R., Takano,E. and Weber,T. (2013) antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204-W212.
7. Weber,T., Blin,K., Duddela,S., Krug,D., Kim,H.U., Bruccoleri,R., Lee,S.Y., Fischbach,M.A., Muller,R., Wohlleben,W., *et al.* (2015) antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237–W243.
8. Frey,M., Chomet,P., Glawischnig,E., Stettner,C., Grün,S., Winklmair,A., Eisenreich,W., Bacher,A., Meeley,R.B., Briggs,S.P., *et al.* (1997) Analysis of a chemical plant defense mechanism in grasses. *Science*, **277**, 696–699.
9. Qi,X., Bakht,S., Leggett,M., Maxwell,C., Melton,R. and Osbourn,A. (2004) A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 8233–8288.
10. Nützmann,H.-W. and Osbourn,A. (2014) Gene clustering in plant specialized metabolism. *Curr. Opin. Biotechnol.*, **26**, 91–99.
11. Nützmann,H.W., Huang,A. and Osbourn,A. (2016) Plant metabolic gene clusters - from genetics to genomics. *New Phytol.*, **211**, 771-789.
12. Yu,N., Nützmann,H.-W., MacDonald,J.T., Moore,B., Field,B., Berriri,S., Trick,M., Rosser,S.J., Kumar,S.V., Freemont,P.S., *et al.* (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.*, **44**, 2255-2265.
13. VanBuren,R., Bryant,D., Edger,P.P., Tang,H., Burgess,D., Challabathula,D., Spittle,K., Hall,R., Gu,J., Lyons,E., *et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature*, **527**, 508–511.
14. Medema,M.H. and Osbourn,A. (2016) Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.*, **33**, 951–962.
15. Shi,M.-Z. and Xie,D.-Y. (2014) Biosynthesis and metabolic engineering of anthocyanins in Arabidopsis thaliana. *Recent Pat. Biotechnol.*, **8**, 47–60.
16. Keller,B. and Feuillet,C. (2000) Colinearity and gene density in grass genomes. *Trends Plant Sci.*, **5**, 246–51.
17. Kellogg,E.A. and Bennetzen,J.L. (2004) The evolution of nuclear genome structure in seed plants. *Am. J. Bot.*, **91**, 1709–1725.
18. Sandhu,D. and Gill,K.S. (2002) Gene-containing regions of wheat and the other grass genomes. *Plant Physiol.*, **128**, 803–811.
19. Ibarra-Laclette,E., Lyons,E., Hernández-Guzmán,G., Pérez-Torres,C.A., Carretero-Paulet,L., Chang,T.-H., Lan,T., Welch,A.J., Juárez,M.J.A., Simpson,J., *et al.* (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94–98.

20. Itkin,M., Heinig,U., Tzfadia,O., Bhide,A.J., Shinde,B., Cardenas,P.D., Bocobza,S.E., Unger,T., Malitsky,S., Finkers,R., *et al.* (2013) Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, **341**, 175–179.
21. Shang,Y., Ma,Y., Zhou,Y., Zhang,H., Duan,L., Chen,H., Zeng,J., Zhou,Q., Wang,S., Gu,W., *et al.* (2014) Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science*, **346**, 1084–1088.
22. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
23. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
24. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J., *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
25. Wu,F., Zhu,H., Sun,L., Rajendran,C., Wang,M., Ren,X., Panjikar,S., Cherkasov,A., Zou,H. and Stöckigt,J. (2012) Scaffold tailoring by a newly detected Pictet-Spenglerase activity of strictosidine synthase: from the common tryptoline skeleton to the rare piperazino-indole framework. *J. Am. Chem. Soc.*, **134**, 1498–1500.
26. Lee,E.-J. and Facchini,P. (2010) Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. *Plant Cell*, **22**, 3489–3503.
27. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
28. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N., *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
29. Rajniak,J., Barco,B., Clay,N.K. and Sattely,E.S. (2015) A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. *Nature*, **525**, 376-379.
30. Serin,E.A.R., Nijveen,H., Hilhorst,H.W.M. and Ligterink,W. (2016) Learning from Co-expression Networks: Possibilities and Challenges. *Front. Plant Sci.*, **7**, 444.
31. Boutanaev,A.M., Moses,T., Zi,J., Nelson,D.R., Mugford,S.T., Peters,R.J. and Osbourn,A. (2014) Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci.*, **112**, E81-E88.
32. Field,B., Fiston-Lavier,A.-S., Kemen,A., Geisler,K., Quesneville,H. and Osbourn,A.E. (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 16116–16121.
33. Blondel,V., Guillaume,J., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J Stat Mech*, **10**, P10008.
34. Buchfink,B., Xie,C. and Huson,D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
35. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C., *et al.* (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
36. Winzer,T., Gazda,V., He,Z., Kaminski,F., Kern,M., Larson,T.R., Li,Y., Meade,F., Teodor,R., Vaistij,F.E., *et al.* (2012) A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, **336**, 1704–1708.
37. Takos,A.M., Knudsen,C., Lai,D., Kannangara,R., Mikkelsen,L., Motawia,M.S., Olsen,C.E., Sato,S., Tabata,S., Jørgensen,K., *et al.* (2011) Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in Lotus japonicus and suggests the repeated evolution of this chemical defence pathway. *Plant J.*, **68**, 273–286.
38. Kagale,S., Koh,C., Nixon,J., Bollina,V., Clarke,W.E., Tuteja,R., Spillane,C., Robinson,S.J., Links,M.G., Clarke,C., *et al.* (2014) The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.*, **5**, 3706.
39. Chalhoub,B., Denoeud,F., Liu,S., Parkin,I.A.P., Tang,H., Wang,X., Chiquet,J., Belcram,H.,

Tong,C., Samans,B., *et al.* (2014) Plant genetics. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, **345**, 950–953.
40. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J., *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
41. Pfalz,M., Vogel,H. and Kroymann,J. (2009) The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in Arabidopsis. *Plant Cell*, **21**, 985–999.
42. Grubb,C.D., Zipp,B.J., Ludwig-Müller,J., Masuno,M.N., Molinski,T.F. and Abel,S. (2004) Arabidopsis glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. *Plant J.*, **40**, 893–908.
43. Liu,W., Yuan,J.S. and Stewart,C.N. (2013) Advanced genetic tools for plant biotechnology. *Nat. Rev. Genet.*, **14**, 781-793.
44. Patron,N.J. (2014) DNA assembly for plant biology: techniques and tools. *Curr. Opin. Plant Biol.*, **19**, 14–19.
45. Thimmappa,R., Geisler,K., Louveau,T., O'Maille,P. and Osbourn,A. (2014) Triterpene biosynthesis in plants. *Annu. Rev. Plant Biol.*, **65**, 225–257.
46. Patron,N.J., Orzaez,D., Marillonnet,S., Warzecha,H., Matthewman,C., Youles,M., Raitskin,O., Leveau,A., Farré,G., Rogers,C., *et al.* (2015) Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytol.*, **208**, 13–19.
47. Casini,A., Storch,M., Baldwin,G.S. and Ellis,T. (2015) Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.*, **16**, 568–576.



SI Figure 1: Lack of correlation between candidate BGC size and coexpression correlation. The cluster size (measured by the number of genes in a BGC) is plotted against the 75th percentile of the Pearson correlation coefficients (PCCs) between genes in a cluster. No strong positive correlation is observed. The same is true when cluster size is plotted against median or average PCCs (data not shown).