MSc Thesis Report

Updating plantiSMASH gene cluster mining and implementing family subgrouping

Ziqiang Luo 1369903

Supervisors: Elena Del Pup, Prof. Marnix Medema

Bioinformatics 19-08-2024

Abstract

Since the beginning of human history, plants have been explored to discover secondary metabolites with significant environmental and health implications. The advent of plantiSMASH in 2017 speeded up the exploration by using biosynthetic gene clusters (BGCs) phenomena in plant genomes. However, with the rapid growth of genomic data and the characterisation of new BGCs and enzyme families, the need for an updated and more functional plantiSMASH is evident.

This project presents a comprehensive update to plantiSMASH, focusing on three key improvements: the refinement of existing rules to incorporate newly discovered plant BGCs, the implementation of an enzyme family subgroup identification module to predict substrate/product types, and the expansion of the pre-calculated reference BGC database to encompass a broader spectrum of homologous clusters.

In conclusion, the update in plantiSMASH realized in this project not only elevates the software's predictive accuracy but also broadens its applicability in the field of plant secondary metabolites research. These improvements are expected to accelerate the discovery of novel plant metabolites and their biosynthetic pathways, paving the way for innovative applications in medicine, agriculture, and environmental science.

Contents

1. Introduction

1.1. Plant-specialized metabolism is worth exploring

A vast array of known and unknown plant natural products important for the environment and human health, such as salicylic acid and paclitaxel (Owen et al., 2017). Mining plant biosynthetic genes allows researchers to find new metabolites and understand biosynthetic pathways required for large-scale biosynthesis in industrial hosts such as yeast. For example, there are still unclear steps in the complete biosynthetic pathway of the anticancer drug paclitaxel with a complex molecular structure, leading to the need for industrial synthesis to obtain its synthetic precursors from plants (Liu et al., 2024). In the early 21st century, mining relied on characterising individual pathway steps. The advent of sequencing and the increasing number of published plant genomes has allowed high-throughput genome mining approaches.

There are increasing numbers of examples of plant genomes which harbour physical clustering of biosynthetic genes (termed Biosynthetic Gene Clusters, or BGCs) that encode enzymes constituting a major part of or an entire biosynthetic pathway (Owen et al., 2017). Mining all possible (not just homologous) BGCs in genomes requires discovering regularities from known BGCs. The state-of-the-art software antiSMASH has used the different enzyme/domain composition rules to identify bacterial and fungal BGCs with respective product types since 2011 (https://antismash.secondarymetabolites.org/). The plant version of antiSMASH, plantiSMASH was developed in 2017 and uses rules based on plant BGCs. It identified >2,000 BGCs from 47 genomes (Kautsar et al., 2017). Its web server has thus far processed over 18,000 jobs.

1.2. Genome mining and the features of plantiSMASH version 1

PlantiSMASH uses profile Hidden Markov Models (pHMMs) to identify biosynthetic enzymes' domains in genomes. The core enzymes (Anarat-Cappillino & Sattely, 2014), such as terpene synthases, determine the chemical class of the end compound of pathways. For reducing noise, the pHMMs collections and rule formulations were based on reported plant BGCs that at least have a core enzymes gene and any two biosynthetic genes with identity less than 50% to take as having different functions (Fig. 1).

In plantiSMASH, two modules can be used to find homologous BGCs in databases. Using ClusterBLAST, users can get an overview of similar BGCs in other plant species, using a pre-calculated database of plantiSMASH results generated from publicly available high-quality plant genomes. Figure 2 shows one of the BGCs in the ClusterBLAST database found previously by plantiSMASH. The KnownClusterBLAST module can be used to quickly identify clusters resembling previously known ones stored in MIBiG databases (http://mibig.secondarymetabolites.org/) (Terlouw et al., 2023). plantiSMASH also provides users with modules to further validate BGCs by

combining gene expression data. All these make plantiSMASH the state-of-the-art software for identifying plant BGCs (Kautsar et al., 2018).

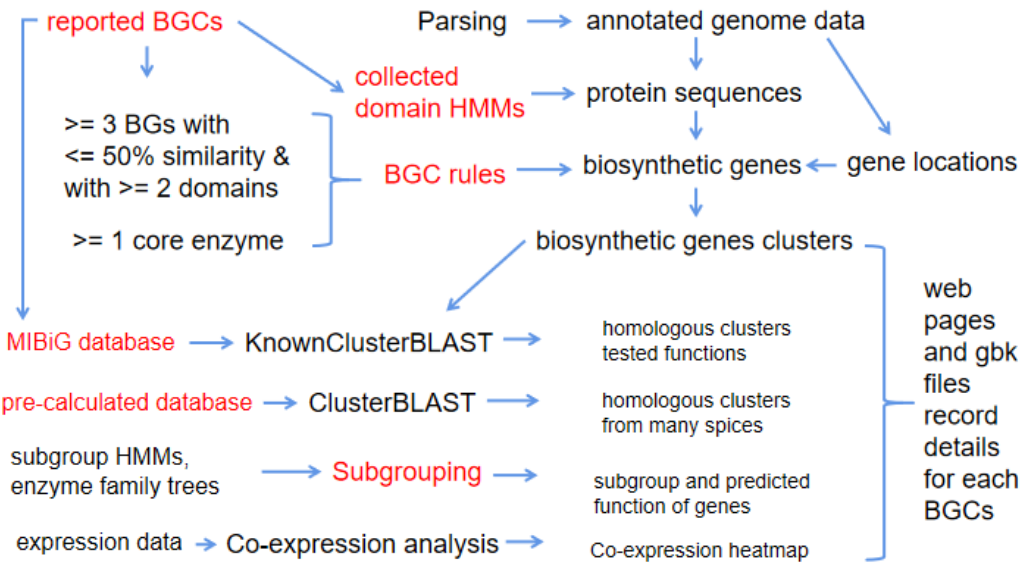

Figure 1|How plantiSMASH mines BGCs on a genome. First plantiSMASH processes the input genomic data to obtain the protein sequence and location of the genes. The collected enzyme domain HMMs are used to identify biosynthetic genes. Clusters formed from closely located biosynthetic genes that satisfy the BGC rules will be left behind and by other modules, providing more relevant information. The words in red indicate the involvement of this project.

1.3. Three parts to improve plantiSMASH mining BGCs and give more insights

1.3.1. Need to update HMMs library and rules to expand the mining ability

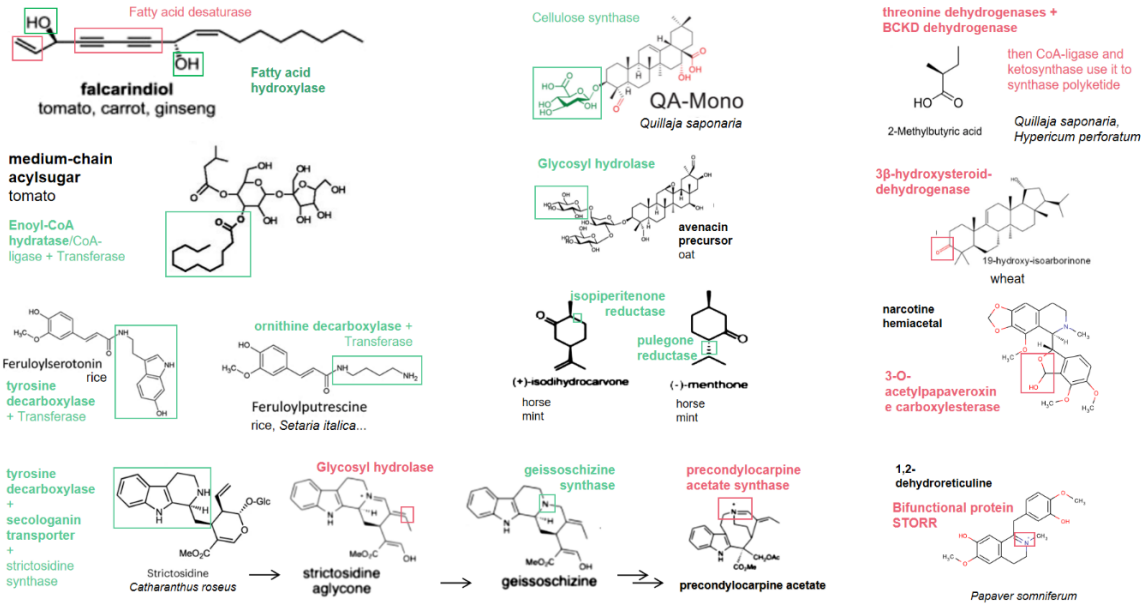

Figure 2| The domains of enzymes (in bold) from newly characterised clusters are used in making new rules. The modified sites on products are framed in the same colour as the enzymes.

Figure 1 shows that the mining is based on the HMMs and rules which set the range

for domain combinations of clusters found. With the increasing publication of plant genomes and the development of multi-omics research, over 50 new plant BGCs have been characterised since 2017 (Supporting file 1, SF1). Some of these are not in the plantiSMASH precalculated results or MIBiG. One reason for the former is that the 62 collected pHMMs could not identify some enzymes involved in these BGCs, such as the ornithine decarboxylase (ODC) and tyrosine decarboxylase in clusters of phenolamides in rice (Fang et al., 2021; Shen et al., 2021). Figure 2 shows the enzymes and their products from clusters considered when updating rules (SF1, SF2).

1.3.2. Identifying subgroups of enzyme families to predict substrate/product types

Predicting substrate/product types of enzymes is useful for distinguishing whether it is likely to be involved in secondary metabolism and thus filtering numerous clusters results, especially for large families` members usually found in clusters. One theory is that primary metabolic enzymes with strong substrate specificity evolved into secondary metabolic enzymes with weak substrate selectivity. This can explain why some homologous enzymes in a clade (subgroup) of an enzyme family tree can modify a group of substrates belonging to the same types or scaffolds.

For example, cellulose synthase superfamily-derived glycosyltransferases (CSyGTs/CslGs) are found in 7 species belonging to phylogenetically distant orders and producing oleanane-type triterpenoid (different saponins in different plant species). They form a clade (subgroup) within the cellulose-synthase-like family CslM branch on a phylogenetic tree of cellulose synthase family (CSLs) in which the CesA family specifically synthesizes cellulose. But all CSyGTs can transfer glucuronic acid from UDP-glucuronic acid to the C-3 position of medicagenic acid, the spinach saponins precursor (Chung et al., 2020; Jozwiak et al., 2020).

It is possible to predict the type of substrate/product of an enzyme by determining which subgroup it belongs to since more and more enzyme families are subgrouped based on the substrate/product types of members, such as CSLs (Chung et al., 2020; Jozwiak et al., 2020), UDP-glucuronosyltransferase (UGTs) Family (Louveau & Osbourn, 2019), short-chain dehydrogenases/reductases (SDRs) family (Moummou et al., 2012), and oxidosqualene cyclase (OSCs) family (unpublished work). Therefore, adding a subgroup identification module to plantiSMASH helps determine whether the cluster is involved in secondary metabolism and the type of product. For example, if a cluster contains a gene identified as CSyGT then the cluster is likely to be involved in the saponins synthesis. In addition, this functional similarity also implies the existence of more conserved sites in sequences from the same subgroup, which is the basis for constructing HMMs for identifying subgroups. This idea was already proposed in the first version of plantiSMASH (Kautsar et al., 2017).

1.3.3. Update the pre-calculated database to find more homologous cluster

To help discover the evolutionary relationships of gene clusters among different species so that the software can predict the generation and change of secondary metabolism pathways, ClusterBLAST module uses the DIAMOND algorithm and

genomic loci of clusters in the database which was built from 93 species previously (Kautsar et al., 2017). More high-quality genomes are available. Just NCBI (www.ncbi.nlm.nih.gov) records 386 plant reference genomes. Also because of new rules, the updated pre-calculated database will reveal more evolutionary clues of BGCs they found.

After the update, plantiSMASH will provide state-of-the-art plant BGCs prediction ability to accelerate mining unknown plants' biosynthetic pathways and metabolites. Clear biosynthetic pathways are required for developing more efficient microbial cell factory biosynthesis of important plant-derived drug compounds. It can be expected that after the release of the updated version, more people will be focused on the field of plant BGC prediction and the development of more strategies and modules.

2. Materials & Methods

2.1. Updating rules based on recently discovered plant BGCs

By reading papers in the list collaborator Dr. Charlotte Owen provided, I collated the domain compositions of 86 clusters, 30 of which were not BGCs (containing three different biosynthetic genes) (SF1, SF2). A total of 24 domains and 17 enzymes (Fig. 2) are not in the identification range of the HMMs library of plantiSMASH version 1. Based on the characteristics of these domains, I updated and made five BGC rules and customized two rules that can save clusters only containing two genes. SF5 records how I make and use new rules. 67 clusters are not in the MIBiG database or need to be updated (SF1). The genbank files (SF3) for reference are generated for 37 clusters, which have orthologous relationships to the remaining clusters. SF2 records which genomes are used.

Old and new default rules can be found in SF1. To compare the effect on the number of clusters after updating, run plantiSMASH with old and new rules on the *Arabidopsis thaliana* genome (GCF_000001735.4) and rice genome (GCF_001433935.1) files on NCBI. SF4 records the original result and the commands.

2.2. Building a subgroup identification module based on pplacer and hmmer

The module first collects protein sequences of each specific family from rules-saved clusters. For example, a gene with a Cellulose synthase domain pHMM match is considered to belong to CSLs and will be identified to which subgroup of this family it belongs. PlantiSMASH can identify subgroups of proteins from CSLs (Chung et al., 2020; Jozwiak et al., 2020), UDP-glucuronosyltransferase (UGTs) Family (Louveau & Osbourn, 2019), short-chain dehydrogenases/reductases (SDRs) family (Moummou et al., 2012), and oxidosqualene cyclase (OSCs) family (unpublished work). The files required to run pplacer and hmmer have been produced based on the subgroup classification literature. Details on how to make these files and customize this module are provided in SF5.

A subgroup is a clade of a family phylogenetic tree and the phylogenetic placement

tool pplacer can place the target protein sequence on a fixed reference tree (Matsen et al., 2010), which is quick and retains the subgroup classification information compared to rebuild the tree with targets. If the other members under the target parent node belong to the same subgroup, the target is considered to belong to the subgroup. In addition, GraPhlAn (Graphical Phylogenetic Analysis) (Asnicar et al., 2015) is used to generate a tree image of the placement result. The HMM of each subgroup used by hmmer scan (http://hmmer.org/) is made based on the full-length protein sequences of members in the subgroup. HMMs find matches based on the conserved positions of subgroup members, so they are also an efficient and simple method to identify subgroups of targets, especially those sequences with big differences. When the subgroup represented by the HMM with the highest match bitscore is consistent with the results from pplacer, it will be reported on the overview page that the target may have the same substrate type as those members of the subgroup. If it belongs to the product type that is predicted by the existence of core enzymes, it is specifically marked with * ahead to make it stand out.

For example, the target sequence is placed on the Glycosyl Transferase Family 2 phylogenetic tree which was used in the paper reporting the CSyGT subgroup (Chung et al., 2020; Jozwiak et al., 2020). If other members under the parent node of the target in the tree are CSyGT and the target matches the CSyGT HMM with the highest bitscore, the target is considered to be CSyGT and the product is predicted to be oleanane-type triterpenoid. At the same time, if the cluster contains a core enzyme for terpenoid synthesis, this prediction will also be marked with *. For situations where the results of these two tools are inconsistent, or the target is not placed in a subgroup, users need to make their judgment based on results shown on the webpage of each cluster. SF7 records input and raw results of testing this module.

2.3. Making the pre-calculated database based on NCBI Streptophyta genomes

The genefinding module was not used when using these well-annotated reference genome files. To save every homologous cluster of BGCs found later, set CD-HIT cut-off=0.9 and all rules minimum=2 as before (Kautsar et al., 2017) to save clusters that only have two genes with different domains when running plantiSMASH. In addition, I added a command, which can automatically update the results of this run to the database. Details and results are shown in SF8.

To verify whether ClusterBLAST can find homologous clusters of the recently reported phenolamide gene cluster, I tested it on 6 monocotyledon plants as this paper (Fang et al., 2021). The SF8 records the method and raw results.

3. Results & Discussion

3.1. The updated rules can find known BGCs and potential clusters

The new rules can find all newly reported BGCs (SF1). Comparing plantiSMASH results found on the *A. thaliana* and rice genomes using the old and new rules separately, shows that the new rules could find more BGC with few non-BGC (Table

1). The new rules contributed 16 new BGCs on *A. thaliana* and 24 new BGCs on rice including recently tested phenolamide clusters (Fang et al., 2021; Shen et al., 2021).

There are potential BGCs within these BGCs worth testing. For example, the potential BGC with 3 disease-resistance-related genes (based on annotations), rice cluster 40 is detected because the transporter domain can be detected, which is between naringenin-chalcone synthase and naringenin O-glucosyltransferase (Yang et al., 2021) tandem repeats might form a disease-resistance-related pathway.

Table 1 | New BGCs found on rice and Arabidopsis. The core domains pHMMs names with * were used before, but the clusters are found now because of updated non-core domains pHMMs.

Rice	*A. thaliana*	Cluster Type	Core domain	Enzyme name
1	3	Fatty acid	FA_hydroxylase	Fatty acid hydroxylase
2	2		ECH_2	Enoyl-CoA hydratase
3	0	Phenolamide	Pyridoxal_deC	Tyrosine decarboxylase
1	0		Orn_Arg_deC_N	Ornithine decarboxylase
4	3	Mate	Mate	Transporter
2	3	Alkaloid	BBE	Precondylocarpine acetate synthase
1	0		Pyridoxal_deC	Tyrosine decarboxylase
4	3	Saccharide	Glyco_hydro_1	Transglucosidase
1	1		Cellulose_synt	Cellulose synthase
0	1		UDPGT*	Glycosyltransferase
1	0	Polyketide	Chal_sti_synt_N*	Ketosynthase
1	0	Lignan	Dirigent*	Dirigent enzyme
1	0	Terpene	Terpene_synth*	Terpene synthase
2	0	Putative	Abhydrolase_3	3-O-acetylpapaveroxine carboxylesterase

3.2. The subgroup module predicts functions and highlights promising clusters

The product/substrate types based on subgroups are seen from the overview web page, with numbers representing the number of occurrences showing how many genes in the cluster might be involved. The * hallmark appears when associated with the product type based on the core enzyme prediction. This highlights promising clusters as if multiple genes in a cluster involve the same substrate types, this cluster is more likely to form a complete pathway and thus be easily verified. For example, as shown in Fig. 3A, the avenacin BGC of *Avena strigosa* can be easily found in 109 clusters, which contain an OSC producing triterpenes and UGTs using oleananes-type triterpenoid as substrate (Li et al., 2021).

The subgroup results for each gene can be seen in the subgroup column on the web page of each cluster (Fig. 3B); without "(E: value)", meaning that the two subgroup strategies (tree and HMMs) agree with each other. By clicking on the subgroup name, the developmental tree is shown, including the name of the node closest to the query (as in Fig. 3E). The function of members in the OSCs and UGTs tree is known. The functional common parts of members are annotated as subgroups. Therefore,

users can predict the more specific function of the query by searching for the function of the node closest to the query. For example, users can predict that the query close to GmSGT2/UGT73P2 (Shibuya et al., 2010) can also elongate the glycosidic branch on the subgroup D sharing substrate type, oleananes-type triterpenoid (Fig. 3C).

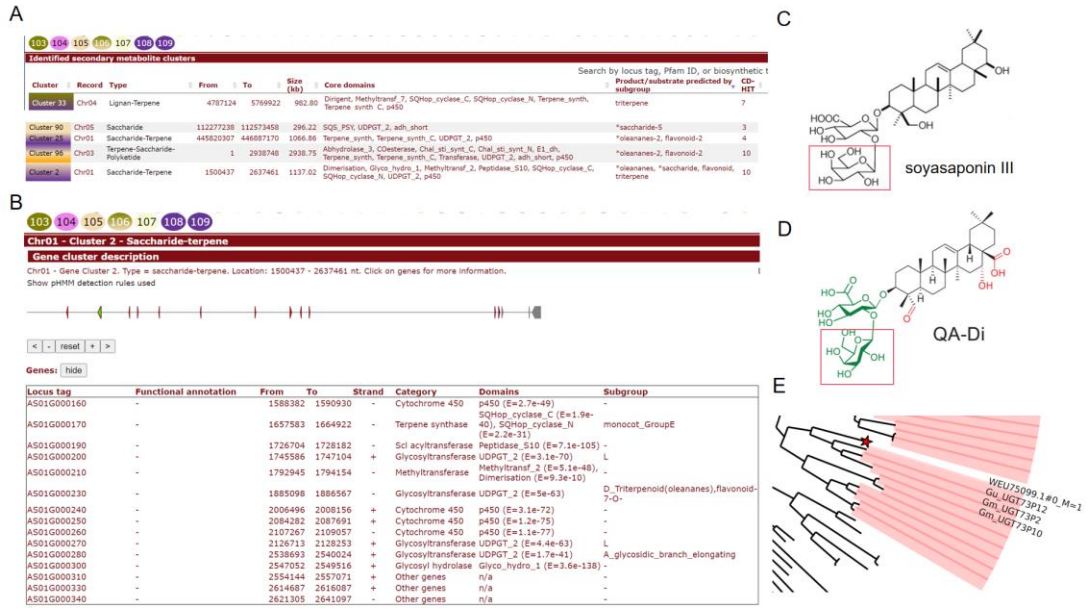

Figure 3| A. The overview web page of plantiSMASH results of *A. strigosa*; B. The cluster web page of the avenacin BGC; C, D. UGT73P2 and UGT73CU3 adding a sugar group (in the red frame) respectively; E. Part of the tree picture shows UGT73CU3 (GenBank, WEU75099.1) is placed in subgroup D (coloured in red) on the referent tree.

To test the module's ability, I used it to subgroup 20 UGT sequences with known functions but not used for building the module (SF7). Two subgroup ways agree on every tested UGT. Functions predictions are all consistent with the actual functions. In the 20 UGTs, UGT73CU3 (Reed et al., 2023) is close to UGT73P2 (Fig. 3E) and does elongate the glycosidic branch (Fig. 3D).

3.3. ClusterBLAST module finds homologous clusters from more types and species

There are over 5,000 genomic loci found in 165 plant species (Streptophyta) reference genomes from NCBI (SF8). These species belong to 10 different Classes, mostly Magnoliopsida (angiosperms).

To verify whether ClusterBLAST can find homologous clusters of the recently reported phenolamide gene cluster containing OsODCs and OsPHTs (Transferase) (Fang et al., 2021), I tested it on 6 monocotyledon plants as that paper did. Unsurprisingly, all homologous clusters (Fig. 4) were found. However, the gene composition of clusters is somewhat different from the paper reported, possibly due to NCBI genomes with annotations being different to the genomes the paper used.

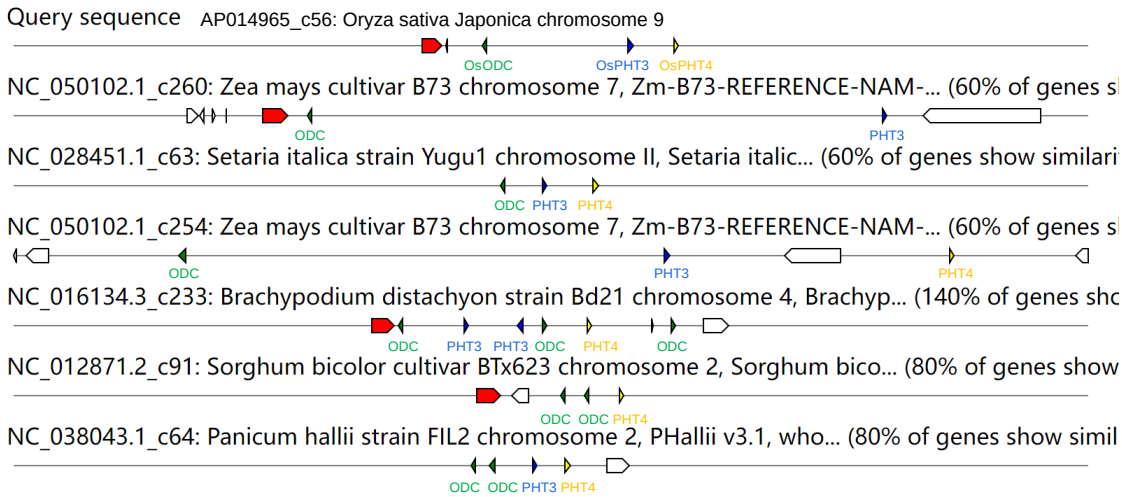

Figure 4| The homologous gene clusters of rice phenolamide found by the ClusterBLAST module.

4. Conclusions and future perspectives

The update is based on 17 enzymes (Fig. 2, SF1) and follows the old rule product classification system. New rules can find known BGCs and potential clusters, so help to update the MIBiG database. However, it is necessary to carefully study all the pHMMs used to rebuild the system to give more detailed and hierarchical type names (Table 2), such as making flavonoids rule by taking the Chalcone-flavanone isomerase (pHMM, Chalcone) as the core enzyme.

Table 2 | Classification of plant secondary metabolites. KEGG (www.genome.jp/brite/br08901)

(Sharma et al., 2022)	KEGG	MIBiG	plantiSMASH
Terpenoids	Terpenoids	Terpene	Terpene
			Sesterterpene
Cyanogenic Glycosides, Glucosinolates	\	Saccharide	Saccharide
Phenolics, Polyphenolics	Flavonoids	\	\
	Shikimate/Acetate-Malonate Pathway-Derived Compounds	\	\
	Phenylpropanoids	\	Lignan
	Polyketides	Polyketide	Polyketide
Nitrogen-Containing Compounds	Alkaloids	Alkaloid	Alkaloid, Phenolamide, Strictosidine Like
	Amino Acid Related Compounds	Nonribosomal Peptide	\
		Ribosomal Peptide	Cyclopeptide
\	\	\	Fatty Acid
\	Other	Other	Putative, Mate

The subgroup module uses hmmer and pplacer to divide CSLs, UGTs, OSCs, and SDRs

in the cluster into subgroups to predict possible substrates or products. It has been tested that 20 UGTs can be placed in a single subgroup on the tree and predicted with the right substrate type. At present, SDR subgroups cannot give a lot of information, and the subgroups of more families are also worth using, such as the subgroup in glycosyl hydrolase family 1 glycosylates substrate using acyl sugars as sugar donors (Orme et al., 2019). In the future, HMMs and reference trees will be updated or added via the option to add custom modules or to customize existing modules by users or developers. By that, more subgroups of more families will be able to be identified, providing more information for inferring the function of clusters. At the same time, combined with predicted information from other modules, clusters forming continuous pathways can be highlighted to help discover specific types of plant metabolites.

As tested in the phenolamide gene clusters of monocotyledons, the ClusterBLAST module can find the homologous clusters to give evolutionary insights without extra effort. There are other species genome data or different versions and varieties with good quality outside NCBI that can be used to build the pre-calculated database, some of which record newly discovered BGCs. However, plantiSMASH may encounter some problems (SF2) using gff3 for annotated genomic data as it is not a strict format. The error on parsing sequences with phase 2 now needs to be fixed (SF6).

This update also solves some errors when installing and using plantiSMASH. It is tested to work on Windows Subsystem for Linux (SF5). In the future, plantiSMASH will be upgraded to python3 and use more advanced algorithm packages to generate results faster, more stably, universally and accurately. For example, the recently published AlphaFold3 can accurately predict the interaction between proteins and small molecules (Abramson et al., 2024), which may become a universal tool to cooperate with for predicting enzyme substrate types in the future.

5. References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C. C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature 2024 630:8016*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Anarat-Cappillino, G., & Sattely, E. S. (2014). The chemical logic of plant natural product biosynthesis. *Current Opinion in Plant Biology*, *19*, 51–58. https://doi.org/10.1016/J.PBI.2014.03.007

Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., & Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, *2015*(6), e1029. https://doi.org/10.7717/PEERJ.1029/SUPP-3

Chung, S. Y., Seki, H., Fujisawa, Y., Shimoda, Y., Hiraga, S., Nomura, Y., Saito, K., Ishimoto, M., & Muranaka, T. (2020). A cellulose synthase-derived enzyme catalyses 3-O-glucuronosylation in saponin biosynthesis. *Nature Communications 2020 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-19399-0

Fang, H., Shen, S., Wang, D., Zhang, F., Zhang, C., Wang, Z., Zhou, Q., Wang, R., Tao, H., He, F., Yang, C., Peng, M., Jing, X., Hao, Z., Liu, X., Luo, J., Wang, G. L., & Ning, Y. (2021). A monocot-specific hydroxycinnamoylputrescine gene cluster contributes to immunity and cell death in rice. *Science Bulletin*, *66*(23), 2381–2393. https://doi.org/10.1016/J.SCIB.2021.06.014

Jozwiak, A., Sonawane, P. D., Panda, S., Garagounis, C., Papadopoulou, K. K., Abebie, B., Massalha, H., Almekias-Siegl, E., Scherf, T., & Aharoni, A. (2020). Plant terpenoid metabolism co-opts a component of the cell wall biosynthesis machinery. *Nature Chemical Biology 2020 16:7*, *16*(7), 740–748. https://doi.org/10.1038/s41589-020-0541-x

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*(W1), W55–W63. https://doi.org/10.1093/NAR/GKX305

Kautsar, S. A., Suarez Duran, H. G., & Medema, M. H. (2018). Genomic identification and analysis of specialized metabolite biosynthetic gene clusters in plants using plantiSMASH. *Methods in Molecular Biology*, *1795*, 173–188. https://doi.org/10.1007/978-1-4939-7874-8_15/FIGURES/8

Li, Y., Leveau, A., Zhao, Q., Feng, Q., Lu, H., Miao, J., Xue, Z., Martin, A. C., Wegel, E., Wang, J., Orme, A., Rey, M. D., Karafiátová, M., Vrána, J., Steuernagel, B., Joynson, R., Owen, C., Reed, J., Louveau, T., … Osbourn, A. (2021). Subtelomeric assembly of a multi-gene pathway for antimicrobial defense compounds in cereals. *Nature Communications 2021 12:1*, *12*(1), 1–13. https://doi.org/10.1038/s41467-021-22920-8

Liu, J. C. T., De La Peña, R., Tocol, C., & Sattely, E. S. (2024). Reconstitution of early paclitaxel biosynthetic network. *Nature Communications 2024 15:1*, *15*(1), 1–11. https://doi.org/10.1038/s41467-024-45574-8

Louveau, T., & Osbourn, A. (2019). The Sweet Side of Plant-Specialized Metabolism. *Cold Spring Harbor Perspectives in Biology*, *11*(12), a034744. https://doi.org/10.1101/CSHPERSPECT.A034744

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538/FIGURES/10

Moummou, H., Kallberg, Y., Tonfack, L. B., Persson, B., & van der Rest, B. (2012). The Plant Short-Chain

Dehydrogenase (SDR) superfamily: Genome-wide inventory and diversification patterns. *BMC Plant Biology*, *12*(1), 1–17. https://doi.org/10.1186/1471-2229-12-219/FIGURES/7

Orme, A., Louveau, T., Stephenson, M. J., Appelhagen, I., Melton, R., Cheema, J., Li, Y., Zhao, Q., Zhang, L., Fan, D., Tian, Q., Vickerstaff, R. J., Langdon, T., Han, B., & Osbourn, A. (2019). A noncanonical vacuolar sugar transferase required for biosynthesis of antimicrobial defense compounds in oat. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(52), 27105–27114. https://doi.org/10.1073/PNAS.1914652116/-/DCSUPPLEMENTAL

Owen, C., Patron, N. J., Huang, A., & Osbourn, A. (2017). Harnessing plant metabolic diversity. *Current Opinion in Chemical Biology*, *40*, 24–30. https://doi.org/10.1016/J.CBPA.2017.04.015

Reed, J., Orme, A., El-Demerdash, A., Owen, C., Martin, L. B. B., Misra, R. C., Kikuchi, S., Rejzek, M., Martin, A. C., Harkess, A., Leebens-Mack, J., Louveau, T., Stephenson, M. J., & Osbourn, A. (2023). Elucidation of the pathway for biosynthesis of saponin adjuvants from the soapbark tree. *Science*, *379*(6638). https://doi.org/10.1126/SCIENCE.ADF3727/SUPPL_FILE/SCIENCE.ADF3727_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Sharma, A., Sharma, S., Kumar, A., Kumar, V., & Sharma, A. K. (2022). Plant Secondary Metabolites: An Introduction of Their Chemistry and Biological Significance with Physicochemical Aspect. *Plant Secondary Metabolites: Physico-Chemical Properties and Therapeutic Applications*, 1–45. https://doi.org/10.1007/978-981-16-4779-6_1/TABLES/3

Shen, S., Peng, M., Fang, H., Wang, Z., Zhou, S., Jing, X., Zhang, M., Yang, C., Guo, H., Li, Y., Lei, L., Shi, Y., Sun, Y., Liu, X., Xu, C., Tohge, T., Yuan, M., Fernie, A. R., Ning, Y., … Luo, J. (2021). An Oryza-specific hydroxycinnamoyl tyramine gene cluster contributes to enhanced disease resistance. *Science Bulletin*, *66*(23), 2369–2380. https://doi.org/10.1016/J.SCIB.2021.03.015

Shibuya, M., Nishimura, K., Yasuyama, N., & Ebizuka, Y. (2010). Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in Glycine max. *FEBS Letters*, *584*(11), 2258–2264. https://doi.org/10.1016/J.FEBSLET.2010.03.037

Terlouw, B. R., Blin, K., Navarro-Muñoz, J. C., Avalon, N. E., Chevrette, M. G., Egbert, S., Lee, S., Meijer, D., Recchia, M. J. J., Reitz, Z. L., van Santen, J. A., Selem-Mojica, N., Tørring, T., Zaroubi, L., Alanjary, M., Aleti, G., Aguilar, C., Al-Salihi, S. A. A., Augustijn, H. E., … Medema, M. H. (2023). MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, *51*(D1), D603. https://doi.org/10.1093/NAR/GKAC1049

Yang, Z., Li, N., Kitano, T., Li, P., Spindel, J. E., Wang, L., Bai, G., Xiao, Y., McCouch, S. R., Ishihara, A., Zhang, J., Yang, X., Chen, Z., Wei, J., Ge, H., Jander, G., & Yan, J. (2021). Genetic mapping identifies a rice naringenin O-glucosyltransferase that influences insect resistance. *The Plant Journal*, *106*(5), 1401–1413. https://doi.org/10.1111/TPJ.15244