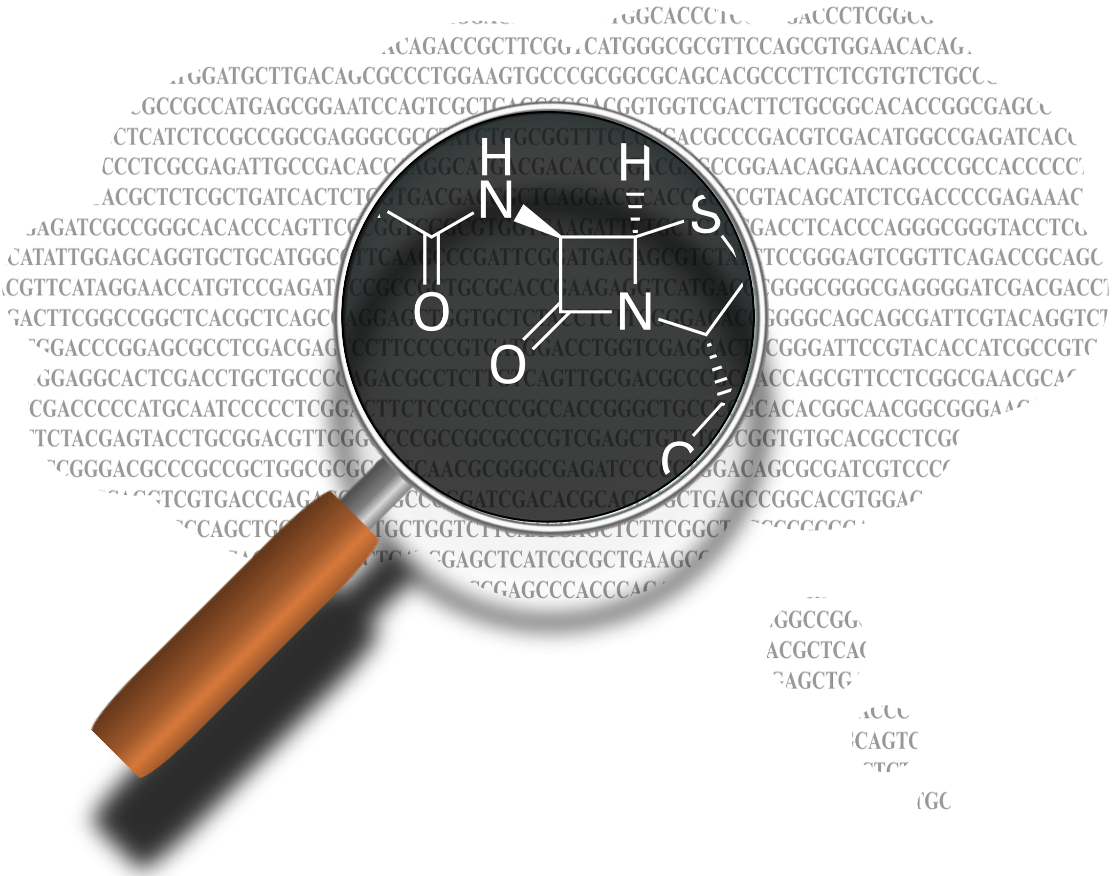
Mapping *Natural Product* Diversity through *Genomics*



Satria A. Kautsar

Propositions

1. (Meta)genomics, not metabolomics, is the most comprehensive approach to chart natural product biosynthetic diversity.
   (this thesis)

2. Highly cultivated microbial taxa still hold sufficient untapped chemical novelty for drug discovery efforts, which lessens the urgency of researching uncultivated microbes.
   (this thesis)

3. Life science journals should organize professional quality assessment of any software or code they publish.

4. In a world without ethics and privacy concerns, the use of intrusive surveillance systems would be the only acceptable method for the social sciences.

5. Public sector integrity is the most important goal for any developing country to transform into a developed nation.

6. Overconsumption, not population growth, is the main cause of the sustainability crisis.

Propositions belonging to the thesis, entitled

Mapping Natural Product Diversity through Genomics

Satria Ardhe Kautsar

Wageningen, 25 May 2021

Mapping Natural Product Diversity through Genomics

Satria A. Kautsar

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promotor
Dr M.H. Medema
Assistant Professor, Bioinformatics
Wageningen University & Research

Other members
Prof. Dr M.E. Schranz, Wageningen University & Research
Dr J. de Ridder, University Medical Center Utrecht
Prof. Dr G.P. van Wezel, Leiden University
Dr. M.S.E. Donia, Princeton University, United States

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Mapping Natural Product Diversity through Genomics

Thesis
Submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 25 May 2021
at 4 p.m. in the Aula.

Satria A Kautsar
Mapping natural product diversity through genomics, 204 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)
With references, with summaries in English, Dutch, and Indonesian

ISBN 978-94-6395-756-4
DOI https://doi.org/10.18174/543999

Table of Contents

Chapter 1

General Introduction

Plants, bacteria, and fungi have historically been an important source of useful natural products. As more organisms were cultivated to search for possible new drugs, a plateau appeared to be reached at some point, as the (re-)discovery rate of known compounds became higher than the rate of discovering unknown ones. However, with advances in DNA sequencing technology, it has become evident that most of an organism's secondary metabolic pathways actually remain dormant during regular cultivation processes. Recent years have seen an explosion of genomic data stored in public databases, providing an opportunity (and challenge) to leverage them for charting and measuring the true potential of this hidden side of the "natural product iceberg". In this opening chapter, I outline the state of the field, and discuss how this has resulted in the challenge that I address in this thesis.

1.1. The importance of natural products

Our life is full of biochemical wonders: from the sip of coffee that wakes us up in the morning, the pleasant smell of "earth" after the rain, to the vast array of clinically important drugs that saved billions of lives throughout centuries (Figure 1.1). Although many commercialized drugs were often synthetically designed and optimized to achieve the efficacy required for clinical usage, a large majority of their origins can actually be traced back to nature. Between 1981 and 2019, it has been estimated that more than 75% of the 1,881 newly approved drugs were biologically derived or inspired by natural products (NPs) [1]. Some notable examples of these NP-derived compounds are the cholesterol-lowering statins and the historically important beta-lactam antibiotics like penicillin and thienamycin (a "model" carbapenem). The first two were originally isolated from *Penicillium* fungi, while the latter was derived from a compound produced by a soil-borne *Streptomyces* species.

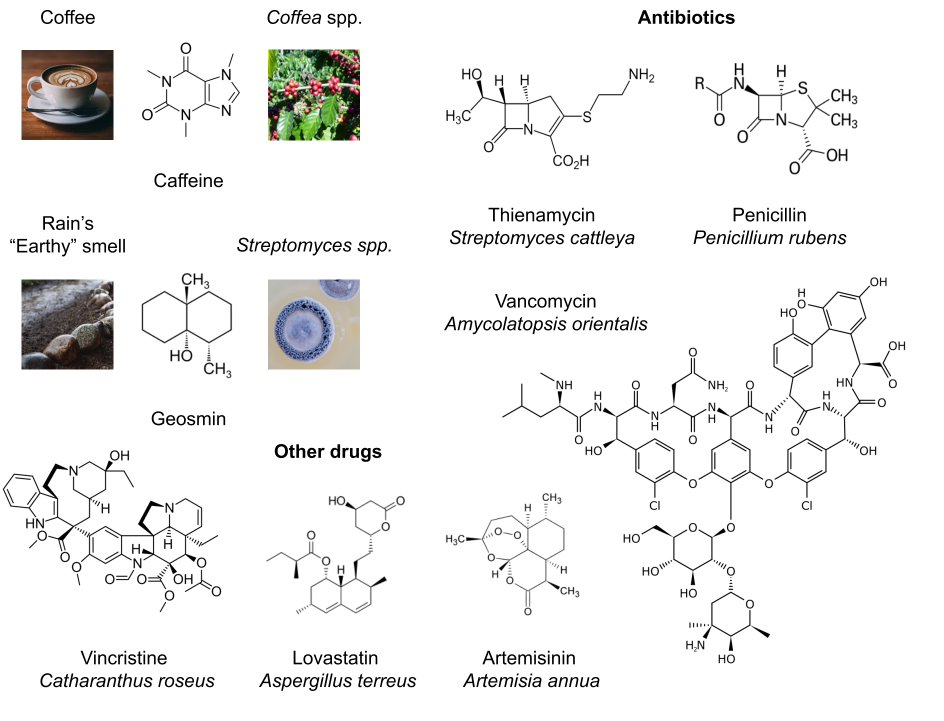


Figure 1.1. Examples of natural products we encounter in our daily life and their producer organisms; examples include clinically relevant drugs that were sourced from or inspired by natural products.

In 2020, the world was suddenly crippled by a global pandemic caused by the highly infectious SARS-CoV-2 coronavirus, which affected more than 51.5 million people and killed at least 1.27 million within the first 11 months of its onset. Yet, at the same time, the problem of antimicrobial resistance (AMR) has continued to snowball without getting as much attention. In fact, AMR seems to get exacerbated even further by the pandemic, first because of funding and resource redirection and second due to the prevalence of pre-emptive antibiotics prescription for COVID-19 patients [2]. When left unaddressed, AMR may cause a catastrophic number of ten million annual deaths in 2050, which is a more than tenfold increase over the 700,000 deaths recorded in 2016 [3]. Although solving the problem would ultimately require a radical change in antibiotic development pipelines and clinical usage policies [4], there is a fundamental need to increase our arsenal of potent but non-cytotoxic novel antimicrobial NP families that we can take into the drug discovery pipeline [5].

1.2. A paradox in the observable NP diversity

Between the 1940s and 1960s, drug discovery was driven mainly by a “top-down” approach: microbes were systematically isolated and screened for their antibiotic activity against known pathogens, and compounds were then purified and fermented on a large scale [6]. By the early 1970s, it was widely assumed that after the discovery of around 29 clinically approved antibiotic families, nature’s repertoire of usable antimicrobial compounds was exhausted, as mostly known compounds were rediscovered in isolation and screening efforts [7,8]. Although more recent innovative ventures beyond the traditional soil-derived *Streptomyces* platforms managed to reveal a new array of potential compounds [9,10], we are still nowhere close to reviving the “golden era” of antibiotics discovery of half a century ago.

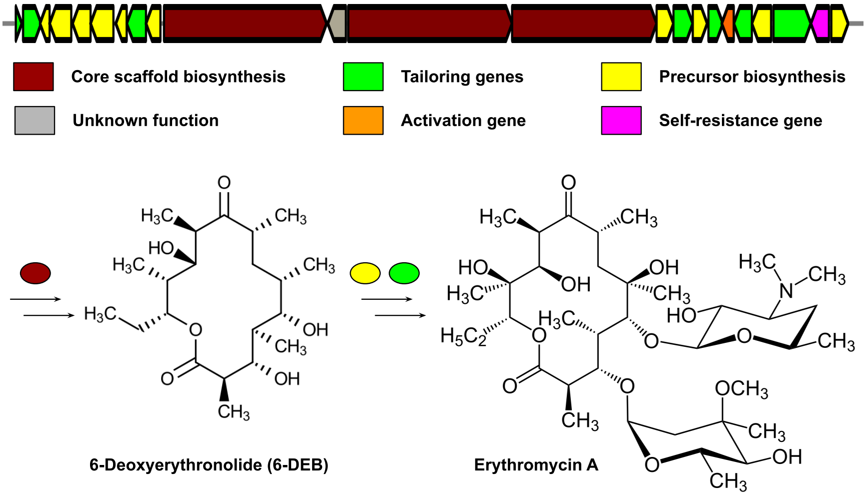
Considering the scale of biodiversity in the microbial world, we would not expect such a plateau to happen so early. We can find an abundance of microbes even in the most uninhabitable places, from highly radioactive nuclear contaminated sites [11], high temperature hydrothermal springs [12], down to extreme pressure in the deep ocean [13]. In these environments, microbial species will not only need to adapt to the environmental factors (i.e., temperature/pressure, nutrient limitation) but also continuously develop and maintain competitive advantages against each other. In the spirit of Darwinian evolution, the variety of conditions across biospheres would create high-level genetic diversity, translating into a similarly high-level NP diversity, even within a single genus such as *Streptomyces*. Although this idea has been formed as early as 1931 [14], a clear answer to this paradox finally came by the early 2000s: most of a microbe’s NP-encoding secondary metabolic pathways were actually never expressed throughout typical cultivation processes [15].

1.3. Genomics and the “bottom-up” approach in drug discovery

After the first bacterial shotgun sequencing project targeting the ~1.8 Mbp (million base pairs) genome of the *Haemophilus influenzae Rd* was completed in 1995 [16], many similar efforts steadily followed suit: five years later, four other bacterial genome assemblies had been deposited in the NCBI GenBank database [17], all disease-causing Gram-negative pathogens. Around 2002, the NP research community caught up with the complete genome sequences of *Streptomyces avermitilis* ATCC 31267 [18,19] and *Streptomyces coelicolor* A3(2) [20], revealing the two largest (~8.6 and 9 MB) bacterial genomes ever recorded at the time, which also harbored the largest number of genes (>7,800 predicted genes).

Box 1.1. Biosynthetic gene clusters (BGCs)

All living organisms possess long stretches (10^5-10^{11} base pairs / bp) of deoxyribonucleic-acid (DNA) in their cells, which act as a blueprint for the organism's biological development and functioning. Canonically termed the genome, the DNA molecules contain enzyme-coding genes that serve as a basis for every essential metabolic pathway for an organism to fully function and thrive in its habitat, which include both primary (directly involved in growth and reproduction) and secondary (involved in ecological function) metabolism. The latter is responsible for the majority of microbial natural products we have known today. Genes that encode these pathways are often co-located together in the genome, forming what are commonly known as Biosynthetic Gene Clusters, or BGCs in short (in some cases, the term may be interchangeable with microbial operons, which are cluster of genes controlled by a single *cis*-regulatory element). A secondary metabolic BGC (referred to simply as "BGC" from now on) typically includes enzymatic (biosynthetic) genes and other relevant genes, such as those encoding transporters and transcriptional regulators.



A classic example of a complex secondary metabolic BGC is the erythromycin synthase locus (pictured) in the genome of *Streptomyces erythraeus* (later re-assigned as *Saccharopolyspora erythraea*) [21,22]. The ~54,000 bp locus includes a multi-domain set of genes (dark brown) that encodes a giant synthase enzyme to build the core polyketide scaffold (6-DEB) from a starter propionyl-coenzyme A (CoA) unit and six (2S)-methylmalonyl-CoA extender units. Multiple tailoring reactions then take place (catalyzed by enzymes shown in green, using precursors provided from enzymes depicted in yellow) to achieve the final double-glycosylated structure of erythromycin A [23]. The BGC also harbors a methyltransferase gene, ErmE (light purple), to confer self-resistance against the antimicrobial activity of the produced compound [23].

More importantly, an upward of twenty uncharacterized secondary metabolic gene clusters (BGCs) putatively involved in stress-related NP production were discovered within each of these genomes (see Box 1.1 for a detailed explanation on BGCs). Even with the large-scale manual labor required to perform the Sanger-based shotgun sequencing during these early days of genomics, many complete bacterial genomes became available at a rapid pace: by 2005, 151 complete bacterial genomes had been submitted to NCBI, thirteen of which were actinobacterial. Using genome mining to measure the secondary metabolic potential of the actinobacteria, it was estimated that on average, more than 80-90% of BGCs could not be matched to any observable NP in the lab [15]. This propelled a new paradigm in NP discovery, known as the "bottom-up" approach.

In this genome-guided discovery, after an initial marker-based amplicon screening/redundancy filtering [24], potential producer microbes are selected based on their expected genetic (and chemical) novelty. Detailed information about each sequenced genome can then be used to guide the activation and subsequent characterization of BGCs within their native host. Activation can be based on a simple approach such as the manipulation of growth conditions [25,26] or a more sophisticated one, e.g., using genetic engineering techniques to knock out a repressor gene [27,28] or over-express positive regulatory genes [28,29] associated with the BGC. Alternatively, an entire BGC can be exported to [29–31] or synthesized in [32,33] an engineered host that is more suitable for the expression of the encoded secondary metabolic pathways.

Although bioactivity assays and clinical testing are ultimately still needed to confirm the newly expressed compounds' viability as drugs and antibiotics, this new paradigm seems to have been successful in introducing many compound classes and producer organisms that were overlooked or missed before [34]. One of the earliest applications of this approach was in 2000, when Challis et al. [35] managed to predict the biosynthesis of coelichelin, a cryptic nonribosomal peptide (NRP) siderophore from *Streptomyces coelicolor* and subsequently managed to induce the compound's production by growing the bacteria in an iron-deficient medium [26]. More recently, Shen et al. [36] performed a large-scale genome mining of 3,400 actinomycetes and managed to unearth new compound analogues and a super producer strain of lidamycin (also known as C-1027 enediyne), one of the most potent cytotoxic NPs ever discovered in bacteria [37–39]. In this current age of massive-scale genomics and metabolomics, such a genome-oriented approach is expected to be a major driving force behind the revitalization of antibiotic developments and discoveries in the coming decades [40,41].

1.4. Entering the era of "big omics"

Sparked by the first introduction of an automated and highly-parallel sequencing-by-synthesis technology in 2005 by 454 Life Sciences [42] along with other similar approaches [43,44], sequencing devices have improved at an exceptionally rapid pace, allowing much faster and cheaper sequencing at a massive scale [45]. This has resulted in an explosion of genomic data: by November 2020, the NCBI GenBank database had accumulated more than 770,000 bacterial genomes (~25,000 of which are complete assemblies). Additionally, although more steadily developed over the course of four to five decades [46], improvements in analytical chemistry instrumentation has led to a massive increase in characterized compounds, with chemical databases like PubChem now holding information on more than 111 million unique molecules [47].

As the number of sequenced bacterial genomes significantly expanded over the years, so has taxonomic coverage (Figure 1.2). Between 2004 and 2019, there was a >200 fold increase (from 118 to 24,589) in the number of unique species with a genome deposited in NCBI RefSeq [48]. Moreover, in addition to several well-known single-cell amplified genome (SAG) [49] recovery studies [50–55], the ever-increasing number of large-scale shotgun sequencing projects to generate metagenome-assembled genomes (MAGs) have significantly expanded the microbial tree of life beyond cultivable organisms [56–63]. Although initially limited to draft-level and error-prone assemblies, with proper standard and quality control, complete MAGs can now easily be generated to allow for a reliable metabolic and evolutionary study of the organisms [64,65].

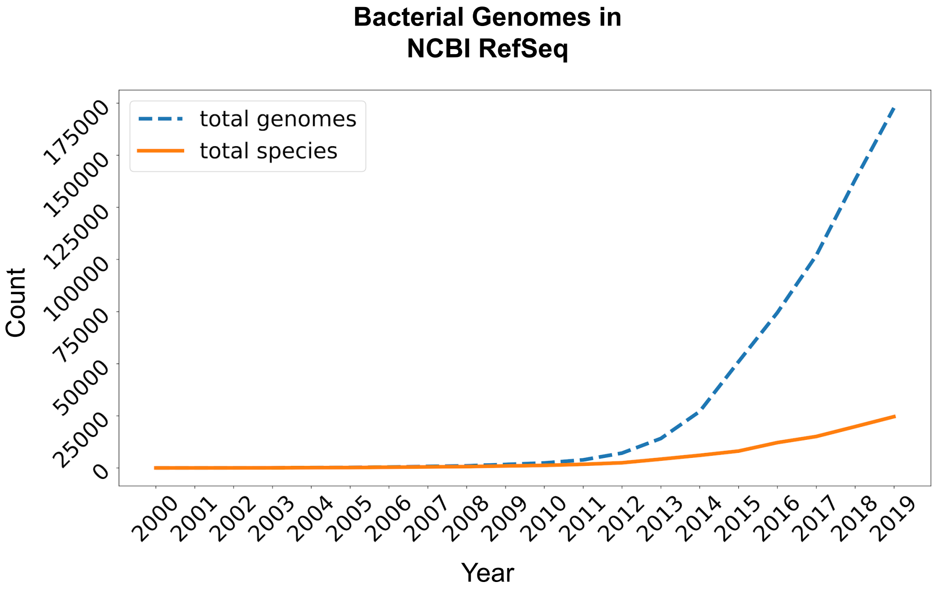


Figure 1.2. Growth of bacterial genome assembly data and the cumulative species counts deposited in NCBI RefSeq from 2000 to 2019. Original image, data, and scripts for this figure are available via figshare [66].

With such abundant biological information comes an excellent opportunity to answer both fundamental and practical questions on NP discovery. As BGCs can be thought of as a direct proxy to an organism's secondary metabolic pathways, analyzing them may offer us a glimpse at the largely hidden chemical space of microbial life. Questions such as: ***”How to prioritize our genome sequencing***

and BGC characterization efforts? Which genera, species, and strains harbor the greatest potential for novel discoveries? How much NP diversity is yet to be explored? Have we really exhausted the NP repertoire of cultivable bacteria? and How many new compound families may we expect from targeting uncultured microbes?” can theoretically be addressed by a comprehensive inspection of all organism (and metagenome-assembled) genomes at the same time, i.e. by performing a global meta-analysis. To do this, it is essential to develop scalable bioinformatics tools and databases that can efficiently generate, process, store and analyze the collective terabyte-to-petabyte scale data.

1.5. Genomic approaches to chart NP diversity

Bioinformatics, a recently emerging [67] field at the intersection of computer science and biology, has played a central role in transforming biological data (such as protein sequences) into actionable insights and knowledge. One major subfield in bioinformatics is computational genomics, which covers the development and usage of software tools to process or analyze genomic sequences. Ever since the first sequenced actinobacterial genome became available in 2002, bioinformatic-driven genome mining has become central to (microbial) NP discovery efforts, with (open source) tools and databases continuing to play pivotal roles that enable new ways to mine information from the ever-increasing amounts of data.

1.5.1. Development of genome mining tools for NP discovery

During the early days of NP genomics, annotations and predictions of BGCs largely relied upon BLAST [68] searches in protein/domain databases such as SwissProt [69] or NCBI’s Genbank [17] and the Conserved Domain Database (CDD) [70] to find core biosynthetic genes and capture their flanking regions. Several web services like NRPS-PKS [71] and ASMPKS [72] offered automated analyses to help perform this manually laborious task. Starting from early 2007, with the publication of a new method by Minowa et al. [73], BLAST-based approaches were slowly phased out in favor of more powerful, probability-based profile Hidden Markov Model (pHMM) protein domain detection [74], especially after the introduction of a new matching algorithm with speed comparable to BLAST [75]. Within the space of five years (2007-2012), many pHMM-based BGC prediction tools were developed [76–81], most of them focusing on the identification and structure prediction of assembly-like type I polyketide synthase (T1PKS) and nonribosomal peptide synthetase (NRPS) BGCs.

In 2011, Medema et al. [82] published antiSMASH, a new BGC prediction tool that, for the first time, allowed the comprehensive identification and analysis of a wide range of known secondary metabolic classes (polyketides, NRPs, terpenes, siderophores, bacteriocins, and at least seven others). A defining feature of this tool is the computational annotation of not only core scaffold genes but also putative tailoring and regulatory elements found within the BGCs. Intended for both regular and advanced users, antiSMASH was offered both as command-line software and wrapped as a web service. Several other prediction tools came after, which offered complementary feature sets [83] or focused analysis on specific BGC classes [84–87] and taxonomy [88,89]. However, due to its comprehensive coverage, continuous updates and user-friendly web service, antiSMASH enjoys the most widespread usage, being cited >4,800 times for its five published versions [82,90–93] and having processed a total of >810,000 online submissions as of November 2020.

Although already a leap forward in terms of coverage, antiSMASH was mainly designed for a high-confidence prediction of known (i.e., experimentally characterized) bacterial/fungal BGC classes, which did not cover all (relevant) NPs [94]. For example, while BGC prediction and analysis has become the default approach in microbial-NP discovery, it has never been applied for plant-based NPs, whose medicinal usage have historical traces as old as the civilization itself [95]. Aside from the significantly higher difficulty in sequencing and assembling their huge (hundreds to billions of base pairs) and complex (many repetitive sequences, often polyploid) genomes, plants were never known to possess any operon-like arrangements that are typical to microbial secondary metabolism. Recently however, at least 21 NP-encoding BGC-like loci were characterized in more than a dozen plant genomes [96], which presents an opportunity and a challenge (addressed in this thesis) to develop bioinformatics methods to efficiently mine BGCs in the increasing number of high-quality plant genomes [97,98]. As we sought to map the global NP diversity with genomics, extending the scope of our BGC prediction tools toward these unexplored metabolic classes will give us the most complete picture. This leads into the first specific question of this thesis: ***"How can we increase the detection coverage of current BGC prediction tools and comprehensively cover secondary metabolism across the tree of life (in this case, plant BGCs)?"***.

1.5.2. Increasing BGC data volumes require a structured storage system

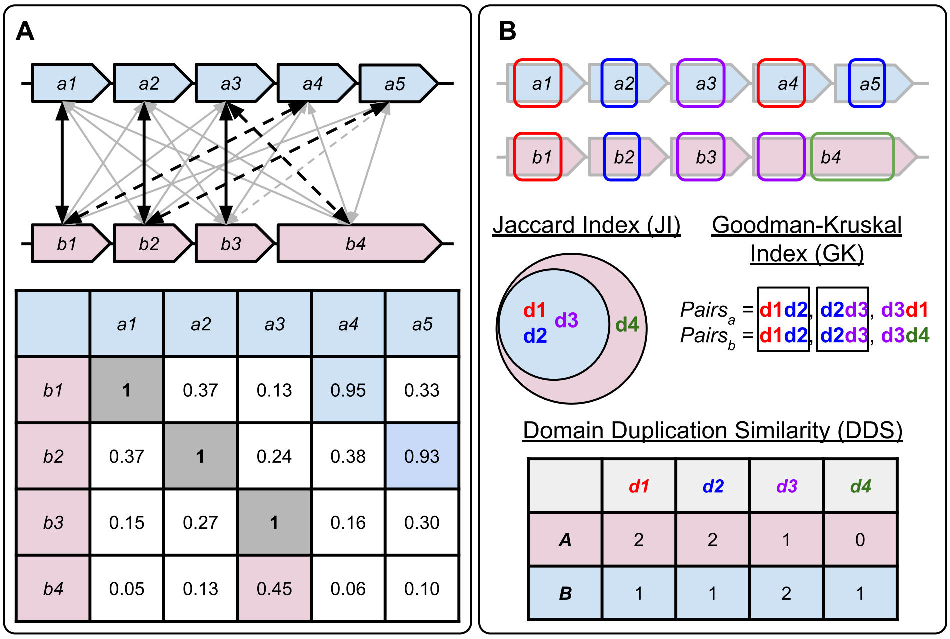
Back when researchers just started to mine genomes for NP discovery, reference databases of known (i.e., experimentally validated) NRPS/PKS/RiPP genes [71,81,99–103] became essential to facilitate mining and functional prediction of their BGCs. Shortly after the introduction of antiSMASH, several BGC-focused databases went online, such as ClusterMine360 [104], DoBISCUIT [105], and StreptomeDB [106]. These resources can potentially be used alongside antiSMASH for rapid dereplication and characterization of newly predicted BGCs, but direct postprocessing was complicated by the lack of a standardized file format. In 2015, the Minimum Information about a Biosynthetic Gene cluster (MIBiG) data standard was announced [107], which proposed not only a data schema for the annotation of BGCs but also a reference database for 1,170 manually curated BGCs that can be queried using antiSMASH's "KnownClusterBlast" module [91]. A year later, the antiSMASH database (antiSMASH-DB) was also published [108], containing 22,292 predicted BGCs from 3,907 dereplicated, reference-quality bacterial genomes from NCBI, allowing direct BGC homology searches using antiSMASH's "ClusterBlast" module.

Even with the simple and static architecture that it has, the initial version of MIBiG database was widely adopted by the community, with the paper having been cited more than 450 times within the span of five years. However, the rate at which new BGCs are being characterized means that the database should be designed to keep up with the rapidly growing data demand and stay relevant as a reference source for BGC-related analyses. This requires not only a scalable software architecture, but also a robust and streamlined data input (i.e., annotation) and database maintenance. Moreover, there is an increasing demand for FAIR databases [109], which dictates that the stored data should be Findable, Accessible, Interoperable and Reusable based on information-rich metadata that allows flexible yet powerful programmatic access into the database. In terms of a global-scale BGC analysis, these BGCs with known compounds serve to infer the function of related unknown BGCs, which brings us to the next specific question of the thesis: ***"How to develop and maintain a high-quality reference database that is scalable and comprehensive enough to handle the full scale of a global BGC analysis?"***

1.5.3. Early attempts in mapping global BGC diversity

With comprehensive prediction tools and programmatically accessible databases in place, we may start to try answering the final question of this thesis: ***"How to leverage the massive amounts of available genomic data (in databases) to map global NP diversity?"***. However, while the 3,907 dereplicated reference genomes in antiSMASH-DB served well to capture highly conserved BGCs, e.g., at genus level, they do not cover the true global strain- and species-level diversity of, for example, the 770,000 bacterial genomes in NCBI GenBank (of which ~200,000 are in RefSeq [48]). This is important, as several studies [110] have pointed out that there can be a big discrepancy between phylogenetic and secondary metabolic diversity, especially in well-known NP-rich genera like *Streptomyces* [111]. A major reason underlying the limited scope of antiSMASH-DBwas the expensive gene-to-gene sequence similarity calculation required by its query algorithm ClusterBlast (Figure 1.3A); such pairwise matching of millions of BGCs causes a large computational bottleneck in the antiSMASH pipeline. Furthermore, while this method is rather effective at detecting a high level of homology between pairs of BGCs from related strains, it may fall short when used to compare the higher-level hierarchy of Gene Cluster Families (GCFs), for example on homologous BGCs from distantly related species [112]. This makes it difficult to fully cover the multi-layered breadth and depth of NP diversity.

One promising approach that can be used as an alternative for the alignment-based BGC comparisons was originally proposed by Lin et al. in 2006 [114], who introduced several measures (termed "Jaccard Index (JI)", "Goodman-Kruskal Index (GK)", and "Domain Duplication Similarity (DDS)") based on the architectural comparison of Pfam [113] domains in order to improve the functional predictions of promiscuous multi-domain proteins. By treating an entire BGC as an equivalent of a giant multi-domain protein (Figure 1.3B), Fischbach and coworkers [94] demonstrated that this method can readily be used to perform BGC homology analysis, grouping similar BGCs into GCFs. However, while it was successful at charting the global biosynthetic diversity of 11,456 BGCs predicted from 1,154 genomes available at the time, scaling the algorithm to the level of millions of BGCs is difficult if not impossible.


A
a1
a2
a3
a4
a5
b1
b2
b3
b4
B
a1
a2
a3
a4
a5
b1
b2
b3
b4
Jaccard Index (JI)
Goodman-Kruskal Index (GK)
d1
d2
d3
d4
Pairs$_a$ = d1d2, d2d3, d3d1
Pairs$_b$ = d1d2, d2d3, d3d4
Domain Duplication Similarity (DDS)

	a1	a2	a3	a4	a5
b1	**1**	0.37	0.13	0.95	0.33
b2	0.37	**1**	0.24	0.38	0.93
b3	0.15	0.27	**1**	0.16	0.30
b4	0.05	0.13	0.45	0.06	0.10

	d1	d2	d3	d4
A	2	2	1	0
B	1	1	2	1

Figure 1.3. A. Multiple sequence similarity approach to compare BGCs. In this case, an all-to-all BLAST-based comparison between the genes (a1-a5, b1-b4) is taken and the cumulative percentage identity of best-matching gene pairs (solid arrows for reciprocal best matches) is calculated to arrive at a final BGC similarity value. B. Domain architecture similarity approach to compare BGCs. First, a full pHMM scan of all genes within the BGCs is performed to annotate protein domains (d1-d4) according to, e.g., the Pfam database [113]. Each detected domain can then be used as "tokens" for class absence-presence comparison (JI), synteny (GK), and domain copy number variation (DDS), each outputting distance values from 0 to 1. A weighted composite value of the three measurements is then calculated to determine the overall BGC similarity level.

The original clustering strategy used by these algorithms relies on a full pairwise distance comparison between BGCs (or genes), which scales quadratically and therefore may not work well on such a massive dataset. As an anecdotal example, if clustering a thousand BGCs takes an hour, clustering ten thousand will take 10^2 = 100 hours, a somewhat manageable runtime, but clustering a million will take $1{,}000^2$ = 1,000,000 hours, or roughly 114 years; not to mention the similarly scaled memory requirements to perform such an operation. Moreover, as Pfam-derived domains can be too broad to cover the specificity of some protein families (e.g., the ketosynthase domain), researchers in some cases resort to adopting alignment-based approaches [115], which adds significant runtime costs. To address this challenge, this thesis offers a new computational method that employs a radically different strategy, allowing a near-linear BGC homology calculation and ultimately facilitating truly global BGC diversity analysis.

1.6. Scope and outline of the thesis

The chapters in this thesis center on answering the aforementioned questions. The main chapters (**chapters 2-8**) are split into two major themes: *“bigger data, better data(bases)”* and *“charting NP diversity through genomics”*. In **chapter 9**, I provide a general discussion on the significance of the databases and methods proposed and how they have moved the field forward to answer the questions. I also discuss some future perspectives on how we can build upon these methods and other relevant emerging technological/scientific developments to make progress in NP discovery.

Theme 1: Bigger data, better data(bases) **(chapters 2-5)**

As more genomes are sequenced, a set of BGC prediction tools that covers not only well-studied BGC classes and taxa but also other emerging (and rare) pathways would facilitate the most comprehensive view over the true global NP diversity. In **chapter 2**, I introduce plantiSMASH, a new tool to complement antiSMASH to predict BGCs from plant genomes. I demonstrate the utility of plantiSMASH in **chapter 3,** using it to mine fungal-like BGC loci that produce sesterterpene compound families in multiple *Brassicaceae* species. Subsequently, as more BGC data gets predicted from genomes, a programmatically accessible high-quality reference database of known BGCs will play a crucial role in providing insights for the functional delineation of the putative BGCs. In **chapter 4**, I present significant quantitative and qualitative improvements to the MIBiG database, with an architecture that provides more robust programmatic access and allows better scaling of annotation data. Then in **chapter 5**, together with the researchers behind the NP-focused database NPAtlas [116], we provide a comprehensive review of relevant databases for microbial NP research along with some suggestions on how to improve database curation and development in the era of large-scale omics.

Theme 2: Mapping NP diversity through genomics **(chapters 6-8)**

Although BGC prediction tools like antiSMASH have been available for almost a decade, a well-developed BGC homology analysis tool is still lacking that can efficiently perform the large-scale grouping of BGCs into GCFs to assess their global biosynthetic diversity. In **chapter 6**, I present my contributions to the new BiG-SCAPE tool through building a user-friendly output visualization engine, which finally allows for a network-based GCF analysis from thousands of genomes to be done under a practical time and resource constraint. While the quadratic algorithm behind BiG-SCAPE is suitable to perform a sensitive homology analysis of up to 50-70 thousand BGCs, it is not equipped to handle the millions of BGCs estimated to be present in all publicly available microbial genomes to date. In **chapter 7**, I present BiG-SLiCE, a highly scalable tool that allows for significantly more efficient calculations of GCFs. With its near-linear clustering algorithm, BiG-SLiCE is designed to process millions of input BGCs using reasonable runtime and resources, which provides us with the first navigable map of global NP diversity. Finally, to allow for widespread adoption and exploration of this global analysis, in **chapter 8** I introduce BiG-FAM, the first online GCF database with many useful features such as comprehensive visualization with direct links to other BGC databases, multi-feature search functionalities, and rapid GCF queries of newly processed antiSMASH BGCs.

References

1. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. J Nat Prod. 2020;83: 770–803.
2. Strathdee SA, Davies SC, Marcelin JR. Confronting antimicrobial resistance beyond the COVID-19 pandemic and the 2020 US election. Lancet. 2020;396: 1050–1053.
3. Antimicrobial resistance : tackling a crisis for the health and wealth of nations / the Review on Antimicrobial Resistance chaired by Jim O'Neill. [cited 11 Nov 2020]. Available: https://wellcomecollection.org/works/rdpck35v
4. Cooper MA, Shlaes D. Fix the antibiotics pipeline. Nature. 2011;472: 32.
5. Talkington K, Shore C, Kothari P. A scientific roadmap for antibiotic discovery. The Pew Charitable Trust, Philadelphia, PA. 2016. Available: https://www.pewtrusts.org/en/research-and-analysis/reports/2016/05/a-scientific-roadmap-for-antibiotic-discovery
6. Schatz A, Bugle E, Waksman SA. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria.*. Exp Biol Med. 1944;55: 66–69.
7. Neelameghan A. Discovery, duplication, and documentation: a case study. Lib Sci Slant Doc. 1968;5: 264–288.
8. Waksman SA. Successes and failures in the search for antibiotics. Adv Appl Microbiol. 1969;11: 1–16.
9. Schäberle TF, Lohr F, Schmitz A, König GM. Antibiotics from myxobacteria. Nat Prod Rep. 2014;31: 953–972.
10. Tortorella E, Tedesco P, Palma Esposito F, January GG, Fani R, Jaspars M, et al. Antibiotics from Deep-Sea Microorganisms: Current Discoveries and Perspectives. Mar Drugs. 2018;16. doi:10.3390/md16100355
11. Suzuki Y, Banfield JF. Resistance to, and Accumulation of, Uranium by Bacteria from a Uranium-Contaminated Site. Geomicrobiol J. 2004;21: 113–121.
12. Saiki T, Kimura R, Arima K. Isolation and Characterization of Extremely Thermophilic Bacteria from Hot Springs. Agric Biol Chem. 1972;36: 2357–2366.
13. Jørgensen BB, Boetius A. Feast and famine--microbial life in the deep-sea bed. Nat Rev Microbiol. 2007;5: 770–781.
14. Weinberg ED. Biosynthesis of Secondary Metabolites: Roles of Trace Metals. In: Rose AH, Wilkinson JF, editors. Advances in Microbial Physiology. Academic Press; 1969. pp. 1–44.
15. Nett M, Ikeda H, Moore BS. Genomic basis for natural product biosynthetic diversity in the actinomycetes. Nat Prod Rep. 2009;26: 1362–1384.
16. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269: 496–512.
17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res. 2002;30: 17–20.
18. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, et al. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat Biotechnol. 2003;21: 526–531.
19. Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takahashi C, Shinose M, et al. Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci U S A. 2001;98: 12215–12220.
20. Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature. 2002;417: 141–147.
21. Donadio S, Staver MJ, McAlpine JB, Swanson SJ, Katz L. Modular organization of genes required for complex polyketide biosynthesis. Science. 1991;252: 675–679.
22. Oliynyk M, Samborskyy M, Lester JB, Mironenko T, Scott N, Dickens S, et al. Complete genome sequence of the erythromycin-producing bacterium Saccharopolyspora erythraea NRRL23338. Nat Biotechnol. 2007;25: 447–453.
23. Weber JM, Leung JO, Maine GT, Potenz RH, Paulus TJ, DeWitt JP. Organization of a cluster of erythromycin genes in Saccharopolyspora erythraea. J Bacteriol. 1990;172: 2372–2383.

24. Ayuso A, Clark D, González I, Salazar O, Anderson A, Genilloud O. A novel actinomycete strain de-replication approach based on the diversity of polyketide synthase and nonribosomal peptide synthetase biosynthetic pathways. Appl Microbiol Biotechnol. 2005;67: 795–806.

25. Dimise EJ, Widboom PF, Bruner SD. Structure elucidation and biosynthesis of fuscachelins, peptide siderophores from the moderate thermophile Thermobifida fusca. Proc Natl Acad Sci U S A. 2008;105: 15311–15316.

26. Lautru S, Deeth RJ, Bailey LM, Challis GL. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat Chem Biol. 2005;1: 265–269.

27. Bunet R, Song L, Mendes MV, Corre C, Hotel L, Rouhier N, et al. Characterization and manipulation of the pathway-specific late regulator AlpW reveals Streptomyces ambofaciens as a new producer of Kinamycins. J Bacteriol. 2011;193: 1142–1153.

28. Chen R, Zhang Q, Tan B, Zheng L, Li H, Zhu Y, et al. Genome Mining and Activation of a Silent PKS/NRPS Gene Cluster Direct the Production of Totopotensamides. Org Lett. 2017;19: 5697–5700.

29. Stevens DC, Conway KR, Pearce N, Villegas-Peñaranda LR, Garza AG, Boddy CN. Alternative sigma factor over-expression enables heterologous expression of a type II polyketide biosynthetic pathway in Escherichia coli. PLoS One. 2013;8: e64858.

30. Kaysser L, Bernhardt P, Nam S-J, Loesgen S, Ruby JG, Skewes-Cox P, et al. Merochlorins A-D, cyclic meroterpenoid antibiotics biosynthesized in divergent pathways with vanadium-dependent chloroperoxidases. J Am Chem Soc. 2012;134: 11988–11991.

31. Chiang Y-M, Oakley CE, Ahuja M, Entwistle R, Schultz A, Chang S-L, et al. An efficient system for heterologous expression of secondary metabolite genes in Aspergillus nidulans. J Am Chem Soc. 2013;135: 7720–7731.

32. Casini A, Chang F-Y, Eluere R, King AM, Young EM, Dudley QM, et al. A Pressure Test to Make 10 Molecules in 90 Days: External Evaluation of Methods to Engineer Biology. J Am Chem Soc. 2018;140: 4302–4316.

33. Shao Z, Rao G, Li C, Abil Z, Luo Y, Zhao H. Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. ACS Synth Biol. 2013;2: 662–669.

34. Winter JM, Behnken S, Hertweck C. Genomics-inspired discovery of natural products. Curr Opin Chem Biol. 2011;15: 22–31.

35. Challis GL, Ravel J. Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS Microbiol Lett. 2000;187: 111–114.

36. Yan X, Ge H, Huang T, Hindra, Yang D, Teng Q, et al. Strain Prioritization and Genome Mining for Enediyne Natural Products. MBio. 2016;7. doi:10.1128/mBio.02104-16

37. Hu JL, Xue YC, Xie MY, Zhang R, Otani T, Minami Y, et al. A new macromolecular antitumor antibiotic, C-1027. I. Discovery, taxonomy of producing organism, fermentation and biological activity. J Antibiot . 1988;41: 1575–1579.

38. Otani T, Minami Y, Marunaka T, Zhang R, Xie MY. A new macromolecular antitumor antibiotic, C-1027. II. Isolation and physico-chemical properties. J Antibiot . 1988;41: 1580–1585.

39. Zhen YS, Ming XY, Yu B, Otani T, Saito H, Yamada Y. A new macromolecular antitumor antibiotic, C-1027. III. Antitumor activity. J Antibiot . 1989;42: 1294–1298.

40. Luo Y, Cobb RE, Zhao H. Recent advances in natural product discovery. Curr Opin Biotechnol. 2014;30: 230–237.

41. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015;14: 111–129.

42. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380.

43. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005;309: 1728–1732.

44. Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A. 2003;100: 3960–3964.

45. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351.

46. Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanal Rev. 2010;2: 23–60.

47. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2020. doi:10.1093/nar/gkaa971

48. O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–45.
49. Stepanauskas R. Single cell genomics: an individual look at microbes. Curr Opin Microbiol. 2012;15: 613–620.
50. Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science. 2011;333: 1296–1300.
51. Eloe-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, et al. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. Nat Commun. 2016;7: 10476.
52. Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, et al. Assembling the marine metagenome, one cell at a time. PLoS One. 2009;4: e5299.
53. Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Richter RA, Valas R, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 2012;6: 1186–1199.
54. Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG, et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. Nat Commun. 2013;4: 1854.
55. Wilson MC, Mori T, Rückert C, Uria AR, Helf MJ, Takada K, et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature. 2014;506: 58–62.
56. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth’s microbiomes. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0718-6
57. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2: 1533–1542.
58. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5: 170203.
59. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 2016;7: 13219.
60. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1: 16048.
61. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176: 649–662.e20.
62. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568: 499–504.
63. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568: 505–510.
64. Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res. 2020;30: 315–333.
65. Alneberg J, Karlsson CMG, Divne A-M, Bergin C, Homa F, Lindh MV, et al. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome. 2018;6: 173.
66. Kautsar S. Growth of bacterial genomes in RefSeq. 2020. doi:10.6084/m9.figshare.13363613.v1
67. Moody G. Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. John Wiley & Sons; 2004.
68. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410.
69. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28: 45–48.
70. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 2002;30: 281–283.
71. Ansari MZ, Yadav G, Gokhale RS, Mohanty D. NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res. 2004;32: W405–13.
72. Tae H, Kong E-B, Park K. ASMPKS: an analysis system for modular polyketide synthases. BMC Bioinformatics. 2007;8: 327.
73. Minowa Y, Araki M, Kanehisa M. Comprehensive analysis of distinctive polyketide and

nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol. 2007;368: 1500–1517.

74. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14: 755–763.

75. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23: 205–211.

76. Starcevic A, Zucko J, Simunkovic J, Long PF, Cullum J, Hranueli D. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res. 2008;36: 6882–6892.

77. Li MHT, Ung PMU, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 2009;10: 185.

78. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol. 2009;140: 13–17.

79. Bachmann BO, Ravel J. Chapter 8 Methods for In Silico Prediction of Microbial Polyketide and Nonribosomal Peptide Biosynthetic Pathways from DNA Sequence Data. Complex Enzymes in Microbial Natural Product Biosynthesis, Part A: Overview Articles and Peptides. Elsevier; 2009. pp. 181–217.

80. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47: 736–741.

81. de Jong A, van Hijum SAFT, Bijlsma JJE, Kok J, Kuipers OP. BAGEL: a web-based bacteriocin genome mining tool. Nucleic Acids Res. 2006;34: W273–9.

82. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339–46.

83. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster ALH, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 2015;43: 9645–9662.

84. Kim J, Yi G-S. PKMiner: a database for exploring type II polyketide synthases. BMC Microbiol. 2012;12: 169.

85. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res. 2017;45: W80–W88.

86. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai H-C, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol. 2017;13: 470–478.

87. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics. 2017;33: 3202–3210.

88. Wolf T, Shelest V, Nath N, Shelest E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. Bioinformatics. 2016;32: 1138–1143.

89. Andersen MR, Nielsen JB, Klitgaard A, Petersen LM, Zachariasen M, Hansen TJ, et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proc Natl Acad Sci U S A. 2013;110: E99–107.

90. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87.

91. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015;43: W237–43.

92. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45: W36–W41.

93. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013;41: W204–12.

94. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.

95. Mushtaq S, Abbasi BH, Uzair B, Abbasi R. Natural products as reservoirs of novel therapeutic agents. EXCLI J. 2018;17: 420–451.

96. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters - from genetics to genomics. New Phytol. 2016;211: 771–789.
97. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The Sequenced Angiosperm Genomes and Genome Databases. Front Plant Sci. 2018;9: 418.
98. Chen F, Song Y, Li X, Chen J, Mo L, Zhang X, et al. Genome sequences of horticultural plants: past, present, and future. Hortic Res. 2019;6: 112.
99. Caboche S, Pupin M, Leclère V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. Nucleic Acids Res. 2008;36: D326–31.
100. Hammami R, Zouhir A, Ben Hamida J, Fliss I. BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol. 2007;7: 89.
101. Whitmore L, Wallace BA. The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols. Nucleic Acids Res. 2004;32: D593–4.
102. Anand S, Prasad MVR, Yadav G, Kumar N, Shehara J, Ansari MZ, et al. SBSPKS: structure based sequence analysis of polyketide synthases. Nucleic Acids Res. 2010;38: W487–96.
103. Tae H, Sohng JK, Park K. MapsiDB: an integrated web database for type I polyketide synthases. Bioprocess Biosyst Eng. 2009;32: 723–727.
104. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 2013;41: D402–7.
105. Ichikawa N, Sasagawa M, Yamamoto M, Komaki H, Yoshida Y, Yamazaki S, et al. DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2013;41: D408–14.
106. Lucas X, Senger C, Erxleben A, Grüning BA, Döring K, Mosch J, et al. StreptomeDB: a resource for natural compounds isolated from Streptomyces species. Nucleic Acids Res. 2013;41: D1130–6.
107. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11: 625–631.
108. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2017;45: D555–D559.
109. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.
110. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc Natl Acad Sci U S A. 2014;111: E1130–9.
111. Vicente CM, Thibessard A, Lorenzi J-N, Benhadj M, Hôtel L, Gacemi-Kirane D, et al. Comparative Genomics among Closely Related Streptomyces Strains Revealed Specialized Metabolite Biosynthetic Gene Cluster Diversity. Antibiotics (Basel). 2018;7. doi:10.3390/antibiotics7040086
112. Bilyk O, Brötz E, Tokovenko B, Bechthold A, Paululat T, Luzhetskyy A. New Simocyclinones: Surprising Evolutionary and Biosynthetic Insights. ACS Chem Biol. 2016;11: 241–250.
113. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432.
114. Lin K, Zhu L, Zhang D-Y. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics. 2006;22: 2081–2086.
115. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
116. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019;5: 1824–1833.

Chapter 2

PlantiSMASH: A New Tool to Explore Operon-like Gene Clusters in Plants

While BGCs and operons are considered to be a hallmark of microbial secondary metabolism, they are not so commonly associated with higher eukaryotes such as plants. Nevertheless, until 2017, at least 21 "operon-like" gene clusters have been reported to encode the biosynthesis of diverse compound classes and are harbored by a wide range of plant species, posing intriguing questions regarding their true prevalence in nature. In this chapter, we introduce plantiSMASH, a new bioinformatics tool to identify and predict such clustered loci in plant genomes. Using plantiSMASH, we identified more than two thousand putative BGCs from 47 high-quality plant genomes, some of which harbor a highly diverse combination of biosynthetic enzymes. Finally, to aid in the investigation of the predicted BGCs, plantiSMASH provides a user-friendly interface to examine the co-expression pattern of genes within BGCs and across distant loci.

Published as:

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*.

* shared first authorship

2.1. Introduction

Across Planet Earth, bacteria, fungi and plants produce an immense diversity of specialized metabolites, each with their own specific ecological roles in the multi-organismal 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 this genome mining process, BGCs are computationally identified in genome sequences and then linked to molecules 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 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: (i) finding biosynthetic genes for important known compounds to allow heterologous production through fermentation in industrial strains, and (ii) identifying novel natural product chemistry guided by biosynthetic gene cluster diversity. Altogether, this development has appropriately been termed the “gene cluster revolution” [1].

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 *A. 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 eukaryotic genome sequencing [13] are finally making complete plant genome sequencing feasible at larger scales: high-quality plant genome sequences for almost a hundred species are already publicly available, and more or less complete genomes can be sequenced for as little as ten to fifty thousand 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 (i) plant biosynthetic pathways involve unique enzyme families not found in bacteria and fungi; (ii) not all plant biosynthetic pathways are clustered (e.g. anthocyanins [15]), so identification of a biosynthetic gene does not equal identification of a BGC; (iii) intergenic distances in plant genomes are larger and much more variable [16–19]; (iv) plant genomes contain clustered groups of genes (e.g. tandem arrays) whose products do not constitute a pathway; (v) 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.

2.2. Methods and Implementation

2.2.1. A procedure for the identification of candidate plant biosynthetic gene clusters

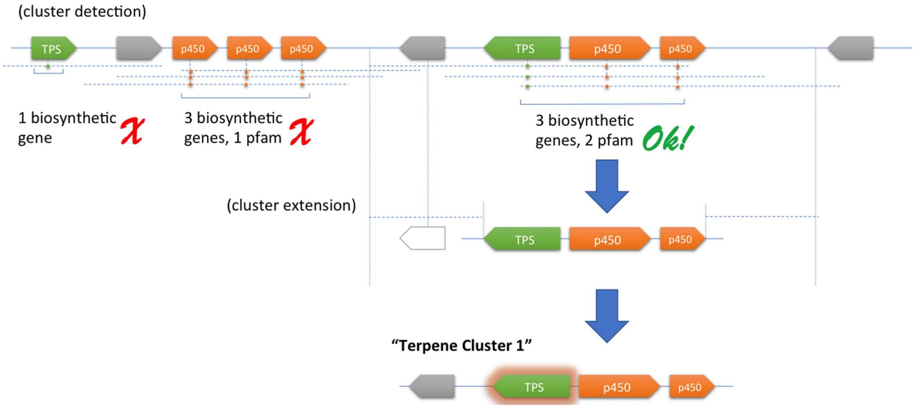


Figure 2.1. General strategy followed by plantiSMASH for the identification of plant BGCs. First, plantiSMASH identifies biosynthetic genes (having a hit on one of the 62 pHMMs) that are located in close proximity to each other. Subsequently, it will look for the co-occurrence of at least three biosynthetic enzyme-coding genes, comprising at least two different enzyme types. (Based on the results of the CD-HIT clustering of encoded protein sequences, closely related duplicate genes will only be counted once). Afterward, identified clusters are extended to incorporate any flanking genes. Finally, each cluster is classified based on the presence of core enzymes (see Supplementary Table S1). In this example, the detected cluster is assigned to the "Terpene" class due to the presence of a terpene synthase-encoding gene.

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 in bacteria and fungi, this procedure is unfit to detect biosynthetic gene clusters in plants, for the reasons described above. To address this, a novel detection strategy was chosen (Figure 2.1): instead of identifying BGCs through the identification of core scaffold-generating genes alone, plantiSMASH identifies them by looking for all genes predicted to encode 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)" [14] (Albeit arbitrary, this definition correctly describes all known plant BGCs at the moment and is open to improvement as more are discovered). Accordingly, 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 pHMMs specific for each class (Supplementary Table S1); 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. This successfully distinguishes potentially real BGCs from tandem repeat regions that are also frequently found in genomes (Supplementary Table S2).

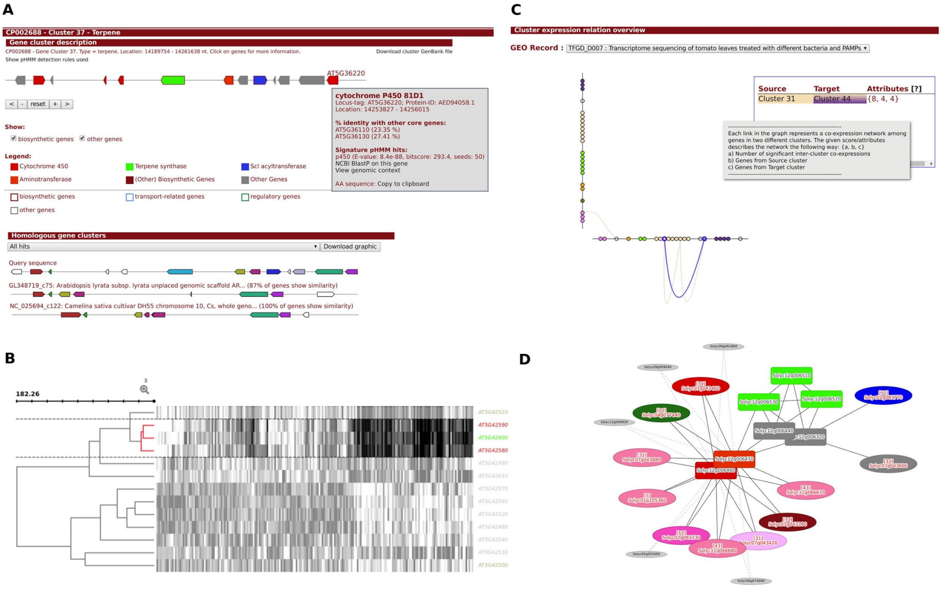
In order to identify all classes of biosynthetic enzymes known to be involved in plant specialized metabolic pathways, we performed a comprehensive literature search of previously characterized plant biosynthetic pathways, which resulted in a list of 62 protein domains that have been associated with specialized metabolic pathways in plants (see Supplementary Table S1). Fifty-seven of these protein domains are represented by pHMMs from the Pfam database [24], and custom pHMMs were generated for five 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 Supplementary Table S3 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-v-1 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 cutoff 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 (see Supplementary Table S2 and Supplementary Figures S1 and S2 for results illustrating calibration of the defaults).

2.2.2. Flexible and user-friendly input and output

To obtain a reliable BGC prediction, high-quality annotations 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 such a procedure. 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 Generic Feature Format version 3 (GFF3) parsing package (http://biopython.org/wiki/GFF_Parsing) capable of combining the CDS features from the input sequence, if any, with those of a file compliant to the GFF3 standard. 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 2.2), which is largely reminiscent of the output of microbial antiSMASH jobs. 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.



CP002688 - Cluster 37 - Terpene

Gene cluster description

CP002688 - Gene Cluster 37. Type = terpene. Location: 14189754 - 14261638 nt. Click on genes for more information.

Download cluster GenBank file

Show pHMM detection rules used

AT5G36220

< | - | reset | + | >

Show:

☑ biosynthetic genes ☑ other genes

Legend:

🟥 Cytochrome 450	🟩 Terpene synthase	🟦 Scl acyltransferase
🟥 Aminotransferase	🟥 (Other) Biosynthetic Genes	■ Other Genes
□ biosynthetic genes	□ transport-related genes	□ regulatory genes
□ other genes		

cytochrome P450 81D1
Locus-tag: AT5G36220; Protein-ID: AED94058.1
Location: 14253827 - 14256015

% identity with other core genes:
AT5G36110 (23.35 %)
AT5G36130 (27.41 %)

Signature pHMM hits:
p450 (E-value: 8.4e-88, bitscore: 293.4, seeds: 50)
NCBI BlastP on this gene
View genomic context

AA sequence: Copy to clipboard

Homologous gene clusters

All hits ▾ | Download graphic

Query sequence

GL348719_c75: Arabidopsis lyrata subsp. lyrata unplaced genomic scaffold AR... (87% of genes show similarity)

NC_025694_c122: Camelina sativa cultivar DH55 chromosome 10, Cs, whole geno... (100% of genes show similarity)

Cluster expression relation overview

GEO Record : TFGD_D007 : Transcriptome sequencing of tomato leaves treated with different bacteria and PAMPs ▾

Source	Target	Attributes [?]
Cluster 31	Cluster 44	{8, 4, 4}

Each link in the graph represents a co-expression network among genes in two different clusters. The given score/attributes describes the network the following way: {a, b, c}
a) Number of significant inter-cluster co-expressions
b) Genes from Source cluster
c) Genes from Target cluster

Figure 2.2. 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* (47) 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* (48) (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 α-tomatine biosynthesis in *S. lycopersicum* (20) are highlighted (clusters 31 and 44). (D) Example ego network that summarizes coexpression correlations between members of the α-tomatine gene (cluster 44), as well as with genes in other gene clusters (including the other α-tomatine biosynthetic locus, cluster 31) and with genes elsewhere on the genome.

2.2.3. 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 was 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 2-fold: 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.

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 (i) are members of another candidate BGC and show high gene expression correlation (>0.9 PCC) with at least one gene in the BGC, or (ii) contain a “biosynthetic” domain (defined as being one of the domains in Supplementary Table S1) 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: (i) all nodes belong to the same Louvain community [31], as determined by analyzing the full coexpression network of all candidate clusters’ genes; (ii) all nodes have a transitivity greater than zero; (iii) the subnetwork contains at least two genes from each cluster; (iv) the subnetwork contains at least one gene per cluster that has a biosynthetic domain; and (v) The subnetwork contains at least three genes with a biosynthetic domain. This highlights arrangements of pairs of clusters that may be linked functionally via coexpression, and is reminiscent of the characterized α-tomatine biosynthetic pathway in *S. lycopersicum*, which is encoded in two separate clusters that are highly coexpressed [20].

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 does 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.

2.2.4. 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 [32]. 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 [33] 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 [34], which allows users to find similarities between newly identified and known clusters with the KnownClusterBlast module of antiSMASH.

2.2.5. 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 47 high-quality plant genomes were precomputed and made available online at http://plantismash.secondarymetabolites.org/precalc. 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: *A. thaliana*, *S. lycopersicum*, *O. sativa*, *Z. mays* and *G. max* (Supplementary Tables S4–S7). As an indication for web server users: the computations took about 24 minutes per genome on average, depending on the size of the genome and pre-selected additional analyses including the coexpression analysis (see further details in Supplementary Table S4).

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.

2.3. Results and Discussion

2.3.1. PlantiSMASH successfully detects all experimentally characterized plant biosynthetic gene clusters

Even though only a relatively small set of plant BGCs has been discovered, 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 [21,35], to relatively simple ones that only encode a couple of enzymes, such as the dhurrin and linamarin/lotaustralin BGCs [36]. Of this set, only nineteen BGCs have annotated sequence information publicly available. When plantiSMASH was run on a multi-GenBank file containing accurately annotated versions of these BGCs, all clusters were successfully detected with default settings. When run 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.

When run on the 47 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 2.3). 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 *C. sativa* [37] with 88 candidate BGCs, *B. napus* [38] with 68 candidate BGCs and *G. max* [39] with 76 candidate BGCs.

2.3.2. Plant genomes contain large numbers of complex biosynthetic gene clusters

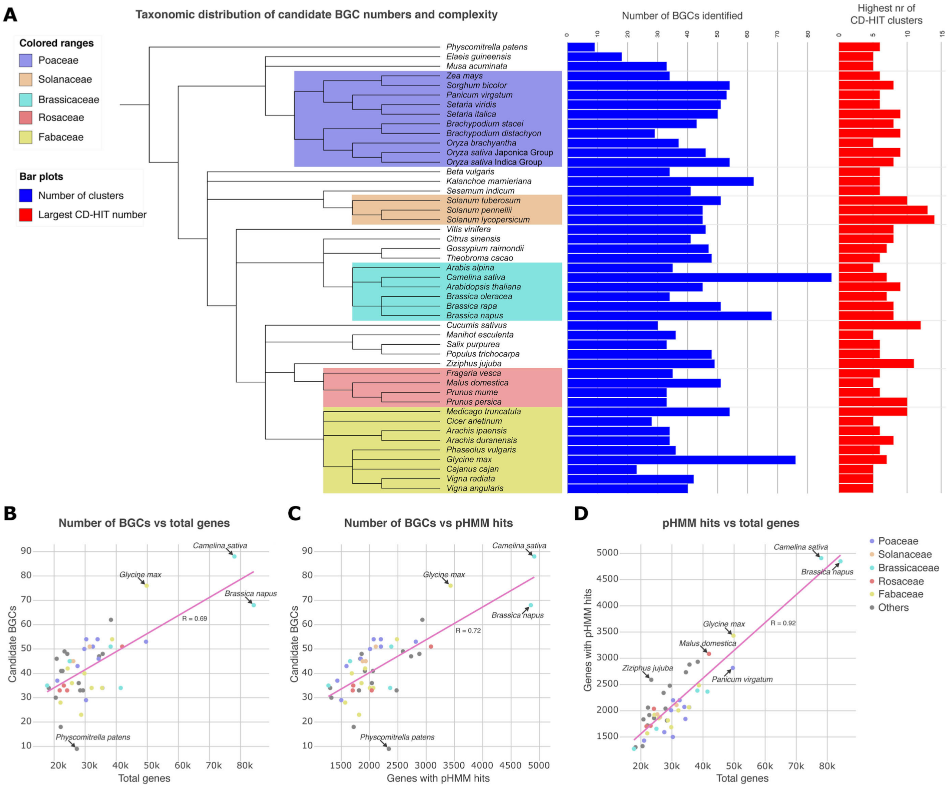


Figure 2.3. Numbers of candidate BGCs identified across the plant kingdom. (A) plantiSMASH BGC predictions plotted onto a phylogenetic tree of plant species for which chromosome-level genome assemblies are available. The blue bars indicate the number of candidate BGCs per genome, the red bars indicate 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.

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 2.4.

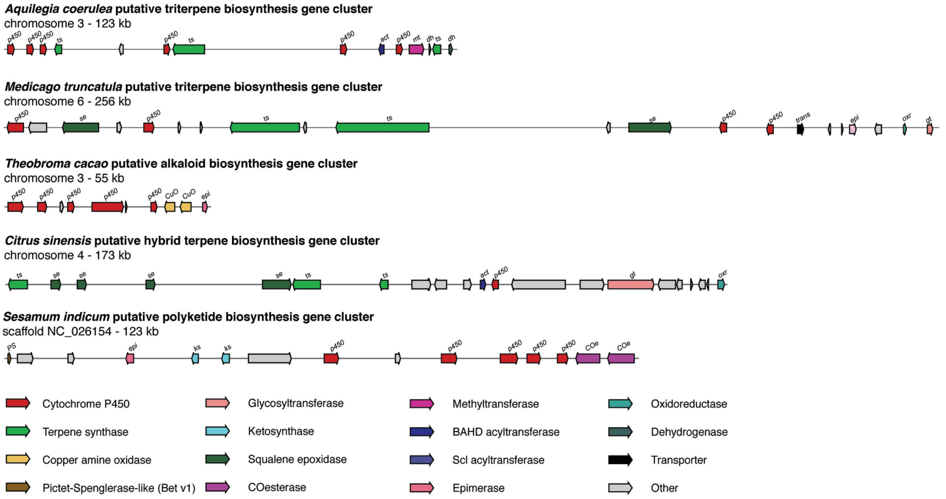

Aquilegia coerulea putative triterpene biosynthesis gene cluster
chromosome 3 - 123 kb
p450 p450 p450 ts p450 ts p450 act p450 mt dh ts dh
Medicago truncatula putative triterpene biosynthesis gene cluster
chromosome 6 - 256 kb
p450 se p450 ts ts se p450 p450 trans epi oxr gt
Theobroma cacao putative alkaloid biosynthesis gene cluster
chromosome 3 - 55 kb
p450 p450 p450 p450 p450 CuO CuO epi
Citrus sinensis putative hybrid terpene biosynthesis gene cluster
chromosome 4 - 173 kb
ts se se se se ts ts act p450 gt oxr
Sesamum indicum putative polyketide biosynthesis gene cluster
scaffold NC_026154 - 123 kb
PS epi ks ks p450 p450 p450 p450 p450 COe COe
Cytochrome P450
Terpene synthase
Copper amine oxidase
Pictet-Spenglerase-like (Bet v1)
Glycosyltransferase
Ketosynthase
Squalene epoxidase
COesterase
Methyltransferase
BAHD acyltransferase
Scl acyltransferase
Epimerase
Oxidoreductase
Dehydrogenase
Transporter
Other

Figure 2.4. 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.

2.3.3. Coexpression patterns can guide BGC prioritization

We subjected the candidate BGCs identified in the genome of *A. thaliana* to a more detailed statistical analysis using within-cluster coexpression in a merged transcriptomic dataset. For this, we compiled two sets of gene expression datasets, one containing transcriptomic experiments of biological treatments (defense; Supplementary Table S5) and one containing experiments of hormone treatments and non-biological stress inductions (Supplementary Table S6). Together, these datasets comprise transcriptomic measurements of 1,047 samples. The Mann–Whitney U one-sided test was selected to test which of the *A. thaliana* BGCs have a statistically greater within-cluster coexpression distribution than the genome's background coexpression distribution. Given a BGC consisting of x genes, the background distribution for the statistical test of this cluster contains all PCCs between pairs of genes that are x-1, x-2, …, 0 genes away from each other across the entire genome (except predicted BGCs). Only genes observed in all transcriptomic experiments were allowed in the test, and only PCCs between genes that each have a >0 median absolute deviation are added to the distributions.

Lastly, the CD-HIT algorithm was run on the entire *A. thaliana* proteome at a 0.5 identity cutoff (same as plantiSMASH's default) to cluster all similar enzymes. The same statistical tests were repeated afterward, but this time discarding PCCs between genes that code for enzymes within the same CD-HIT cluster, ensuring both distributions only include coexpression of genes that produce enzymes of different classes, which more accurately resembles the type of interactions desired in a bona fide BGC. The results of these analyses (Supplementary Table S8 and Supplementary Figure S3) show that at a significance level of 0.05, eleven predicted BGCs showed a statistically higher within-cluster coexpression than their respective background distribution even when discarding coexpression between genes in the same CD-HIT cluster. This list includes the four known *A. thaliana* BGCs, encoding the biosynthetic pathways for arabidiol/baruol (P = 2.92e^{-40}), thalianol (P = 1.94e^{-17}), marneral (P = 7.03e^{-10}) and tirucalla-7,24-dien-3β-ol (P = 1.10e^{-4}), which corroborates that coexpression is a valid criterion to prioritize functional BGCs.

There are several explanations for the fact that strong coexpression is observed for some candidate BGCs but not others. A first explanation is that their coordinated expression is induced by conditions not included in these transcriptomic experiments; in other words, the absence of evidence of coexpression is not evidence of an absence of coexpression. A second 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 (Supplementary Table S9); e.g., AT1G24100 and AT5G57220, which occur in two different candidate BGCs, are known to each be involved in a different branch of glucosinolate biosynthesis [40,41], 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=0.004, and P=0.64 when fitting linear regression) of coexpression with cluster size, which suggests that the default plantiSMASH BGC prediction cut-offs are not set too inclusively.

All in all, coexpression analysis provides a powerful tool to prioritize the candidate BGCs detected by plantiSMASH that are most likely to encode functional pathways.

2.4. Conclusion

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 toward 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 [42–46], 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 [47,48]. 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 mimic 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 Supplementary Table S9. We thank the research groups of Elizabeth Sattely and Sue Rhee for useful feedback and discussions.

Funding

Netherlands Organization for Scientific Research (NWO) VENI Grant [863.15.002 to M.H.M.]; Graduate School for Experimental Plant Sciences (EPS); Novo Nordisk Foundation (to K.B.); UK Biotechnological and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Grant “Understanding and Exploiting Plant and Microbial Metabolism” [BB/J004561/1 to A.O.]; John Innes Foundation, the joint Engineering and Physical Sciences Research Council/BBSRC-funded OpenPlant Synthetic Biology Research Centre grant [BB/L014130/1 to A.O.]; National Institutes of Health Genome to Natural Products Network award [U101GM110699 to A.O., M.H.M.]. Funding for open access charge: Netherlands Organization for Scientific Research (NWO).

Conflict of interest statement

None declared.

Supplementary Material

Supplementary figures and tables can be downloaded from https://bit.ly/3s2nCq7.

References

1. Jensen PR. Natural Products and the Gene Cluster Revolution. Trends Microbiol. 2016;24: 968–977.
2. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nat Chem Biol. 2015;11: 639–648.
3. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nat Rev Microbiol. 2015;13: 509–523.
4. Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes - a review. Nat

Prod Rep. 2016;33: 988–1005.

5. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339–46.
6. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013;41: W204–12.
7. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015;43: W237–43.
8. Frey M, Chomet P, Glawischnig E, Stettner C, Grün S, Winklmair A, et al. Analysis of a chemical plant defense mechanism in grasses. Science. 1997;277: 696–699.
9. Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. Proc Natl Acad Sci U S A. 2004;101: 8233–8238.
10. Nützmann H-W, Osbourn A. Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 2014;26: 91–99.
11. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters - from genetics to genomics. New Phytol. 2016;211: 771–789.
12. Yu N, Nützmann H-W, MacDonald JT, Moore B, Field B, Berriri S, et al. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res. 2016;44: 2255–2265.
13. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature. 2015;527: 508–511.
14. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Nat Prod Rep. 2016;33: 951–962.
15. Shi M-Z, Xie D-Y. Biosynthesis and metabolic engineering of anthocyanins in Arabidopsis thaliana. Recent Pat Biotechnol. 2014;8: 47–60.
16. Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang T-H, et al. Architecture and evolution of a minute plant genome. Nature. 2013;498: 94–98.
17. Keller B, Feuillet C. Colinearity and gene density in grass genomes. Trends Plant Sci. 2000;5: 246–251.
18. Kellogg EA, Bennetzen JL. The evolution of nuclear genome structure in seed plants. Am J Bot. 2004;91: 1709–1725.
19. Sandhu D, Gill KS. Gene-containing regions of wheat and the other grass genomes. Plant Physiol. 2002;128: 803–811.
20. Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science. 2013;341: 175–179.
21. Shang Y, Ma Y, Zhou Y, Zhang H, Duan L, Chen H, et al. Plant science. Biosynthesis, regulation, and domestication of bitterness in cucumber. Science. 2014;346: 1084–1088.
22. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.
23. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.
24. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42: D222–30.
25. Wu F, Zhu H, Sun L, Rajendran C, Wang M, Ren X, et al. 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. 2012;134: 1498–1500.
26. Lee E-J, Facchini P. Norcoclaurine synthase is a member of the pathogenesis-related 10/Bet v1 protein family. Plant Cell. 2010;22: 3489–3503.
27. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20: 2878–2879.
28. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40: D1178–86.
29. Rajniak J, Barco B, Clay NK, Sattely ES. A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. Nature. 2015;525: 376–379.
30. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. Learning from Co-expression Networks:

Possibilities and Challenges. Front Plant Sci. 2016;7: 444.

31. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008: P10008.

32. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, et al. Genome structure of the legume, Lotus japonicus. DNA Res. 2008;15: 227–239.

33. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12: 59–60.

34. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11: 625–631.

35. Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, et al. A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science. 2012;336: 1704–1708.

36. Takos AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, et al. Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicu*s and suggests the repeated evolution of this chemical defence pathway. Plant J. 2011;68: 273–286.

37. Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, et al. The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun. 2014;5: 3706.

38. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, et al. Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science. 2014;345: 950–953.

39. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463: 178–183.

40. Grubb CD, Zipp BJ, Ludwig-Müller J, Masuno MN, Molinski TF, Abel S. *Arabidopsis* glucosyltransferase UGT74B1 functions in glucosinolate biosynthesis and auxin homeostasis. Plant J. 2004;40: 893–908.

41. Pfalz M, Vogel H, Kroymann J. The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in Arabidopsis. Plant Cell. 2009;21: 985–999.

42. Casini A, Storch M, Baldwin GS, Ellis T. Bricks and blueprints: methods and standards for DNA assembly. Nat Rev Mol Cell Biol. 2015;16: 568–576.

43. Liu W, Yuan JS, Stewart CN Jr. Advanced genetic tools for plant biotechnology. Nat Rev Genet. 2013;14: 781–793.

44. Patron NJ. DNA assembly for plant biology: techniques and tools. Curr Opin Plant Biol. 2014;19: 14–19.

45. Patron NJ, Orzaez D, Marillonnet S, Warzecha H, Matthewman C, Youles M, et al. Standards for plant synthetic biology: a common syntax for exchange of DNA parts. New Phytol. 2015;208: 13–19.

46. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A. Triterpene biosynthesis in plants. Annu Rev Plant Biol. 2014;65: 225–257.

47. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al. Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 2015;112: E81–8.

48. Field B, Osbourn AE. Metabolic diversification--independent assembly of operon-like gene clusters in different plants. Science. 2008;320: 543–547.

Addendum

Genome Mining of Putative Plant BGCs with PlantiSMASH

This addendum provides a step-by-step procedure covering the installation, input preparation, and the running of plantiSMASH on a newly sequenced plant genome. It also briefly discusses the interpretation of the analysis results, helping users prioritize the most interesting BGCs for follow-up studies.

Published as:

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.
In *Plant Chemical Genomics (pp. 173-188)*. Humana Press, New York, NY.

* shared first authorship

2.A1. Introduction

For many centuries already, plant-derived natural products have played key roles in medicine. Now, the rapidly growing availability of plant genomes and transcriptomes is opening up opportunities to apply computational strategies toward the discovery of novel plant molecules [1]. Of specific interest in these endeavors is the recent discovery that the genes encoding plant biosynthetic enzymes for a given pathway are frequently found in close physical proximity to each other on the chromosome in biosynthetic gene clusters (BGCs) [2–4].

Growing interest in plant BGCs has led to the development of multiple bioinformatic methods and tools [5–7] to identify them; one of these is plantiSMASH [8]. Derived from the widely used microbial and fungal BGC prediction tool antiSMASH [9], it shares similar distinct characteristics, namely ease of use, feature-rich analysis and visualization, and a modular architecture.

In order to use plantiSMASH, users need to supply genomic data (with or without annotations), preferably accompanied by transcriptomic data. PlantiSMASH offers a wide array of options to set up an analysis, and offers multiple visual outputs that require careful expert interpretation.

In this chapter, we provide detailed guidelines for installing plantiSMASH, setting up plantiSMASH analyses, and interpreting its output.

2.A2. Materials

2.A2.1. Hardware and Operating System

Web Server Version:

- Computer and OS capable of running fairly recent web browsers (preferably Firefox version 40.0 or later), equipped with at least 2 GB of RAM.
- Internet connection.

Stand-Alone Version:

- Computer with at least 4GB of RAM and 5GB free hard disk space.
- Unix-derived OS (Linux, MacOS X with Homebrew) or Windows running a Linux Virtual Machine.

2.A2.2. Software

Web Server Version:

- Web browser (preferably Firefox version 40.0 or later).

Stand-Alone Version:

- PlantiSMASH stand-alone source code downloaded from http://plantismash.secondarymetabolites.org/download.html.
- Required packages and libraries (as of plantiSMASH version 1.0.0: for an updated list of requirements, please refer to the download link above):
    - Python v2.7.0
    - Python2 pip-installer
    - Glimmer v3.02
    - GlimmerHMM v3.0.4
    - HMMer v2.3.2
    - HMMER v3.1b2
    - FastTree v2.1.7
    - MUSCLE v3.8.31
    - Prodigal v2.6.1
    - NCBI Blast+ v2.2.31
    - XZ Utils 5.1.1
    - Libxml 2.9.1
    - Argparse
    - Straight.plugin v1.4.0-post-1
    - Cssselect
    - Pyquery v1.2.9
    - Numpy
    - Biopython v1.65
    - Helperlibs
    - Pysvg
    - PyExcelerator
    - Backports.lzma
    - Networkx v1.11
    - Python-louvain v0.5
    - Bcbio-gff v0.6.2

2.A2.3. Genomic Data

PlantiSMASH accepts two types of genomic data: unannotated (FASTA) and feature-annotated (GBK/EMBL/GFF+FASTA) sequence files. Exact assembly quality requirements vary between genomes, but as a rule of thumb scaffolds/contigs should be large enough to contain at least three adjacent genes. By default, scaffolds smaller than one thousand base pairs are ignored by the algorithm. Additionally, the quality of the BGC predictions is highly dependent on the quality of the assembly and annotations provided as input; this makes choosing and preparing the input data beforehand an important step before using plantiSMASH.

To make this process easier, plantiSMASH accepts sequences and annotations in different formats, as shown in Figure 2.5.

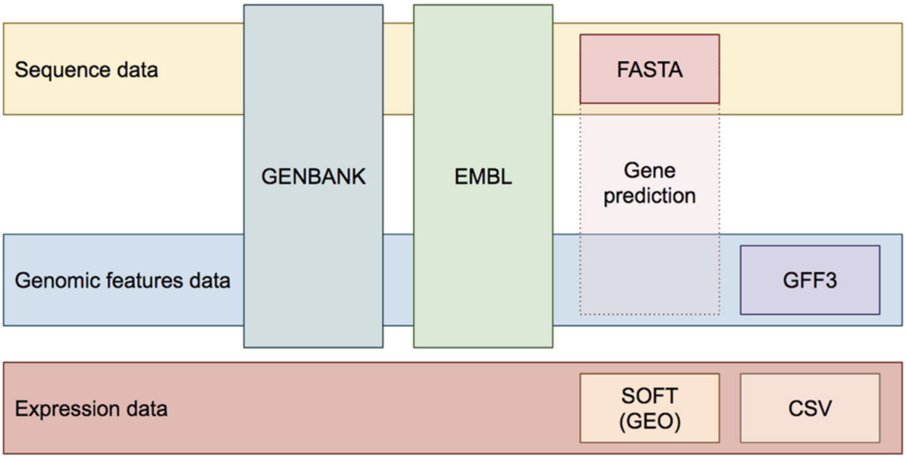


Figure 2.A1. Input data for plantiSMASH.

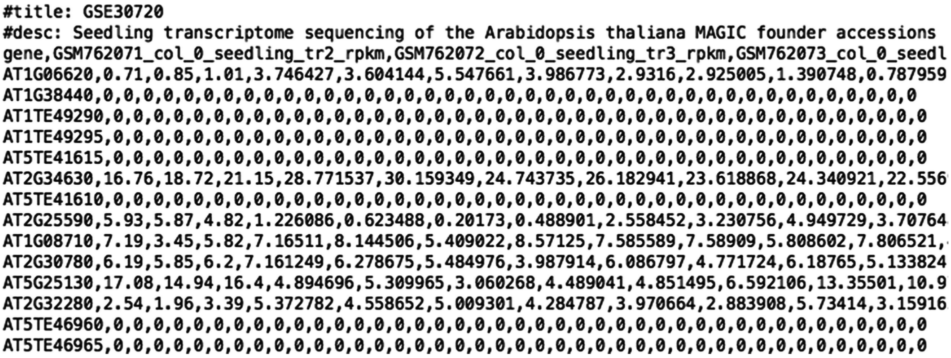
Sequence input is mandatory, and can be provided in FASTA, GENBANK, or EMBL format. The latter two formats have the advantage of also allowing the inclusion of genomic features and annotation data within the same file. Genomic features can also be provided independently in GFF3 format, or not at all, in which case plantiSMASH will use its own gene-finding module to annotate the genome (however, given the importance of high-quality annotations for obtaining optimal results, a previously annotated genome is highly recommended over this latter option).

2.A2.4. Expression Data

PlantiSMASH includes a gene expression analysis module to facilitate the study of coexpression patterns in the BGCs predicted by the algorithm. Multiple gene expression datasets may be used as input at the same time in all versions of plantiSMASH for independent analysis, allowing the user to compare results among different experiments in a simple manner. Gene expression data must be provided alongside sequence and genomic feature data. As previously seen in Figure 2.A1, plantiSMASH accepts two formats for this: SOFT files and CSV files.

SOFT (Simple Omnibus Format in Text): With over 85,000 series and more than 2,000,000 samples combined, NCBI’s Gene Expression Omnibus/GEO (https://www-ncbi-nlm-nih-gov.ezproxy.library.wur.nl/geo/) is one of the most widely used repositories for microarray, next-generation sequencing, and high-throughput functional genomic data. The database stores raw data as submitted by the authors and data prepared in the SOFT format, which can hold the expression data along with descriptive information regarding the experiment in a machine-readable format. SOFT files can be retrieved for both GEO Dataset (GDS) and GEO Series (GSE); as long as it contains complete expression values data, it can be used by plantiSMASH. Being a flat file format, SOFT files can be opened and edited in any plain text editor.

CSV (Comma-Separated Values): For expression data not in SOFT files, or not yet publicly available, plantiSMASH can also read gene expression data in CSV format. In this format, data is stored tabularly, with genes represented in rows, and samples in columns. An example can be seen in Figure 2.A2.



```
#title: GSE30720
#desc: Seedling transcriptome sequencing of the Arabidopsis thaliana MAGIC founder accessions
gene,GSM762071_col_0_seedling_tr2_rpkm,GSM762072_col_0_seedling_tr3_rpkm,GSM762073_col_0_seedl
AT1G06620,0.71,0.85,1.01,3.746427,3.604144,5.547661,3.986773,2.9316,2.925005,1.390748,0.787959
AT1G38440,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT1TE49290,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT1TE49295,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT5TE41615,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT2G34630,16.76,18.72,21.15,28.771537,30.159349,24.743735,26.182941,23.618868,24.340921,22.556
AT5TE41610,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT2G25590,5.93,5.87,4.82,1.226086,0.623488,0.20173,0.488901,2.558452,3.230756,4.949729,3.70764
AT1G08710,7.19,3.45,5.82,7.16511,8.144506,5.409022,8.57125,7.585589,7.58909,5.808602,7.806521,
AT2G30780,6.19,5.85,6.2,7.161249,6.278675,5.484976,3.987914,6.086797,4.771724,6.18765,5.133824
AT5G25130,17.08,14.94,16.4,4.894696,5.309965,3.060268,4.489041,4.851495,6.592106,13.35501,10.9
AT2G32280,2.54,1.96,3.39,5.372782,4.558652,5.009301,4.284787,3.970664,2.883908,5.73414,3.15916
AT5TE46960,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
AT5TE46965,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

Figure 2.A2. Example of a self-curated coexpression CSV file.

Like SOFT files, CSV files can be opened and modified in all text editors, but, given their tabular nature, Microsoft Excel is also a good alternative. However, users of Excel should be wary of autocorrect functions that may inadvertently change gene names or other data [10].

In addition to the expression data, the CSV file should contain:

- Title or name of the dataset as a comment in the file's header, e.g., `#title: mydataset1`.
- Short description of the dataset as a comment in the file's header, e.g., `#desc: mydescription1`.

2.A3. Methods

2.A3.1. Selecting a PlantiSMASH Version

Currently, plantiSMASH is available in two versions: web and stand-alone. Both versions offer all basic and default functionalities, with the stand-alone version providing extra fine-tuning of the parameters and thresholds for all the analyses.

PlantiSMASH Web Server Advantages:

- No installation required.

- Minimal knowledge of the tool's parameters and significance is needed.

PlantiSMASH Stand-Alone Advantages:

- Customizable cluster-calling rules and pHMM models for targeted predictions.
- The values of parameters and thresholds for several analyses may be modified, such as the CD-HIT [11] identity threshold for cluster prediction, or the maximum MAD for coexpression analysis, among others.
- Safer for protection of proprietary data (potentially, depending on private server setup).
- No upload time: Large genomes can take a long time to upload to the web server, which has a strict upload timeout of around 300 seconds to prevent service denial.
- No job-queuing waiting time, which may vary according to the web server's load.

2.A3.2. Preparing Input Files

Some of plantiSMASH's multiple input types may require additional preparation to ensure that the algorithm works as intended.

Annotated GenBank (.gb, .gbk) or EMBL file: Properly annotated genomes in GenBank or EMBL format need no preparation; plantiSMASH will process them out of the box.

FASTA sequence with GFF3 annotations: The annotation file must adhere to the GFF3 specification standards as developed and maintained by The Sequence Ontology Project. To confirm that a GFF file complies to the GFF3 format, third-party tools such as "GFF3Validator" (http://genometools.org/cgi-bin/gff3validator.cgi) can be used. The record identifiers of the FASTA and GFF3 files must match. If only one record is present in both inputs, it is assumed that they refer to the same scaffold/contig/chromosome, and only coordinates must match (e.g., no annotation should point to coordinates beyond the sequence's length). Currently, plantiSMASH only accepts one GFF3 file. To submit annotations for multiple sequences, GFF3 files can be concatenated into one; special consideration must be given to ensure that no redundant record identifiers are present in the resulting file, and that gene IDs within an entry are unique.

FASTA sequence only: Without an annotation file as input, plantiSMASH will use its gene-finding module to annotate the genome before executing other steps. On the stand-alone version of plantiSMASH, gene finding may be skipped by considering all possible ORFs more than sixty base pairs long to define the CDS instead. This can be done with the `all-orfs` option. However, users should note that this microbially inspired option will frequently lead to artifacts in plant genomes, as multiple exons may be called as separate genes.

Gene Expression Data: Features, or locus tags, listed in the gene expression dataset must have the same nomenclature as the annotations provided to ensure that plantiSMASH matches them correctly. This problem may arise when annotations change nomenclature format as versions progress, or when an expression dataset lists transcripts instead of genes.

Additional Considerations:

- **CSV (Comma-Separated Values):** For proper coexpression analysis, it is important that expression data is normalized beforehand by the user according to their needs [12]. No gene or locus tag in the expression file may appear more than once. This can be an issue with microarray experiments, in which the relation of probes and genes is not always one to one. The simplest solution is to remove rows containing duplicate identifiers beforehand. Comments notwithstanding, the CSV format must describe a rectangular array or matrix: all rows must have the same number of columns, and all columns must have the same number of rows.

- **NCBI GEO SOFT file:** First, one needs to check whether the SOFT file contains a complete gene expression dataset. The fastest way to achieve this is to open the SOFT file with a text editor and find the string `!sample_table_begin`, below which the expression data should appear.

2.A3.3. Choosing Parameters

Enable Additional Analyses:

- **Coexpression:** We still do not know the extent to which genomic clustering of biosynthetic genes is indicative of their co-involvement in a plant secondary metabolic pathway , and prioritization of high-potential BGCs is important to guide further analysis. Coexpression has previously proven to be a powerful technique for this [13,14]. PlantiSMASH gives its users a way to prioritize those candidate BGCs by leveraging multiple coexpression datasets and visualizing them in a user-friendly way. To do that, the user needs to supply gene expression data (either from microarray or transcriptomic experiments) in a specific format alongside the genome file.

- **ClusterBLAST:** While the definite origins of plant BGCs are still unclear, many traces of evolutionary relationships between BGCs in closely related plant species have been observed, such as for those encoding the biosynthesis of the antimicrobial diterpene phytoalexins in the *Oryza* family [15]. Using ClusterBLAST, users can get an overview of similar BGCs in other plant species, using a precalculated database of plantiSMASH results generated from publicly available high-quality plant genomes.

- **KnownClusterBLAST:** In addition, the KnownClusterBLAST module can be used for a quick identification of clusters resembling previously known ones. To do this, plantiSMASH refers to a database of known BGCs (http://mibig.secondarymetabolites.org/) and performs a regular ClusterBLAST against it.

Adjust Algorithm Stringency:

- **CD-HIT cutoff:** To prevent the inclusion of duplicated tandem arrays, plantiSMASH by default uses a 50% identity threshold (0.5 CD-HIT cutoff) to cluster similar protein-coding genes into groups that represent the “unique classes of enzyme-coding genes” within a cluster. While this works well for regular usage, users can alter this behavior by applying a more stringent threshold (lower value down to 0.2) or more relaxed one (higher, up to 1.0).

- **Minimum # of unique domains:** By default, plantiSMASH will include clusters containing at least two different classes of enzymes (putatively assigned using the Pfam database [16], and compressed with CD-HIT; see above). To include clusters of one kind (e.g., groups of p450s), you can set this parameter to 1 instead. Alternatively, when the user is only interested in more “complex” clusters, the parameter can be set to a higher value.

- **MAD Cutoff (for coexpression analysis):** The median absolute deviation of an array is a measurement of its variability, and it can be calculated according to a specific equation (see below). Regarding gene expression, MAD is an assessment of how much each gene’s expression changes across all samples, and it is a useful tool to weed out coexpression with, or among, housekeeping genes that are unlikely to be related to specialized metabolite production in a particular experiment. Setting a proper threshold for what constitutes a gene with low or high variability depends on several factors and is a parameter best chosen by the user. By default, plantiSMASH will filter out genes with a MAD of zero (the lowest value possible) from the coexpression analysis. This will remove genes from the analysis for which the expression values remain unchanged (e.g., at zero) across all samples. The MAD is calculated with the following formula:

$$\mathrm{MAD}{=}\mathrm{median}\left(\left\|Xi{-}\mathrm{median}(X)\right\|\right)$$

Advanced Settings (Stand-Alone Version Only):

- **Cutoff Multiplier:** PlantiSMASH uses different distance cutoffs between biosynthetic genes for cluster calling in each genomic region, based on its local gene density. For example, the cluster sizes of Arabidopsis thaliana can range from fifteen to four hundreds kb on the extreme. To alter that calculation, users can use the parameter `--cutoff-multiplier` and specify how large the clusters are allowed to be, e.g., 2 for twice the original values.

- **Full HMMer:** By default, plantiSMASH will only scan for pHMMs listed in its library (around 63 pHMMs). Hits from those pHMMs are used to define the biosynthetic genes, and this information will be retained on the resulting output. Sometimes, it would also be useful to have full information on all protein domains (from the Pfam-A database) present in the cluster (even for the non-biosynthetic genes). Users can enable this mode by using the `--full-hmmer` parameter. As the calculation can take significantly longer, this mode is disabled by default.

- **All ORFs:** In case gene finding did not work, users can also resort to including all possible ORFs in the genome by using `--all-orfs` parameter. Users should be aware that this is likely to yield very-low-quality gene predictions and will increase computation time significantly.

2.A3.4. Running the Analysis Shell:

Web Server Version:

- Access plantiSMASH's site at http://plantismash.secondarymetabolites.org/. Figure 2.A3 shows the input form as available at the time of writing.

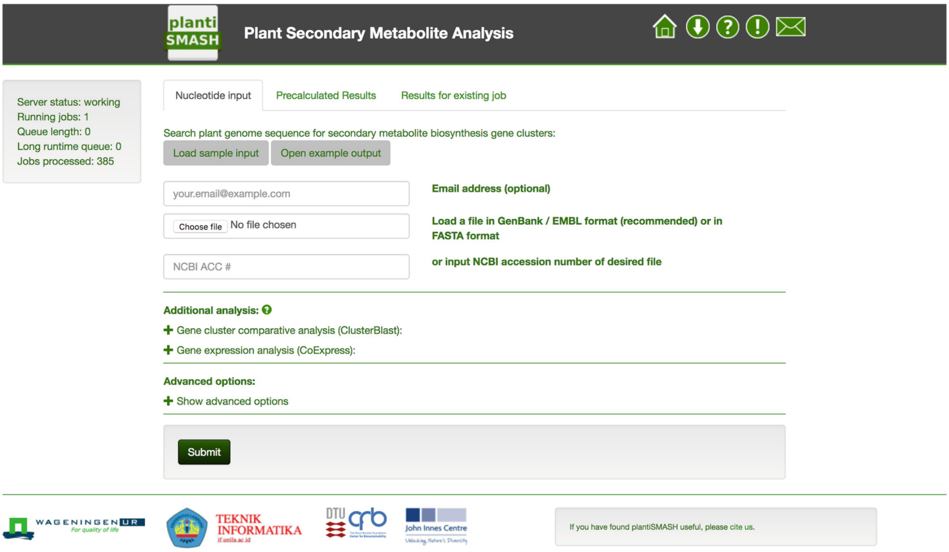
- Fill out your email address. This is an optional step, but is a convenient way to track your plantiSMASH job that may run from minutes to hours.

- Upload EMBL/GBK/Fasta and GFF3 files via the provided column. In case of a public genome data available via NCBI website (https://www-ncbi-nlm-nih-gov.ezproxy.library.wur.nl/genome/browse/), the RefSeq/Genbank accession number may be provided instead. When using a whole-genome shotgun sequencing project entry, one can provide the accession of the master record; the web server will then collect all the corresponding scaffold or contig entries from NCBI.

- Select additional analyses to be performed. For coexpression analysis, also upload the SOFT / CSV file(s) via the provided column.

- Some parameters may be adjusted via the "Advanced options" tab. Leave the columns unchanged to use default settings.

- Click on the "Submit" button.



Server status: working
Running jobs: 1
Queue length: 0
Long runtime queue: 0
Jobs processed: 385

Search plant genome sequence for secondary metabolite biosynthesis gene clusters:

your.email@example.com — Email address (optional)

Choose file No file chosen — Load a file in GenBank / EMBL format (recommended) or in FASTA format

NCBI ACC # — or input NCBI accession number of desired file

Additional analysis:

+ Gene cluster comparative analysis (ClusterBlast):
+ Gene expression analysis (CoExpress):

Advanced options:

+ Show advanced options
Submit
WAGENINGEN UR For quality of life
TEKNIK INFORMATIKA if.unila.ac.id
DTU cfb
John Innes Centre Unlocking Nature's Diversity

If you have found plantiSMASH useful, please cite us.

Figure 2.A3. Input form of plantiSMASH webserver.

Stand-Alone Version:

- Open terminal, and change directory to the working plantiSMASH folder.
- Input to the terminal:

```
python2 run_antismash.py --taxon plants
<additional_parameters> <sequence_files>
```

- Additional parameters:
    - `--gff3 <filename>`: if using fasta+gff3 input, specify the gff3 file path here
    - `--coexpress`: will enable CoExpression analysis, coupled with
    - `--coexpress-soft_file <filenames>`: path to SOFT file(s), separated by comma
    - `--coexpress-csv_file <filenames>`: path to CSV file(s), separated by comma
    - `--clusterblast`: will enable ClusterBlast analysis

- ○ `--knownclusterblast`: will enable KnownClusterBlast analysis
- ○ `--cdh-cutoff <0.2-1.0>`: will set the CD-HIT cutoff (default: 0.5)
- ○ `--min-domain-number <value>`: will set the minimum unique domains threshold (default: 2)
- ○ `--coexpress-min_MAD <0-1.0>`: will set the minimum absolute deviation for CoExpression analysis (default: 0/disabled)
- ○ `--cutoff-multiplier <value>`: will set a multiplier for the cluster kb-size stringency (default: 1.00)
- ○ `--full-hmmer`: will enable full Pfam-A based domain annotation for every genes in the genome
- ○ `--all-orfs`: will use every possible ORFs >60 bp to define the CDS instead of genefinding in the case of no annotation file provided
- ○ `--outputfolder <path>`: will set an alternative folder for the output files (default: a folder with the same name as the input file in current directory)

- When done, plantiSMASH will create a folder in the current directory with the same name as the input files (unless a custom output folder is specified). To view the visualized result directly, open the file index.html in that folder.

2.A4. Interpreting Results

When the run is finished, an HTML file will be generated that provides a visual overview of the results. On this page, all BGC predictions are listed in numerical order. The nucleotide record in which each BGC is located, and their coordinates and size, is also listed. Additionally, the protein domains found in the cluster, and the number of functionally different protein-coding genes (CD-HIT Clusters), are also provided; the overview table can be sorted on these data, allowing quick identification of clusters that may be of interest.

If KnownClusterBLAST was enabled, the most similar known cluster is shown, along with its MIBiG identifier. Clicking on any cluster's MIBiG accession number takes the user to the corresponding cluster view, where each cluster can be analyzed independently.

2.A4.1. Clusters with Complex Architectures

The number of CD-HIT clusters for each gene cluster represents the number of functionally different protein (sub-)families it encodes. This can be used to highlight gene clusters with complex architectures that produce diverse types of enzymes related to specialized metabolite biosynthesis. An example can be seen with Medicago truncatula, containing a predicted saccharide-terpene cluster in chromosome 6 with 10 CD-HIT clusters, pictured in Figure 2.A4.

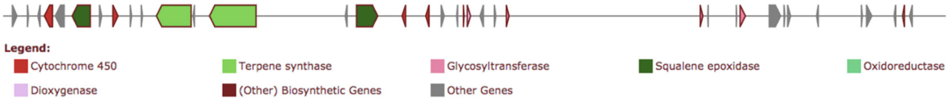


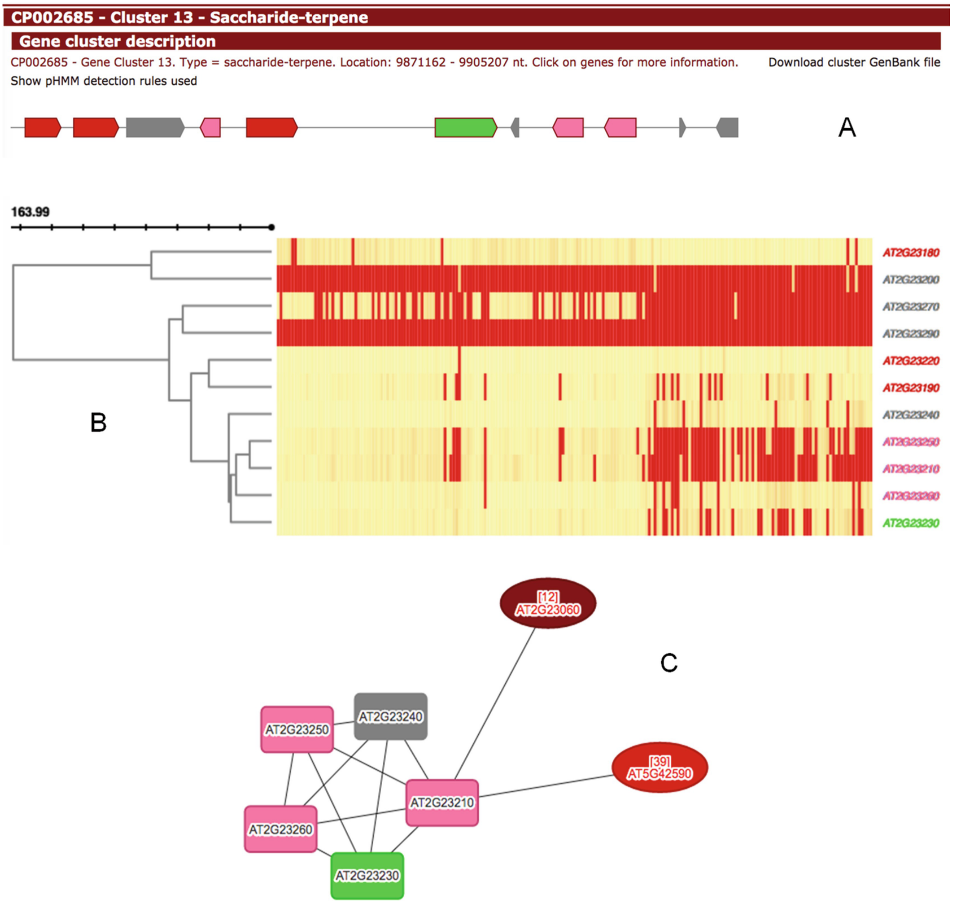
Figure 2.A4. Saccharide-Terpene cluster in chromosome 6 of *M. truncatula*.

2.A4.2. Observing Coexpression Patterns

As demonstrated by Itkin et al. [13] and Boutanaev et al. [14], coexpression analysis can be very valuable for metabolic pathway discovery. When a gene expression file is provided as input, plantiSMASH can guide BGC prioritization to this end.

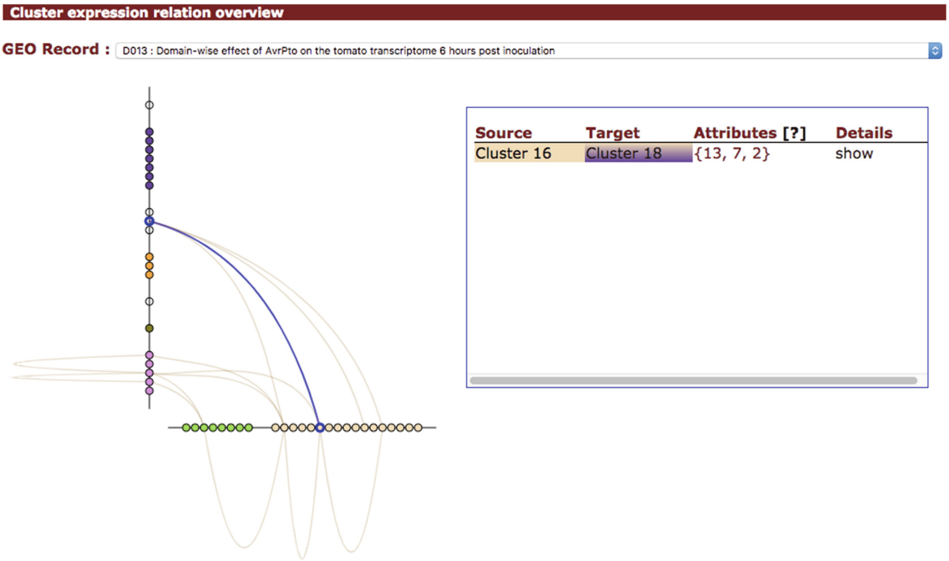
First, the expression of all genes in a cluster can be examined with an expression heatmap, as seen in Figure 2.A5B, which can be normalized per gene (row) or per sample (column). For easy interpretation, the color of the locus tags to the right of the heatmap matches the color of the gene in the cluster view shown in Figure 2.A5A. A coexpression network is also generated automatically, as seen in Figure 2.A5C, which can be examined and redrawn with a new distance threshold to determine which edges are drawn between the nodes. The coexpression between genes in the cluster of interest and genes in other clusters is also shown if significant; in this case, genes from other clusters are drawn as ovals and marked with the cluster number between brackets. Additionally, edges also show the coexpression among gene pairs with the Pearson correlation coefficient (PCC).

In this example, four of the cluster's biosynthetic (non-grey) genes are located in a clade at the bottom with short distances between each other, suggesting that they are coexpressed. This is evidenced by the coexpression network, showing a module of five genes coexpressed with each other, one of which is also coexpressed with biosynthetic genes from other clusters. Similarly, plantiSMASH will also highlight BGC genes that are coexpressed among different clusters. This is shown with a hive plot, where each node in any of the two axes represents a cluster, and edges connect clusters that show significant coexpression. This can highlight possible pathways encoded in more than one BGC, such as the experimentally characterized α-tomatine pathway in *S. lycopersicum* [13]. In Figure 2.A6, we can observe significant coexpression between the two clusters that are necessary for the production of this metabolite.



CP002685 - Cluster 13 - Saccharide-terpene
Gene cluster description
CP002685 - Gene Cluster 13. Type = saccharide-terpene. Location: 9871162 - 9905207 nt. Click on genes for more information.
Download cluster GenBank file
Show pHMM detection rules used
A
163.99
AT2G23180
AT2G23200
AT2G23270
AT2G23290
AT2G23220
AT2G23190
AT2G23240
AT2G23250
AT2G23210
AT2G23260
AT2G23230
B
[12]
AT2G23060
C
AT2G23250
AT2G23240
[39]
AT5G42590
AT2G23210
AT2G23260
AT2G23230

Figure 2.A5. (A) Cluster view, showing the selected cluster's structure, its genes, and their annotations. (B) Gene expression heat map. Each row represents a different gene from the selected cluster, and each column a different sample of the expression dataset input. (C) Co-expression network between the genes of the selected cluster. Genes from other clusters are also included if significantly coexpressed.



Cluster expression relation overview

GEO Record : D013 : Domain-wise effect of AvrPto on the tomato transcriptome 6 hours post inoculation

Source	Target	Attributes [?]	Details
Cluster 16	Cluster 18	{13, 7, 2}	show

Figure 2.A6. Hive plot showing inter-cluster coexpression in a particular expression dataset in *S. lycopersicum*. Clusters are represented by nodes, and edges represent significant co-expression. The highlighted edge shows significant coexpression between cluster 16 and cluster 18, which together encode the α-tomatine biosynthesis pathway.

2.A4.3. Additional Analysis

Several bioinformatics analyses can be done with plantiSMASH results. One example of this is a statistical analysis based on coexpression, as presented in the plantiSMASH paper [8]. Here, we showed that 12 of the 42 predicted *A. thaliana* clusters score within the significance threshold by using the Mann-Whitney U test, seen in Figure 2.A7. The test was performed with the alternative hypothesis that the coexpression among genes in each predicted cluster is stochastically greater than the coexpression among other genes in close physical proximity to each other throughout the genome (source code for this analysis is available from https://bitbucket.org/herl91/testclustercoexpression/).

2.A4.4. Additional Info: Loading GBK and EMBL Results

While useful and informative, the main drawback of the default html output is the inability to browse beyond the clusters. To address that, both the web server and stand-alone version plantiSMASH also output GBK and EMBL files that can be loaded into genome browsers like Artemis [17] or IGV [18].

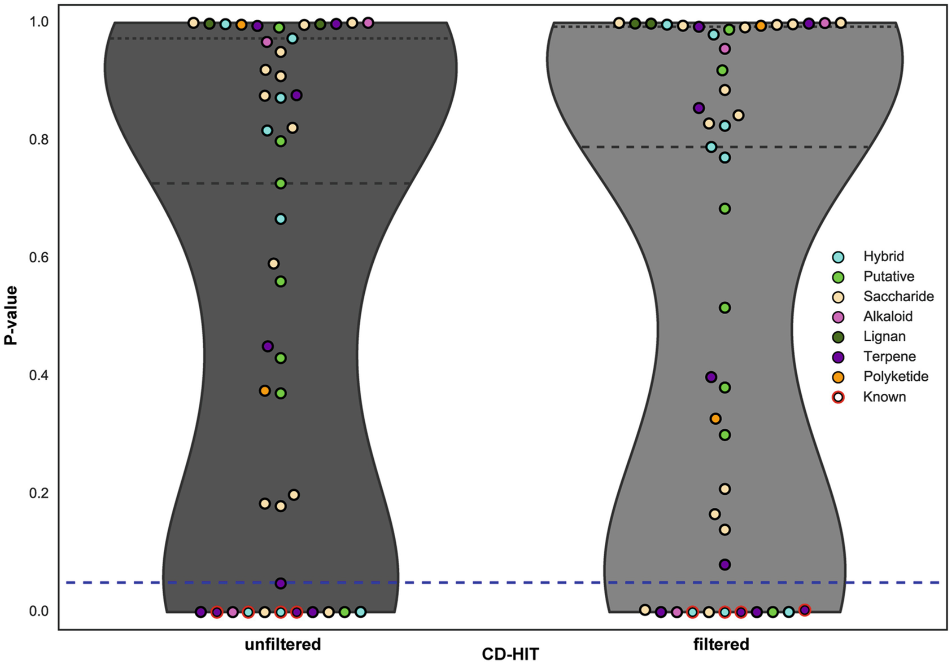


Figure 2.A7. Mann-Whitney U test results of the 42 predicted *A. thaliana* clusters when tested with the alternative hypothesis that the coexpression among the predicted clusters is stochastically greater than the coexpression between genes in close physical distance to each other throughout the genome. Clusters are represented with circles coloured according to their cluster type. Known clusters are shown as circles outlined in red. P=0.05 is defined by the blue dashed line. The black dashed lines are each distribution's second and third quartile. Left: P distribution without discarding gene-pairs in the same CD-HIT cluster. Right: P distribution when discarding gene-pairs in the same CD-HIT cluster. Multiple testing correction was not necessary because each comparison is among different samples, with background distributions chosen for each cluster size. Nonetheless, using Bonferroni correction would only remove one cluster from the list of significantly coexpressed clusters.

Acknowledgments

This work was supported by a VENI grant [863.15.002 to M.H.M.] from The Netherlands Organization for Scientific Research (NWO) and by the Graduate School for Experimental Plant Sciences (EPS).

References

1. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Nat Prod Rep. 2016;33: 951–962.
2. Nützmann H-W, Osbourn A. Gene clustering in plant specialized metabolism. Curr Opin Biotechnol. 2014;26: 91–99.
3. Boycheva S, Daviet L, Wolfender J-L, Fitzpatrick TB. The rise of operon-like gene clusters in plants. Trends Plant Sci. 2014;19: 447–459.
4. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters - from genetics to genomics. New Phytol. 2016;211: 771–789.
5. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiol. 2017;173: 2041–2059.
6. Töpfer N, Fuchs L-M, Aharoni A. The PhytoClust tool for metabolic gene clusters discovery in

plant genomes. Nucleic Acids Res. 2017;45: 7049–7063.

7. Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell. 2017;29: 944–959.
8. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 2017;45: W55–W63.
9. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339–46.
10. Mallona I, Peinado MA. Truke, a web tool to check for and handle excel misidentified gene symbols. BMC Genomics. 2017;18: 242.
11. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.
12. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. Learning from Co-expression Networks: Possibilities and Challenges. Front Plant Sci. 2016;7: 444.
13. Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science. 2013;341: 175–179.
14. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al. Investigation of terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 2015;112: E81–8.
15. Miyamoto K, Fujita M, Shenton MR, Akashi S, Sugawara C, Sakai A, et al. Evolutionary trajectory of phytoalexin biosynthetic gene clusters in rice. Plant J. 2016;87: 293–304.
16. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–85.
17. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28: 464–469.
18. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

Chapter 3

An Investigation of Sesterterpene Gene Clusters in *Brassicaceae*

Sesterterpenoids, a terpene subclass with five-isoprenoid (C_{25}) backbones, are widely distributed (i.e., they can be found in multiple domains of life) but collectively rare in nature. Although their biological roles remain unexplored, many isolated sesterterpenes showed promising biochemical activities. In fungi, these compounds are synthesized by bifunctional enzymes (colloquially termed sesterterpene synthases, or STS) containing both the precursor-producing prenyltransferase (PT) and the cyclization (TPS) domains. As the coupling of these biosynthetic domains allows efficient production of the terpenoid scaffolds, we could expect convergent evolution of the pair, either within a single gene or as co-localized genes (BGC) in the genomes. In this chapter, I used plantiSMASH to explore putative STS-encoding BGCs in 55 publicly available high-quality plant genomes. Together with collaborators, I sought to understand the evolutionary relationships and chemical products of these BGCs. In the end, we confirmed the production of seven structurally diverse sesterterpenes from three Brassicaceae species via transient expression of their TPS enzymes in tobacco.

Originally published as:

Huang, A. C., Kautsar, S. A., Hong, Y. J., Medema, M. H., Bond, A. D.,
Tantillo, D. J., & Osbourn, A. (2017).
Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae
through genome mining reveals convergent evolution.
Proceedings of the National Academy of Sciences, 114(29), E6005-E6014.

3.1. Introduction

Sesterterpenoids are a largely unexplored class of terpenes with only around 1,000 members isolated from nature so far, representing a mere <2% of the reported terpene family members (>70,000). These compounds are structurally diverse and have a wide spectrum of biological activities, ranging from anti-inflammatory, anticancer, cytotoxic, and antimicrobial bioactivities to phytotoxicity and plant defense [1]. Recent work on sesterterpenoids has suggested that this terpene class could be an important new source of anticancer drugs [2,3]. The majority of sesterterpenoids characterized to date are from marine sponges and terrestrial fungi, with only 60–70 being of plant origin. Many of these plant sesterterpenes were isolated from the mint family (*Lamiaceae*) and have been implicated in plant defense [4–7]. Although large transcriptome datasets for the mint family have been generated [8], very limited genome sequences for the Lamiaceae are currently available [9,10].

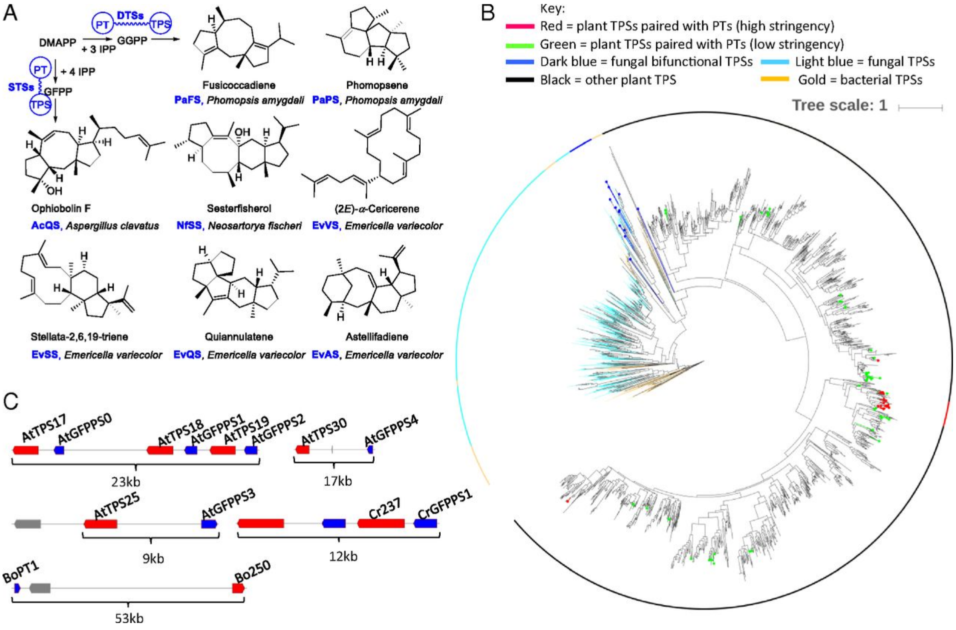
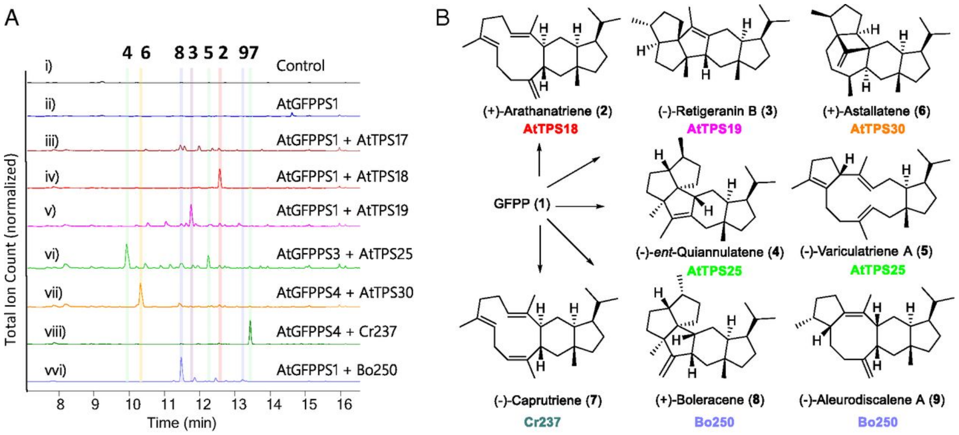


Figure 3.1. Identification of candidate STS genes from plant genomes. (A) Chemical structures of diterpenes and sesterterpenes synthesized by fungal bifunctional DTSs and STSs. (B) Phylogenetic tree constructed using TPSs from 55 plant genomes together with fungal and bacterial TPSs. The TPSs for which the genes were found to be colocalized with PT genes with high stringency are shown in red, and those with low stringency in green. Bifunctional fungal TPSs are shown in dark blue. (C) A graphic representation of PT-TPS gene pairs from *A. thaliana*, *C. rubella*, and *B. oleracea* selected for functional validation in this work.

Like other classes of terpenes, the structural diversity of sesterterpenes largely originates from the very first scaffold-generating step, which in this case is catalyzed by sesterterpene synthases (STSs), a class of terpene synthase (TPS). In fungi, six STSs have recently been identified and shown to synthesize ophiobolin F (in *Aspergillus clavatus*) [11], sesterfisherol (in *Neosartorya fischeri*)

[12], and stellata-2,6,19-triene, (2E)-α-cericerene, quiannulatene, and astellifadiene (in *Emericella variecolor*) [13–16] (Figure 3.1). These fungal enzymes are all bifunctional, containing a C-terminal trans-prenyltransferase (PT) and an N-terminal TPSs domain. The PT domain synthesizes the universal C25 sesterterpene precursor geranylfarnesyl diphosphate (GFPP) (**1**) from dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP), which is then cyclized by the TPS domain to form diverse scaffolds. Some fungal diterpene synthases (DTSs), such as Phomopsis amygdali fusicoccadiene synthase [17,18] and phomopsene synthase [19], are also bifunctional. Fusion of PT and TPS activities has been proposed to contribute to more efficient production of bioactive terpenes from precursor supply [18,20]. In most cases, however, the condensation and cyclization steps for diterpene biosynthesis in fungi are carried out by individual enzymes encoded by genes that are sometimes colocalized within fungal biosynthetic gene clusters [21].

In plants, two TPSs from *A. thaliana* were recently reported to synthesize the sesterterpenes arathanatriene (**2**) and retigeranin B (**3**) [22]. Intriguingly, these two TPSs are colocalized with PTs that synthesize the precursor substrate GFPP. To our knowledge, bifunctional TPSs that contain both PT and TPS domains have not been reported from any plant species. This raises questions regarding the landscape of PT and TPS genes and their related functions in plant genomes. We have previously developed algorithms for mining plant genomes for genes encoding natural product biosynthetic pathways [23,24]. Here, we carried out a systematic search of 55 sequenced plant genomes using a customized version of the plantiSMASH genome mining algorithm [24] to identify genes predicted to encode PT or TPS domains. Although we did not find evidence for bifunctional genes encoding both PT and TPS domains, we identified a pool of colocalized pairs of PT and TPS genes using this algorithm. Phylogenetic analysis of all TPSs from the 55 plant genomes that were investigated identified a single clade containing those TPS genes that were physically clustered with PT genes, greatly expanding on previous observations [25]. Transient expression of selected TPS genes from this group from A. thaliana, C. rubella, and B. oleracea in heterologous host N. benthamiana revealed five additional TPSs that make a variety of previously unknown fungal-type sesterterpenes with diverse scaffolds, including 11/6/5 tricyclic (−)-caprutriene (7), 5/12/5 tricyclic (−)-variculatriene A (5), 5/8/6/5 tetracyclic (−)-aleurodiscalene A (9), 5/5/5/6/5 pentacyclic (−)-ent-quiannulatene (4) and (+)-boleracene (8), and 5/4/7/6/5 pentacyclic (+)-astellatene (6) (Figure 3.2B, Supplementary Tables S1–S8 and Supplementary Dataset S1). Among these scaffolds, (−)-ent-quiannulatene (4) is an enantiomer of the fungal metabolite (+)-quiannulatene. Homology modeling and sequence comparisons implicated a key amino acid as being likely to be important for scaffold diversification. Site-directed mutagenesis of this residue further expanded scaffold diversity and, coupled with quantum chemical calculations, shed light on the cyclization mechanisms leading to the formation of the pentacyclic sesterterpenes (−)-retigeranin B (3) and (+)-astellatene (6). The carbocation cascade sequences following protonation of the GFPP substrate in the active site of plant STSs likely mirror those of the distantly related fungal bifunctional STSs. However, phylogenetic analysis suggests the independent evolution of fungal and plant STSs, indicating that the plant and fungal enzymes have arisen by convergent evolution.



C

Species	TPS	Compounds											
		2	3	4	5	6	7	8	9	14	15	16	Sum %
A. thaliana	AtTPS17	-	-	-	-	-	-	-	-	-	-	-	-
A. thaliana	AtTPS18	89	-	-	-	-	-	-	-	-	-	-	89
A. thaliana	AtTPS19	5.5	42	-	-	-	-	-	-	5.4	4.8	0	57.7
A. thaliana	AtTPS25	3.4	-	44	16	-	-	-	-	-	-	-	63.4
A. thaliana	AtTPS30	5.3	-	-	-	73	-	-	-	-	-	-	78.3
C. rubella	Cr237	-	-	-	-	-	91	-	-	-	-	-	91
B. oleracea	Bo250	-	-	-	-	-	-	81	4.5	-	-	-	85.5
A. thaliana	AtTPS19^{428Asp}	12	30	-	-	-	-	-	-	14	9.7	28	93.7
A. thaliana	AtTPS18^{431Tyr}	86	-	-	-	-	-	-	-	-	-	-	86

Figure 3.2. Functional analysis of candidate STSs using transient tobacco expression. (A) Comparative GC-MS total ion chromatograms (TICs) of extracts from *N. benthamiana* leaves transiently coexpressing different combinations of GFPPS–TPS gene pairs. *A. tumefaciens* LBA4404 carrying the green fluorescent protein gene in pEAQ-HT was used as a control. (B) Chemical structures of compounds **2-9** synthesized by the selected TPSs (AtTPS17, -18, -19, -25, -30, -Cr237, and Bo250). (C) Extracted ion chromatogram (m/z = 340) area percentages of the identified compounds in the total sesterterpene mixtures generated by the different TPSs. Sum area percentages represent the percentages of all characterized compounds in the total sesterterpenes made by each TPS. Note that the sum does not amount to 100% because some uncharacterized and very minor sesterterpenes are not included.

Although mining plant genomes and transcriptomes for individual PTs and TPSs has proven to be effective for uncovering terpene diversity [26–28], to our knowledge systematic genome mining for colocalized PT and TPS genes has not previously been attempted. Collectively, our work sheds light on the genomic organization of plant PT and STS genes and demonstrates an effective pipeline for plant sesterterpene discovery, from finding genes to production by engineering in the heterologous host *N. benthamiana* and further manipulation by protein engineering.

3.2. Results

3.2.1. Landscape of PT and TPS Genes in Plant Genomes and Identification of a Phylogenetic Clade of Candidate Plant STSs Using a Customized Genome Mining Algorithm.

To search for genes with predicted PT and TPS domains across 55 sequenced high-quality plant genomes (Supplementary Dataset S2), we used the Polyprenyl_synt (PT) and Terpene_synth C (TPS) models from the Pfam database [29] in a version of plantiSMASH [24] with gene cluster detection logic customized for this specific purpose. The Terpene_synth_C pHMM was seeded from plant, fungal, and bacterial TPSs and so is applicable across plants and microbes. Although we could not identify any candidate bifunctional TPS genes, we were able to identify 21 colocalized PT-TPS gene pairs using our default (high-stringency) parameters (cutoff 25±5 kb; deviation calculated using gene density/average intergenic distance for the neighboring ten genes as a multiplier) (Supplementary Dataset S2). An additional 75 candidate PT-TPS gene pairs were identified at lower stringency (allowing for distances between the PT and TPS domains that were up to five times larger) (Supplementary Dataset S2). We next constructed a phylogenetic tree (Figure 3.1B) using 2,846 TPS sequences extracted from the 55 plant genomes (Supplementary Dataset S2), together with 724 fungal TPSs (including the known bifunctional TPSs). A total of 278 bacterial TPSs was included as an outgroup.

This phylogenetic analysis enabled us to identify a single TPS clade in which the TPS genes that were colocalized with PT genes as identified using the high-stringency cut-off were heavily represented (Figure 3.1B). Nine of the PTs in these PT-TPS gene pairs have indeed been previously shown to be functional plant GFPPSs [25,30], and two of the TPSs (AtTPS18 and AtTPS19 from *A. thaliana*) were shown to biosynthesize the sesterterpenes (+)-arathanatriene and (−)-retigeranin B [22]. However, two other TPSs from these pairs (AtTPS25/At3g29410 and AtTPS30/At3g32030 *from A. thaliana*) have recently been reported to have diterpene or sesquiterpene activities when expressed in *Escherichia coli* [31].

A total of 18 TPS genes, including AtTPS25 and AtTPS30, that are colocalized with PT genes were grouped together in this clade (marked in red, Figure 3.1B and Supplementary Figure S1A), greatly outnumbering previous observations [25]. Two outlier TPS genes (one from *V. angularis* and one from *B. napus*, marked as red dots in Figure 3.1B) from the 21 PT-TPS gene pairs identified using high-stringency cutoff were located in other clades (Figure 3.1B and Supplementary Figure S1B). One TPS gene from these 21 PT-TPS gene pairs was filtered out during tree construction. The characterized fungal bifunctional TPSs (including STSs and DTSs) grouped together in two closely related clades that were clearly distinct from this plant TPS clade, suggesting that the plant TPS (red in Figure 3.1B) and the fungal bifunctional STS clades (blue in Figure 3.1B) have evolved independently. Other TPS genes that colocalized with PT genes over larger distances (low stringency) were spread across different clades (green dots in Figure 3.1B). To understand the functions of the TPSs that emerged from our search and investigate the significance of this phylogenetic grouping, we chose to functionally test seven TPSs (five from *A. thaliana*, one from *C. rubella*, and one from *B. oleracea*) for which the genes are clustered with PT genes from the main clade (Figure 3.1C and Supplementary Figure S1A), along with two other *A. thaliana* TPSs (AtTPS22/At1g33750 and AtTPS29/At1g31950) (Supplementary Figure S1A) for which the genes are not clustered with a PT gene yet are part of the same monophyletic branch within the TPS phylogeny.

3.2.2. AtTPS17, -18, -19, -25, and -30 from A. thaliana, Cr237 from C. rubella, and Bo250 from B. oleracea Are all Functional STSs.

The selected TPSs, AtTPS17 (At3g14490), AtTPS18 (At3g14520), AtTPS19 (At3g14520), AtTPS25 (At3g29410), and AtTPS30 (At3g32030) from *A. thaliana*, Cr237 (CARUB_v10016237mg) from *C. rubella*, and Bo250 (LOC106343250) from *B. oleracea* all have corresponding clustered PT genes, namely AtGFPPS0 (At3g14510), AtGFPPS1 (At3g14530), AtGFPPS2 (At3g14550), AtGFPPS3 (At3g29430), AtGFPPS4 (At3g32040), CrGFPPS1 (CARUB_v10014047mg), and BoPT1 (LOC106293690), respectively (Figure 3.1C). In contrast, AtTPS22 and AtTPS29 are not paired with any PT genes in the genome. AtGFPPS0 is a pseudo gene encoding a nonfunctional protein with a frame shift, whereas AtGFPPS1-4 and CrGFPPS1 are characterized GFPPSs [25,30]. BoPT1 has not been previously identified or characterized. N-terminal transit peptide (N-tp)–truncated recombinant AtTPS18 and AtTPS19 proteins were recently coexpressed with N-tp AtGFPPS1 and AtGFPPS2 in *E. coli* and reported to synthesize (+)-arathanatriene and (−)-retigeranin B, respectively [22]. Recombinant proteins encoded by N-tp–truncated AtTPS22 and AtTPS25 were also recently expressed in *E. coli* but reported to have sesquiterpene synthase activity, making sesquiterpene mixtures when incubated with E,E-farnesyl diphosphate (E,E-FPP) as a substrate *in vitro* [31]. Recombinant proteins encoded by N-tp–truncated AtTPS30 were found to have DTS activities when assayed with GGPP or ent-copalyl diphosphate (ent-CPP) as substrates *in vitro*, whereas AtTPS29 was not active in these assays [31].

AtTPS17, Cr237, and Bo250 have not been investigated previously. We generated expression constructs for each of seven TPSs from the PT-TPS gene pairs (*A. thaliana* AtTPS17, -18, -19, -25, and 30; *C. rubella* Cr237; *B. oleraceae* Bo250), three AtGFPPSs (*A. thaliana* AtGFPPS1, -3, -4), and the two nonclustered *A. thaliana* TPS AtTPS22 and AtTPS29 genes using a vector designed for transient expression in *N. benthamiana* (pEAQ-HT) [32]. These constructs, each harboring individual genes, were transformed into *A. tumefaciens*. Strains containing the different expression constructs were then used for transient expression in *N. benthamiana* leaves. This tobacco transient-expression platform allows high-level expression of genes of interest, especially those of plant origin, thus enabling us to rapidly test the biochemical functions of individual genes [33]. Furthermore, different combinations of genes can be coexpressed simply by mixing A. tumefaciens strains containing the different expression constructs and co-infiltrating these into the leaves [33].

When we expressed all nine TPSs (AtTPS17, -18, -19, -25, -30, -22, -29, Cr237, and Bo250) individually in *N. benthamiana*, trace compounds 2, 3, 4, 6, 7, and 8 were detected from extracts of leaves expressing AtTPS18, -19, -25, -30, Cr237, and Bo250 (Supplementary Figure S2B). However, we did not observe the formation of sesquiterpenes or diterpenes, even for AtTPS22, -25, -29, and -30, which had previously been reported to synthesize these types of terpenes *in vitro* [31]. This outcome may be because of the low activity of the encoding enzymes, inadequate precursor supply, or the loss of the volatile/semivolatile products in *N. benthamiana*, especially if these products were produced in very low abundance. Expression of AtGFPPS1, -3, and -4 alone in *N. benthamiana* resulted in the accumulation of two minor new peaks (1c and 1d in Supplementary Figure S2B), which were tentatively identified as β-geranylfarnesene and geranylfarnesol, based on their mass spectra (Supplementary Figure S2C). These are likely to be hydrolyzed products of GFPP generated by endogenous *N. benthamiana* enzymes. However, when AtTPS17, -18, -19, -25, -30, Cr237, or Bo250 were coexpressed with a GFPPS (regardless of whether it was the corresponding paired one) in *N. benthamiana*, new peaks with a characteristic sesterterpene mass fragment at m/z = 340 (calculated for C25H40=340) were detected (Figure 3.2A and Supplementary Figure S2E and S3A-H), indicating production of sesterterpenes by these TPSs. The electron-ionization (EI)-MS spectra of compounds 2 and 3 agreed with those of (+)-arathanatriene and (−)-retigeranin B characterized recently [22]. We did not detect any sesterterpenes when AtTPS22 or AtTPS29 were coexpressed with a GFPPS. Apart from AtTPS17, which produced multiple products in very low yields, AtTPS18, -19, -25, -30, Cr237, and Bo250 all produced dominant peaks corresponding to compounds **2**-**8** (Figure 3.2A and 3.2C), suggesting that GFPP was limiting in *N. benthamiana* leaves in the absence of a coexpressed GFPPS. Compounds **2**-**8** were then isolated for structural elucidation.

*3.2.3. AtTPS18, -19, -25, -30, Cr237, and Bo250 Synthesize the Fungal-Type Sesterterpenes (+)-Arathanatriene (**2**), (−)-Retigeranin B (**3**), (−)-Ent-Quiannulatene (**4**), (−)-Varicülartriene A (**5**), (+)-Astellatene (**6**), (−)-Caprutriene (**7**), (+)-Boleracene (**8**), and (−)-Aleurodiscalene A (**9**).*

To isolate compounds **2**-**9** for structural elucidation, we scaled up the transient expression of these different GFPPS–TPS pairs in *N. benthamiana* [34]. Following extraction, repeated chromatography, and recrystallization, we obtained quantities of pure compounds **2** (102.0 mg), **3** (13.0 mg), **4** (3.7 mg), **5** (3.5 mg), **6** (4.2 mg), **7** (21.3 mg), **8** (17.3 mg), and **9** (2.1 mg) for collecting full 1D and 2D NMR datasets (Supplementary Dataset S1). In addition to NMR data, we also used X-ray diffraction analysis to solve these complex structures unambiguously. We synthesized the epoxides **10**, **12**, and **13** of compounds **2**, **4**, and **6** and the enone **11** of compound **3** (Figure 3.3) to obtain single crystals for X-ray diffraction analysis. We also succeeded in obtaining single crystals of **7**, **8**, and **9** for crystallography. The crystal structures of **7**-**13** enabled us to establish the relative structures of (+)-arathanatriene (**2**), (−)-retigeranin B (**3**), (−)-ent-quiannulatene (**4**), and (+)-astellatene (**6**), (−)-caprutriene (**7**), (+)-boleracene (**8**), and (−)-aleurodiscalene A (**9**) unambiguously, as depicted in Figure 3.2B.

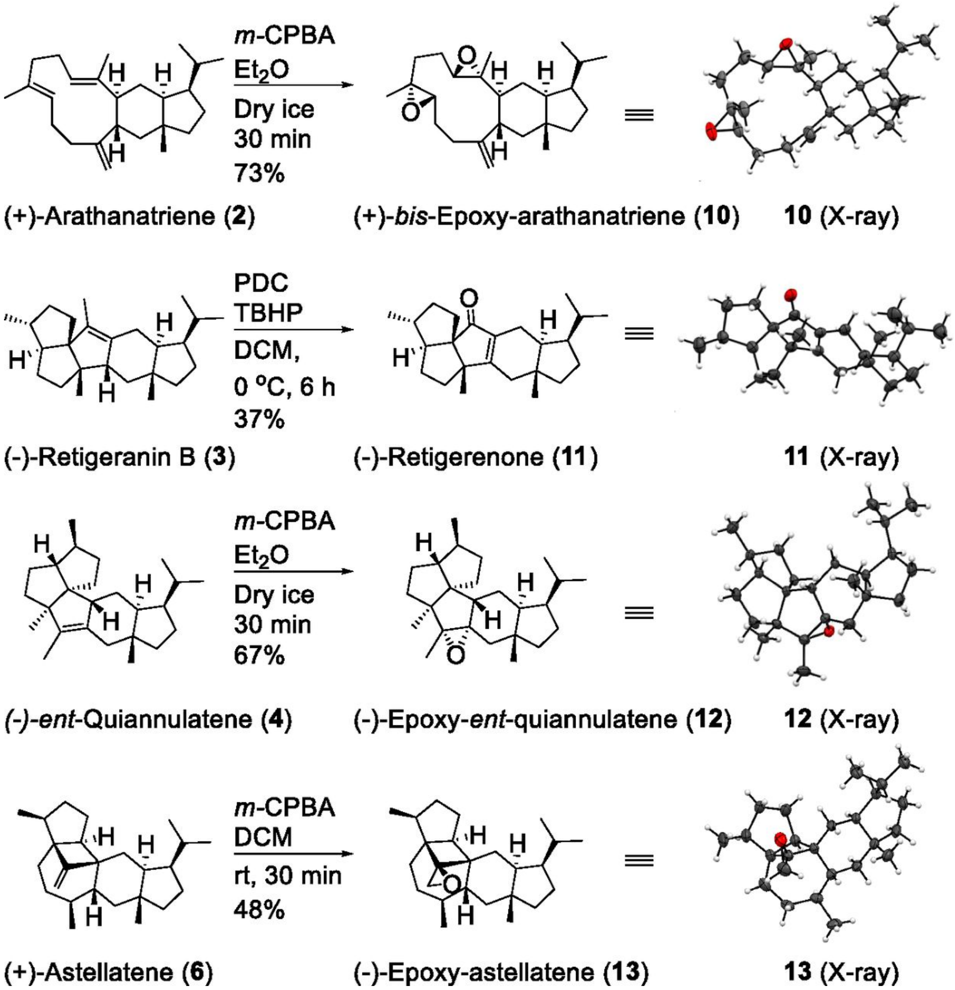


Figure 3.3. Synthesis of sesterterpene derivatives (**10**-**13**) for X-ray diffraction analysis. Epoxides **10**, **12**, and **13** were synthesized by epoxidation with meta-chloroperoxybenzoic acid (m-CPBA). Enone **11** was synthesized by allylic oxidation with pyridinium dichromate (PDC) and tert-butyl hydroperoxide (TBHP) in dichloromethane (DCM). Crystal structures (**10**-**13**) are presented with displacement ellipsoids shown at 50% probability.

The spectroscopic data of compounds **2** and **3** agreed with those reported in the literature [22]. The most likely absolute structures of the compounds (**2**-**4** and **6**-**9**) were established using Bayesian statistics on the Bijovet differences [35] of crystal structures **7**-**13**. (−)-ent-Quiannulatene (**4**) is an enantiomer of the fungal sesterterpene (+)-quiannulatene, sharing an identical relative structure, and its absolute structure was also confirmed by comparing its specific optical rotation ($[\alpha]^{20}_{D}$=−43.9 in benzene) with that of quiannulatene ($[\alpha]^{28}_{D}$=+41.7 in benzene-d_6) reported in the literature [15].

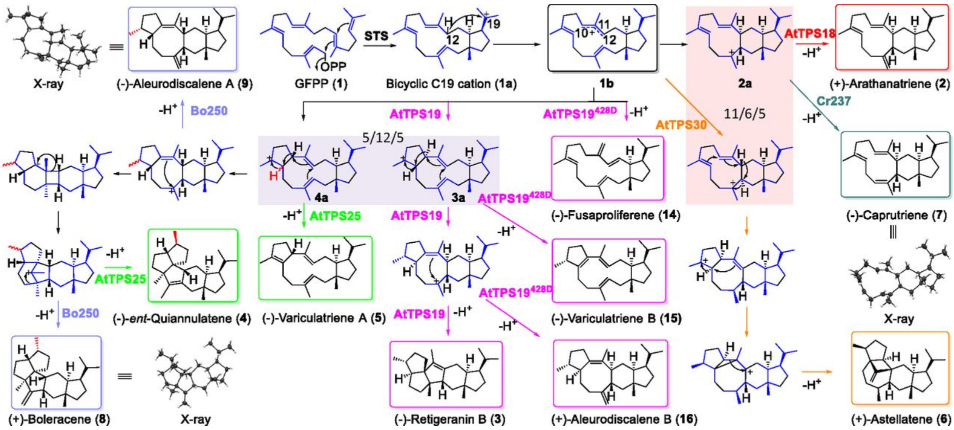


Figure 3.4. Proposed cyclization paths toward the formation of fungal-type sesterterpenes **2**-**9** and **14**-**16** by plant STSs. The universal sesterterpene precursor GFPP is cyclized to form the unified bicyclic C12 cation 1b (black box) following protonation in the active sites of plant STSs and mutated AtTPS19 (AtTPS19428D). Cation 1b diverges to 5/12/5 and 11/6/5 tricyclic carbocations en route to the formation of (+)-arathanatriene (**2**), (−)-retigeranin B (**3**), (−)-ent-quiannulatene (**4**), (−)-variculatriene A (**5**), (+)-astellatene (**6**), (−)-caprutriene (**7**), (+)-boleracene (**8**), (−)-aleurodiscalene A (**9**), (−)-fusaproliferene (**14**), (−)-variculatriene B (**15**), and (+)-aleurodiscalene B (**16**). Compounds isolated and characterized are highlighted in colored boxes. Different colors indicate different cyclization paths. Crystal structures **7**-**9** are presented with displacement ellipsoids shown at 50% probability.

The absolute structure of compound **5**, which is a minor product of AtTPS25, was elucidated by 1D and 2D NMR analysis (Supplementary Table S4), as well as the fact that it is a product derived from one carbocation in the cyclization pathway toward (−)-ent-quiannulatene (**4**) (Figure 3.4). This compound is named (−)-variculatriene A based on the 5/12/5 scaffold of variculanol, which was isolated from *E. variecolor* in 1991 [36]. (−)-Retigeranin B (**3**) and (+)-astellatene (**6**) are direct precursors of fungal sesterterpenoids retigeranic acid B (isolated from lichens in 1972 [37–39]) and astellatol (isolated from *E. variecolor* in 1989 [40]), respectively (Supplementary Figure S4). (−)-Aleurodiscalene A (**9**) shares the same scaffold with the antimicrobial sesterterpene glycoside aleurodiscal isolated from *A. mirabilis* in 1989 [41] (Supplementary Figure S4).

With these pure compounds in hand, we also detected compounds **2**, **3**, **4**, and **6**, (Figure 3.5), which are products of *A. thaliana* TPSs in the plant of origin. We found that these four sesterterpenes are present across different *A. thaliana* tissues at various levels. (+)-Arathanatriene (**2**) and (−)-retigeranin B (**3**) were more abundant in the siliques and seeds, whereas (−)-ent-quiannulatene (**4**) and (+)-astellatene (**6**) were detected only in roots. The metabolite profiles are in good agreement with the expression profiles of the corresponding TPSs, suggesting that AtTPS18, -19, -25, and -30 are active and could function as STSs *in planta* (Figure 3.5 and Supplementary Figure S5 and S6).

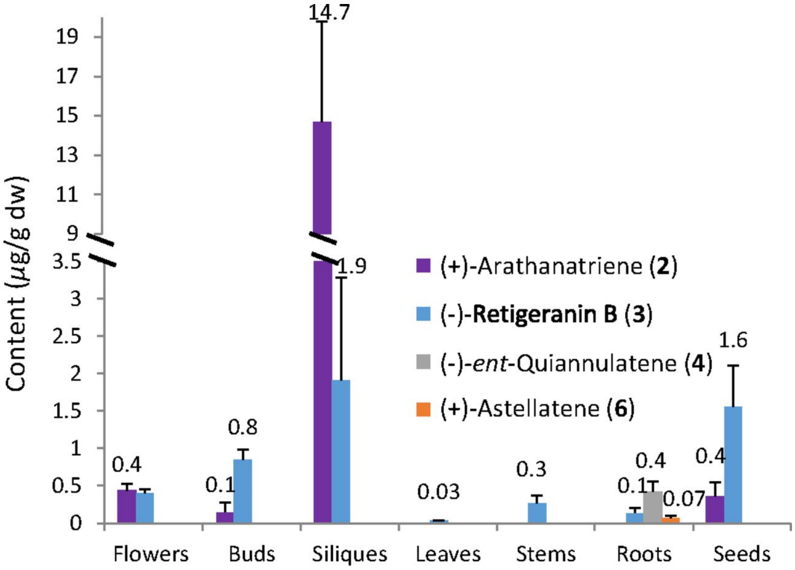


Figure 3.5. Detection of compounds **2**, **3**, **4**, and **6** from different tissues of *A. thaliana* accession Col-0. SD bars for three biological replicates are shown.

*3.2.4. Divergence of Carbocation Cyclization Pathways Gives Rise to the Structural Diversity of Sesterterpenes **2**-**9**.*

The diverse fungal-type sesterterpene products of the plant STSs identified in this work arise from the divergence of carbocation cyclization pathways following the formation of a common ancestral bicyclic C12 carbocation **1b** (Figure 3.4). Feeding experiments with ^{13}C-labeled precursors in fungi have suggested the formation of a quiannulatene scaffold via a 5/12/5 tricyclic carbocation and an intriguing cyclobutane intermediate [15,42]. With the isolation of tricyclic (5/12/5) and tetracyclic (5/8/6/5) carbocation-derived (−)-variculatriene A (**5**) and (−)-aleurodiscalene A (**9**) as minor products of AtTPS25 and Bo250, which produce mainly the quiannulatene scaffold sesterterpenes **4** and **8**, respectively (Figure 3.2C), it is very likely that cyclization of GFPP toward (−)-ent-quiannulatene (**4**) and (+)-boleracene (**8**) in plant STSs adopts the same path as (+)-quiannulatene in fungal bifunctional STS (EvQS), as shown in Figure 3.4. Although the astellatene and retigeranin B scaffolds have recently been proposed to also be derived from 5/12/5 tricyclic carbocations (**12**), previous feeding experiments with ^{13}C-labeled precursor (**43**) and our finding that the 11/6/5 scaffold (**2**) is the major product of AtTPS18 and is present as a very minor product of AtTPS19, -25, and -30 (Figure 3.2C) suggests that the bicyclic C12 carbocation may be diverged into tricyclic 5/12/5 and 11/6/5 cations en route to the formation of pentacyclic sesterterpenes, such as (−)-retigeranin B (**3**) and (+)-astellatene (**6**). Such divergence is driven by inherent energetics of the carbocation reactions involved, the conformations allowed in STS active sites and interactions of carbocations with key amino acid residues in the active sites of STSs.

*3.2.5. Mutation of Tyrosine428 to Aspartic Acid in AtTPS19 Reveals Intermediates in the Cyclization Pathway Toward the Formation of (−)-Retigeranin B (**3**).*

Phylogenetic analysis of the seven STSs identified in this work showed that AtTPS18 is a relatively recently evolved homolog of AtTPS19 (Supplementary Figure S1A). By comparing the protein sequences of these STSs (Figure 3.6A and Supplementary Figure S7), together with homology modeling of the encoding enzymes (Figure 3.6B), we identified a conserved amino acid site present as an aromatic tyrosine (Tyr; Y) in the active sites of AtTPS17, -19, -25, -30, and Bo250. However, in AtTPS18, the corresponding residue is a negatively charged aspartic acid (Asp; D) in close proximity (approximately 2.75 Å) to the C12 of the GFPP substrate (Figure 3.6B), suggesting that this site might be important for driving the formation of the 11/6/5 tricyclic scaffold of (+)-arathanatriene (**2**). We then carried out site-directed mutagenesis of the STSs from A. thaliana. We mutagenized the Tyr at the corresponding amino acid sites of AtTPS17, -19, -25, and -30 to Asp. This approach resulted in inactivation of AtTPS17, -25, and -30, as well as significant functional changes in AtTPS19, with (−)-retigeranin B dropping to 30% from 42% in wild-type AtTPS19 and the appearance of peak **16** and the increase of peaks **14** and **15** (Figure 3.2C and Figure 3.6C), whereas altering Asp to Tyr in AtTPS18 did not lead to metabolite profile change. The Tyr residue in AtTPS17, -25, and -30 therefore appears to be important for enzyme function and changing it to Asp could have possibly caused incorrect folding.

Although AtTPS19 is a closer homolog to AtTPS18, sharing 87% protein sequence identity (Figure 3.5A and Supplementary Figure S8), the mutant form of AtTPS19, AtTPS19428Asp, was still active after changing Tyr428 to Asp428. We then scaled-up co-infiltration of *A. tumefaciens* strains carrying pEAQ-HT/AtTPS19428Asp and pEAQ-HT/AtGFPPS1 in *N. benthamiana*, isolated the compounds corresponding to peaks **14**, **15**, and **16**, and established their structures as (−)-fusaproliferene (**14**), (−)-variculatriene (**15**), and (+)-aleurodiscalene (**16**), as depicted in Figure 3.4 by NMR (Supplementary Table S13-S15). Intriguingly, changing Tyr428 to Asp428 in AtTPS19 resulted in premature termination of the intermediate carbocations at different stages of the cyclization pathway toward (−)-retigeranin B (**3**) and led to the accumulation of the corresponding bi-, tri-, and tetracyclic products (−)-fusaproliferene (**14**), (−)-variculatriene B (**15**), (+)-arathanatriene (**2**), and (+)-aleurodiscalene B (**16**). We hypothesize that Tyr428 of AtTPS19 might play a role in stabilizing the carbocation intermediates in transition states via cation–π interactions (**44**), such that cyclization toward pentacyclic (−)-retigeranin B (**3**) can be completed.

However, it is also possible that the enzyme activity of the AtTPS19428Asp mutant is compromised such that not all of the intermediates could be cyclized to the final pentacyclic product. The fact that the composition of (+)-arathanatriene (**2**) increased from 5.5% of wild-type AtTPS19 products to 12% for the mutagenized AtTPS19428Asp suggests that a common ancestral bicyclic C12 cation may have been diverted in the enzyme active sites into 5/12/5 and 11/6/5 tricyclic carbocations en route to the formation of pentacyclic sesterterpenes, such as (−)-retigeranin B (**3**), and that such divergence could occur within a single enzyme. Although changing Asp431 to Tyr431 in AtTPS18 did not alter its function and we have not yet identified the amino acid sites that drive the exclusive formation of the 11/6/5 tricyclic scaffold of **2**, it is reasonable to hypothesize that AtTPS18 has evolved to be more specific and to have a more compact active site that would impose steric force on the bicyclic C12 cation (**1b**) to allow for the tight formation of the C2–C12 bond and subsequently the 11/6/5 tricyclic scaffold (**2** and **7**).

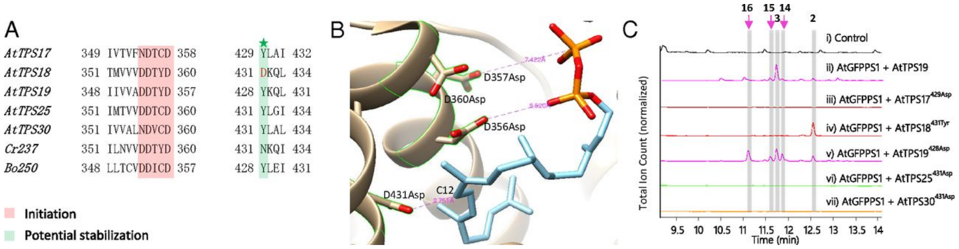


Figure 3.6. Protein sequence analysis, homology modeling and site-directed mutagenesis of A. thaliana STSs. (A) Sequence analysis of AtTPS17, -18, -19, -25, -30, Cr237, and Bo250 identified a conservation site implicated in structural diversification (protein sequences were aligned using MUSCLE). (B) Homology modeling of AtTPS18 with substrate GFPP (carbon backbone in cyan and phosphate in orange and red) docked in the active site showing the motif DDXXD and 431Asp suggested that 431Asp might be involved in making the 11/6/5 tricyclic scaffold of arathanatriene. The homology model of AtTPS18 was generated based on the crystal structure of 5-epi-aristolochene synthase on the Phyre2 server (www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index). (C) Comparative GC-MS total ion chromatograms of extracts of *N. benthamiana* leaves transiently coexpressing AtGFPPS1 with mutagenized AtTPS17429Asp, AtTPS18431Tyr, AtTPS19428Asp, AtTPS25431Asp, and AtTPS30431Asp.

3.3. Discussion

The characterized bifunctional STSs from fungi and the plant STSs identified in this work are very distantly related (around 10% amino acid sequence identity) (Supplementary Figure S8). However, both types of STSs are able to cyclize the universal precursor GFPP into similar or in some cases even identical scaffolds (e.g., (+)-quiannulatene by EvQS and (−)-ent-quiannlatene by AtTPS25) via virtually equivalent cyclization paths. This striking example of convergent evolution of terpenoids in plants and fungi has parallels with the well-known example of the diterpenoid gibberellin phytohormones [43]. By delving into the carbocation cyclization mechanisms of the plant STSs that we have identified, we have demonstrated that cyclization of the acyclic GFPP precursor to pentacyclic sesterterpenes (e.g., (−)-retigeranin B and (−)-ent-quiannulatene) can diverge into 5/12/5 and 11/6/5 tricyclic scaffolds in one single enzyme after the formation of a common ancestral bicyclic C12 cation, supported experimentally by the isolation and characterization of intermediate cation derived compounds **2**, **5**, **9**, and **14**-**16**.

Quantum chemical calculations (see **subchapter 3.3** and Supplementary Dataset S3 for details) were also performed on two reasonable cyclization pathways for the formation of (+)-astellatene (**6**) (Figure 3.7). In path 1, an 11/6/5 ring system is formed first (carbocation **D**), whereas path 2 involves initial formation of a 5/12/5 ring system (carbocation **K**). On the basis of the computed energetics for these two rearrangements (Figure 3.8), both pathways are energetically viable, but path 1 is inherently preferred, given that the highest energy transition-state structure along this pathway is predicted to be 6-kcal/mol lower than that along path 2. In both pathways, instances of the merging of chemical events (e.g., hydride shifts, carbocation-alkene cyclizations) into concerted processes are observed (e.g., **E** → **G**, **G** → **I**, **L** → **N**, **N** → **I**) [44].

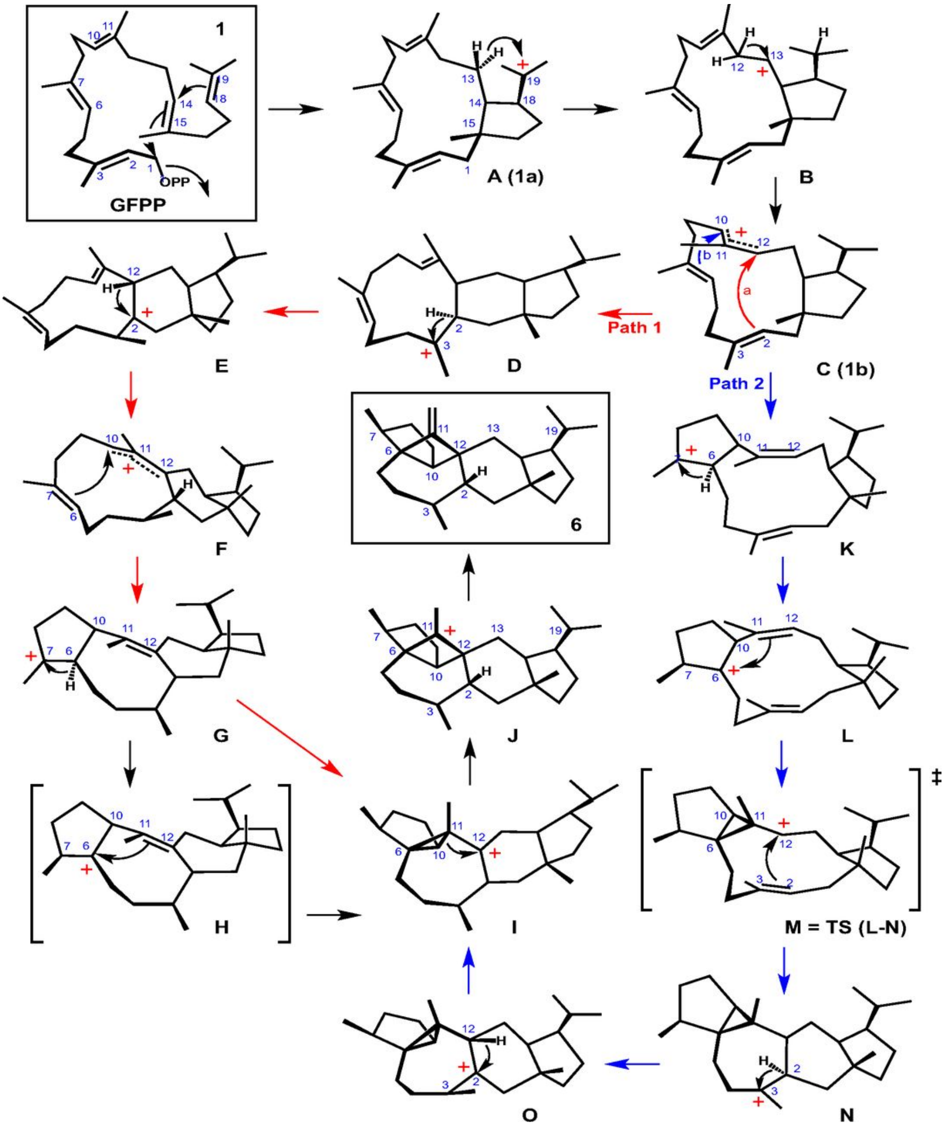


Figure 3.7. (**A**-**O**) Calculated structures for two possible cyclization paths en route to (+)-astellatene (**6**).

The results of these calculations indicate that several putative intermediates are unlikely to have lifetimes long enough for deprotonation to occur, allowing us to predict which carbocations are most likely to lead directly to sesterterpenes.

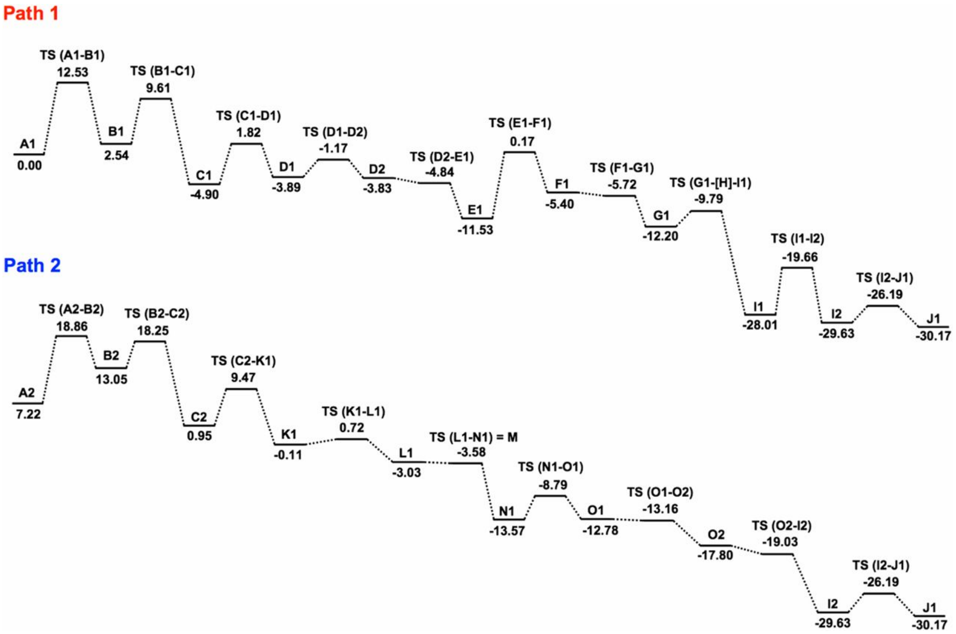


Figure 3.8. Computed energetics for two possible paths to (+)-astellatene (**6**). Energies computed with mPW1PW91/6–31+G(d,p)//B3LYP/6–31+G(d,p) are shown for the two pathways depicted in Figure 3.7. Path 1 is energetically preferred. Numbers following carbocation letters indicate different conformations. See Supplementary Dataset S3 for computed geometries.

By identifying a key amino acid residue Tyr428 in AtTPS19 important for the complete the cyclization toward (−)-retigeranin B (**3**), we have unveiled the bi-/tri-/tetracyclic structures **14**, **15**, **2**, and **16** that are derived from likely carbocations in the cyclization path toward (−)-retigeranin B (**3**), demonstrating the capacity of plant STSs to generate these diverse scaffolds and showcasing the promising potential for directing the functional plasticity of STSs for creating chemical diversity via rational protein engineering (**47**). Solving the crystal structures of these enzymes will be key in guiding such approaches in the future. The fact that (−)-fusaproliferene (**14**) and (+)-aleurodiscalene B (**16**) are direct precursors of the known bioactive sesterterpenoids fusaproliferin (toxic) and aleurodiscal (antimicrobial) (Supplementary Figure S4) illustrates the importance of harnessing the capacity to generate scaffolds of this type.

We present in this work the convergent evolution of STSs in fungi and plants as inferred from both metabolic and genomic data. Both fungi and plants appear to have independently evolved similar strategies for biosynthesizing structurally related sesterterpenes. Although, unlike the situation in fungi, plant STSs and GFPPSs are encoded by separate genes rather than bifunctional ones, it is intriguing that the genes encoding the STSs that we have characterized are colocalized in plant genomes with PT genes. These gene pairs are not obviously coexpressed (Supplementary Figure S5) and the GFPPSs are interchangeable (Figure 3.2A), suggesting that these gene pairs have not undergone intimate coevolution. The genomic colocalization of PT and TPS genes appears to be particularly prevalent in the Brassicaceae based on our analysis of currently available plant genomes. From the topology of the paired TPSs within the phylogenetic tree (Figure 3.1B) and our functional analysis of selected TPSs from *A. thaliana*, *C. rubella*, and *B. oleracea*, it seems that the functions of closely related TPSs from these three species are quite conserved and that close homologs make sesterterpene analogs (Figure 3.9). It is likely that these PT and TPS gene pairs may have originated from a common ancestral PT-TPS gene pair by duplication and functional divergence [23,45–47]. Although the biological roles of these sesterterpenes in plants are as yet unknown and inevitably many downstream products in the biosynthetic pathways remain to be elucidated, we expect that the biosynthetic and mechanistic insights provided by this work will facilitate the understanding and discovery of the biology and chemistry of this important class of compounds in plants.

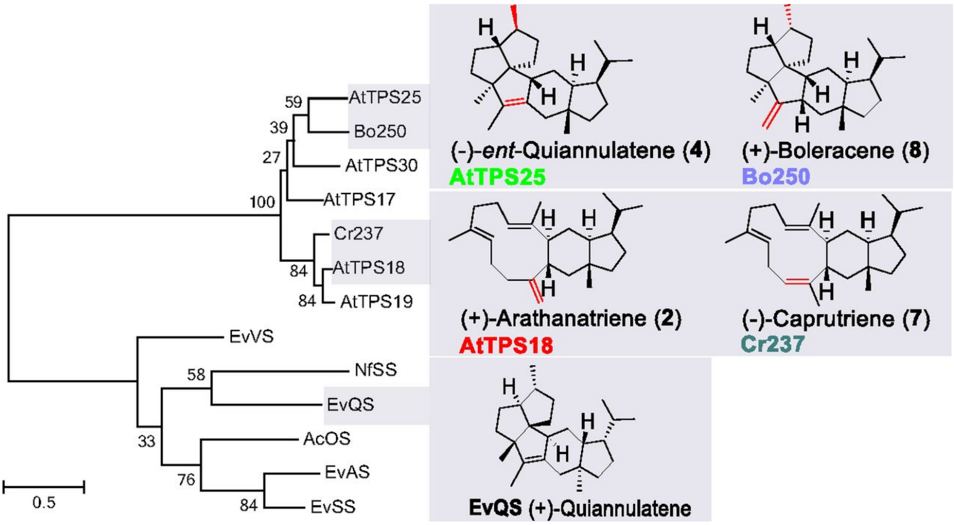


Figure 3.9. Phylogenetic tree of the characterized plant and fungal STSs showing the phylogeny of these distantly related proteins. The chemical structures of analogs synthesized by plant STS homologs and the distantly related fungal STS EvQS are shown to aid in evolutionary interpretation. The tree was constructed using MEGA5.0 with 1,000 bootstrap resampling.

3.4. Materials and Methods

3.4.1. Bioinformatics Algorithms and Construction of Phylogenetic Trees.

A collection of 55 high-quality plant genomes (Supplementary Dataset S2) was used as input for analysis with a customized version of plantiSMASH [24]. Terpene_synth_C (PF03936) and Polyprenyl_synt (PF00348) pHMMs were fitted into a custom gene cluster detection rule (Supplementary Dataset S2).

Subsequently, the PF03936 pHMM was used to collect all TPS-encoding genes from the 55 plant genomes, which resulted in the identification of 2,846 genes. A total of 724 fungal genes, including 17 bifunctional TPS genes [15] were also collected with the same method but queried against the reference proteome database with taxonomic filtering set to "Fungi." Finally, 278 bacterial genes (same method, with filtering set to "Bacteria") were included to act as an outgroup.

All sequences were aligned using the hmmalign program from the HMMer suite [48] with default parameters. Predicted repetitive subsequences were discarded. The remaining sequences were manually trimmed for tree reconstruction, which was performed using FastTree (v2.1) [49]. The phylogenetic tree was visualized using iTOL software [50]. Reliability tests were done using the FastTree-provided Shimodaira–Hasegawa test with 1,000x resamples without branch optimization.

3.4.2. Cloning and Transient Expression.

AtTPS19, AtTPS25, AtTPS30, AtGFPPS1, and AtGFPPS3 were amplified from a root cDNA library of *A. thaliana* accession Col-0. Coding sequences for AtTPS17, AtTPS18, AtTPS22, AtTPS29, and AtGFPPS4 were retrieved from The Arabidopsis Information Resource and those of Cr237 (CARUB_v10016237mg) and Bo250 (LOC106343250) from the National Center for Biotechnology Information and synthesized by Integrated DNA Technologies to include the 5' and 3' attB sites for Gateway cloning. These sequences were cloned into the pEAQ-HT expression vector for transient expression in *N. benthamiana*, as detailed in Supplementary Appendix.

3.4.3. Metabolite Extraction and GC-MS Analysis.

N. benthamiana leaves expressing genes of interest were harvested five days post-infiltration, lyophilized, extracted with EtOAc, and analyzed directly by GC-MS, as described in Supplementary Appendix. Qualitative and quantitative metabolite analysis were performed on an Agilent GC (7890B)-MSD (5977A) equipped with a robotic multipurpose autosampler (MPS) and a Zebra-5HT INFERO capillary column (35-m, 0.25-mm, 0.1-μm film thickness; phenomenex).

3.4.4. Isolation of Sesterterpenes Following Large-Scale Vacuum-Infiltration of N. benthamiana Leaves.

A. tumefaciens LBA4404 harboring expression constructs for the relevant GFPPSs and STSs were infiltrated into *N. benthamiana* leaves as described previously [34], with slight modifications using a custom-built vacuum infiltration system (Supplementary Appendix). Leaves of *N. benthamiana* expressing desired constructs were harvested five days post-infiltration, lyophilized, and extracted with n-hexane. Pure compounds **2**-**9** and **14**-**16** were isolated from the extract as described in Supplementary Appendix.

*3.4.5. Chemical Synthesis of Sesterterpene Derivatives (**10**-**13**) for X-Ray Diffraction Analysis.*

Compounds **10**, **12**, and **13** were synthesized by epoxidation of sesterterpenes **2**, **4**, and **5** with meta-chloroperoxybenzoic acid. Enone **11** was synthesized by allylic oxidation of compound **3** with pyridinium dichromate and tert-butyl hydroperoxide in dichloromethane (Supplementary Appendix).

3.4.6. Structural Characterization of Sesterterpenes by NMR and Other Spectroscopic Techniques.

Standard 1D and 2D NMR spectra, including 1H, 13C, DEPT135, COSY, HSQC, HMBC, and NOESY were acquired on a Bruker 400-MHz Topspin NMR spectrometer. All signals were acquired at 298K. Samples were dissolved in CDCl3 or benzene-d6 for data acquisition and calibrated by referencing to either residual solvent 1H and 13C signal or TMS. Detailed structural assignments and tabulated NMR data for compounds **2**-**16** are presented in Supplementary Appendix and Supplementary Table S1-S15. Optical rotations of compounds **2**-**16** were measured in benzene using a PerkinElmer Polarimeter (Model 341) with a 100-mm path cell (1 mL) at 20°C and converted to specific optical rotation using equation $[\alpha]^{20}{}_{D}$ = 100x (measured optical rotation)/concentration (g/100 mL).

3.4.7. Single Crystal X-Ray Diffraction Analysis.

Single-crystal X-ray analysis was carried out for **7**, **8**, **9**, **10**, **11**, **12**, and **13** on a Bruker D8-QUEST instrument equipped with a PHOTON-100 area detector and Incoatec IμS Cu microsource (wavelength=1.5418 Å, beam diameter at the crystal approximately 100 μm). Crystals were mounted on an X-ray transparent loop (Mitegen) using an inert oil and cooled to 180(2)K using an open-flow N2 cryostat. Data was collected and processed using the APEX3 software package (Bruker). Structures were solved and refined using SHELXT and SHELXL (Bruker). In the absence of any significant anomalous scatterers, reliable indications of the absolute structure could not be obtained by conventional refinement of the Flack parameter. However, Bayesian statistical methods [35] indicate the following probabilities that the absolute structure is correct, under the assumption that the crystal is enantiopure (p2(true)): 7 0.886, 8 0.980, 9 0.682, 10 1.000, 11 1.000, 12 1.000, and 13 0.745. See Supplementary Dataset S4 and S5 for crystal structures and crystallographic raw data of compounds **7**-**13**.

3.4.8. Sequence Comparisons and Homology Modeling.

Amino acid sequences of AtTPS17, -18, -19, -25, -30, Cr237, and Bo250 were aligned using Uniprot (www.uniprot.org/). Homology models of AtTPS18 were generated on modeling server Phyre2 (www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) using the crystal structure of 5-epi-aristolochene synthase as template. Docking of STSs substrate GFPP was performed using Autodock4.0 and AutoDockTool developed by The Scripps Research Institute [51].

3.4.9. Mutagenesis of AtTPSs17, -18, -19, -25, and -30.

Site-directed mutagenesis was performed by PCR amplification using the entry vector pDONR207 harboring the wild-type genes as templates and the mutated complementary sequences as primers, as listed in the Supplementary Appendix, Supplementary Table S17. Mutagenized genes were Gateway-cloned into the pEAQ-HT expression vector, transformed into LBA4404 and coexpressed with LBA4404/GFPPS1 in our transient tobacco expression system, as described in Cloning and Transient Expression, above.

3.4.10. Quantum Chemical Calculations.

All quantum calculations were performed with the GAUSSIAN09 software suite. Geometries were optimized using the B3LYP density functional theory method and the 6–31+G(d,p) basis set. All stationary points were characterized as minima or transition-state structures using frequency calculations. All reported energies include zero-point energy corrections (unscaled) from these frequency calculations. mPW1PW91 single-point energies were calculated for all structures [52]. These methods are well established for examining carbocation rearrangement reactions [44]. Structures of intermediate carbocations, transition state structures, and corresponding coordinates are presented in Supplementary Dataset S3.

Acknowledgments

We thank A. Leveau, H. Nutzmann, R. Thimmappa, and other members of the A.O. laboratory for helpful discussion and suggestions on experiments.

Funding

Research in D.J.T.’s laboratory is supported by US National Science Foundation CHE-1565933 and computing support from CHE030089 (XSEDE). A.O.’s laboratory is supported by the United Kingdom Biotechnological and Biological Sciences Research Council Institute Strategic Programme Grant “Understanding and Exploiting Plant and Microbial Metabolism” (BB/J004561/1) and the John Innes Foundation. This work is supported by the NIH Genome to Natural Products Network Award U101GM110699 (to A.O. and A.C.H.). A.C.H. is supported by European Commission Marie Skłodowska-Curie Individual Fellowship H2020-MSCA-IF-EF-ST-702478-TRIGEM. M.H.M. is supported by VENI Grant 863.15.002 from The Netherlands Organization for Scientific Research. S.A.K. is supported by a grant from the Graduate School for Experimental Plant Sciences (to M.H.M.).

Supplementary Material

All supporting information is available online at https://bit.ly/3s3rRBx.

References

1. Wang L, Yang B, Lin X-P, Zhou X-F, Liu Y. Sesterterpenoids. Nat Prod Rep. 2013;30: 455–473.

2. Evidente A, Kornienko A, Lefranc F, Cimmino A, Dasari R, Evidente M, et al. Sesterterpenoids with Anticancer Activity. Curr Med Chem. 2015;22: 3502–3522.
3. Zhang C, Liu Y. Targeting cancer with sesterterpenoids: the new potential antitumor drugs. J Nat Med. 2015;69: 255–266.
4. Luo S-H, Luo Q, Niu X-M, Xie M-J, Zhao X, Schneider B, et al. Glandular trichomes of Leucosceptrum canum harbor defensive sesterterpenoids. Angew Chem Int Ed Engl. 2010;49: 4471–4475.
5. Li C-H, Jing S-X, Luo S-H, Shi W, Hua J, Liu Y, et al. Peltate glandular trichomes of Colquhounia coccinea var. mollis harbor a new class of defensive sesterterpenoids. Org Lett. 2013;15: 1694–1697.
6. Luo S-H, Weng L-H, Xie M-J, Li X-N, Hua J, Zhao X, et al. Defensive sesterterpenoids with unusual antipodal cyclopentenones from the leaves of Leucosceptrum canum. Org Lett. 2011;13: 1864–1867.
7. Luo S-H, Hua J, Niu X-M, Liu Y, Li C-H, Zhou Y-Y, et al. Defense sesterterpenoid lactones from Leucosceptrum canum. Phytochemistry. 2013;86: 29–35.
8. Ahkami A, Johnson SR, Srividya N, Lange BM. Multiple levels of regulation determine monoterpenoid essential oil compositional variation in the mint family. Mol Plant. 2015;8: 188–191.
9. Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, et al. Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties. BMC Plant Biol. 2015;15: 212.
10. Vining KJ, Johnson SR, Ahkami A, Lange I, Parrish AN, Trapp SC, et al. Draft Genome Sequence of Mentha longifolia and Development of Resources for Mint Cultivar Improvement. Mol Plant. 2017;10: 323–339.
11. Chiba R, Minami A, Gomi K, Oikawa H. Identification of ophiobolin F synthase by a genome mining approach: a sesterterpene synthase from Aspergillus clavatus. Org Lett. 2013;15: 594–597.
12. Ye Y, Minami A, Mandi A, Liu C, Taniguchi T, Kuzuyama T, et al. Genome Mining for Sesterterpenes Using Bifunctional Terpene Synthases Reveals a Unified Intermediate of Di/Sesterterpenes. J Am Chem Soc. 2015;137: 11846–11853.
13. Matsuda Y, Mitsuhashi T, Lee S, Hoshino M, Mori T, Okada M, et al. Astellifadiene: Structure determination by NMR spectroscopy and crystalline sponge method, and elucidation of its biosynthesis. Angew Chem Weinheim Bergstr Ger. 2016;128: 5879–5882.
14. Qin B, Matsuda Y, Mori T, Okada M, Quan Z, Mitsuhashi T, et al. An unusual chimeric diterpene synthase from Emericella variecolor and its functional conversion into a sesterterpene synthase by domain swapping. Angew Chem Weinheim Bergstr Ger. 2016;128: 1690–1693.
15. Okada M, Matsuda Y, Mitsuhashi T, Hoshino S, Mori T, Nakagawa K, et al. Genome-Based Discovery of an Unprecedented Cyclization Mode in Fungal Sesterterpenoid Biosynthesis. J Am Chem Soc. 2016;138: 10011–10018.
16. Matsuda Y, Mitsuhashi T, Quan Z, Abe I. Molecular Basis for Stellatic Acid Biosynthesis: A Genome Mining Approach for Discovery of Sesterterpene Synthases. Org Lett. 2015;17: 4644–4647.
17. Chen M, Chou WKW, Toyomasu T, Cane DE, Christianson DW. Structure and Function of Fusicoccadiene Synthase, a Hexameric Bifunctional Diterpene Synthase. ACS Chem Biol. 2016;11: 889–899.
18. Toyomasu T, Tsukahara M, Kaneko A, Niida R, Mitsuhashi W, Dairi T, et al. Fusicoccins are biosynthesized by an unusual chimera diterpene synthase in fungi. Proc Natl Acad Sci U S A. 2007;104: 3084–3088.
19. Toyomasu T, Kaneko A, Tokiwano T, Kanno Y, Kanno Y, Niida R, et al. Biosynthetic gene-based secondary metabolite screening: a new diterpene, methyl phomopsenonate, from the fungus Phomopsis amygdali. J Org Chem. 2009;74: 1541–1548.
20. Brodelius M, Lundgren A, Mercke P, Brodelius PE. Fusion of farnesyldiphosphate synthase and epi-aristolochene synthase, a sesquiterpene cyclase involved in capsidiol biosynthesis in Nicotiana tabacum. Eur J Biochem. 2002;269: 3570–3577.
21. Fischer MJC, Rustenhloz C, Leh-Louis V, Perrière G. Molecular and functional evolution of the fungal diterpene synthase genes. BMC Microbiol. 2015;15: 221.
22. Shao J, Chen Q-W, Lv H-J, He J, Liu Z-F, Lu Y-N, et al. (+)-Thalianatriene and (-)-Retigeranin B Catalyzed by Sesterterpene Synthases from Arabidopsis thaliana. Org Lett. 2017;19: 1816–1819.
23. Boutanaev AM, Moses T, Zi J, Nelson DR, Mugford ST, Peters RJ, et al. Investigation of

terpene diversification across multiple sequenced plant genomes. Proc Natl Acad Sci U S A. 2015;112: E81–8.

24. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 2017;45: W55–W63.

25. Wang C, Chen Q, Fan D, Li J, Wang G, Zhang P. Structural Analyses of Short-Chain Prenyltransferases Identify an Evolutionarily Conserved GFPPS Clade in Brassicaceae Plants. Mol Plant. 2016;9: 195–204.

26. Keeling CI, Weisshaar S, Ralph SG, Jancsik S, Hamberger B, Dullat HK, et al. Transcriptome mining, functional characterization, and phylogeny of a large terpene synthase gene family in spruce (Picea spp.). BMC Plant Biol. 2011;11: 43.

27. Schilmiller AL, Schauvinhold I, Larson M, Xu R, Charbonneau AL, Schmidt A, et al. Monoterpenes in the glandular trichomes of tomato are synthesized from a neryl diphosphate precursor rather than geranyl diphosphate. Proc Natl Acad Sci U S A. 2009;106: 10865–10870.

28. Falara V, Akhtar TA, Nguyen TTH, Spyropoulou EA, Bleeker PM, Schauvinhold I, et al. The tomato terpene synthase gene family. Plant Physiol. 2011;157: 770–789.

29. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–85.

30. Nagel R, Bernholz C, Vranová E, Košuth J, Bergau N, Ludwig S, et al. Arabidopsis thaliana isoprenyl diphosphate synthases produce the C25 intermediate geranylfarnesyl diphosphate. Plant J. 2015;84: 847–859.

31. Wang Q, Jia M, Huh J-H, Muchlinski A, Peters RJ, Tholl D. Identification of a Dolabellane Type Diterpene Synthase and other Root-Expressed Diterpene Synthases in Arabidopsis. Front Plant Sci. 2016;7: 1761.

32. Sainsbury F, Thuenemann EC, Lomonossoff GP. pEAQ: versatile expression vectors for easy and quick transient expression of heterologous proteins in plants. Plant Biotechnol J. 2009;7: 682–693.

33. Sainsbury F, Lomonossoff GP. Transient expressions of synthetic biology in plants. Curr Opin Plant Biol. 2014;19: 1–7.

34. Geisler K, Hughes RK, Sainsbury F, Lomonossoff GP, Rejzek M, Fairhurst S, et al. Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants. Proc Natl Acad Sci U S A. 2013;110: E3360–7.

35. Hooft RWW, Straver LH, Spek AL. Determination of absolute structure using Bayesian statistics on Bijvoet differences. J Appl Crystallogr. 2008;41: 96–103.

36. Singh SB, Reamer RA, Zink D, Schmatz D, Dombrowski A, Goetz MA. Variculanol: structure and absolute stereochemistry of a novel 5/12/5 tricyclic sesterterpenoid from Aspergillus variecolor. J Org Chem. 1991;56: 5618–5622.

37. Kaneda M, Takahashi R, Iitaka Y, Shibata S. Retigeranic acid, a novel sesterterpene isolated from the lichens of lobaria retigera group. Tetrahedron Lett. 1972;13: 4609–4611.

38. Corey EJ, Desai MC, Engler TA. Total synthesis of (.+-.)-retigeranic acid. J Am Chem Soc. 1985;107: 4339–4341.

39. Kaneda M, Iitaka Y, Shibata S. X-ray studies of C25 terpenoids. IV. The crystal structure of retigeranic acid p-bromoanilide. Acta Crystallogr B. 1974;30: 358–364.

40. Sadler IH, Simpson TJ. The determination by n.m.r. methods of the structure and stereochemistry of astellatol, a new and unusual sesterterpene. Journal of the Chemical Society, Chemical Communications. 1989. p. 1602. doi:10.1039/c39890001602

41. Lauer U, Anke T, Sheldrick WS, Scherer A, Steglich W. Antibiotics from basidiomycetes. XXXI. Aleurodiscal: an antifungal sesterterpenoid from Aleurodiscus mirabilis (Berk. & Curt.) Höhn. J Antibiot . 1989;42: 875–882.

42. Hong YJ, Tantillo DJ. How cyclobutanes are assembled in nature--insights from quantum chemistry. Chem Soc Rev. 2014;43: 5042–5050.

43. Hedden P, Sponsel V. A Century of Gibberellin Research. J Plant Growth Regul. 2015;34: 740–760.

44. Tantillo DJ. Biosynthesis via carbocations: theoretical studies on terpene formation. Nat Prod Rep. 2011;28: 1035–1053.

45. Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci U S A. 2011;108: 16116–16121.

46. Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, et al. Evolution of a complex locus for terpene biosynthesis in solanum. Plant Cell. 2013;25: 2022–2036.

47. Trapp SC, Croteau RB. Genomic organization of plant terpene synthases and molecular evolutionary implications. Genetics. 2001;158: 811–832.
48. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.
49. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5: e9490.
50. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242–5.
51. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30: 2785–2791.
52. Matsuda SPT, Wilson WK, Xiong Q. Mechanistic insights into triterpene synthesis from quantum mechanical calculations. Detection of systematic errors in B3LYP cyclization energies. Org Biomol Chem. 2006;4: 530–543.

Chapter 4

MIBiG 2.0: An Improved Reference Database of Experimentally Validated BGCs

Following the transformative developments in genome sequencing and analytical chemistry technologies at the turn of the 21st century, a large number of experiments characterized the molecular mechanisms (i.e., BGCs) responsible for the production of many natural products. This trove of knowledge was initially scattered across literature, making it difficult to use for systematic analysis of uncharacterized BGCs. In 2015, the Minimum Information about a BGC (MIBiG) repository was created, providing manually curated information of 1,170 BGCs and their chemical products. In this chapter, we present MIBiG 2.0, a large update that includes 851 new entries and a major reworking of the database architecture. Altogether, the new version presents a significant quality improvement of the database, equipping it for the increasingly large numbers of sequenced genomes and metagenomes.

Published as:

Kautsar, S. A.*, Blin, K.*, Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J., ... & Medema, M. H. (2020). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research, 48(D1), D454-D458*.

* shared first authorship

4.1. Introduction

Plants, microbes and fungi produce a large variety of specialized metabolites that are often uniquely found in one or a few species. From the dawn of civilization, humans have tapped into this treasure trove for medicinal, economic or recreational purposes. Within the last decade, genome-based discovery of specialized metabolites has become a widely adopted practice within both the scientific community and commercial settings. The magnitude of these efforts is continuously growing because of the ongoing increase in availability of genome and metagenome assemblies in public databases. These sequences can be mined for the presence of Biosynthetic Gene Clusters (BGCs): multi-enzyme loci that encode the biosynthetic pathways for one or more specific compounds.

Thousands of candidate BGCs have thus been identified using computational tools such as antiSMASH [1] and ClusterFinder [2]. Databases like IMG-ABC [3] and antiSMASH-DB [4] store many thousands of such computationally predicted BGCs, potentially coding for a very diverse range of natural product classes. To unravel the function and novelty of current and future candidate BGCs, knowledge on previously characterized BGCs is essential. This calls for a standardized deposition and extraction of BGCs associated with molecules of known chemical structure, as this relevant knowledge is usually buried inside the text of scientific articles.

A first step to this end was taken in 2013, when ClusterMine360 [5] appeared, the first database of BGCs with known products, containing data on around 300 gene clusters. In 2015, the MIBiG (Minimum Information about a Biosynthetic Gene Cluster) Data Standard and Repository was established, containing 1,170 BGC entries that were manually curated through a community effort, the results of which could be accessed via a fairly simple web application [6]. Now, the MIBiG repository has become a central reference database for BGCs of known function, and provides the basis for comparative analyses in antiSMASH via its KnownClusterBlast module. It has enabled many computational analyses of BGC function and novelty central to both small and large-scale studies of microbes and microbial communities. For example, Crits-Cristoph et al. [7] recently used MIBiG to assess and highlight the exceptional novelty of BGCs across 376 metagenome-assembled genomes of uncultivated soil bacteria from understudied phyla, by showing that most of these BGCs lacked any homology to gene clusters from MIBiG. Similarly, Bahram et al. [8] used homology searches against MIBiG to identify fungal BGCs associated with antibacterial activity across 7,560 metagenomic samples, based on a set of MIBiG gene clusters whose products could be annotated with this activity; thus, they were able to show that the abundance of such “antibacterial” BGCs correlated with the presence of antimicrobial resistance genes across soils. Yet another usage is illustrated by the ClusterCAD tool [9], which sources BGC data from MIBiG as a starting point for the computer-aided design of new biochemical pathways.

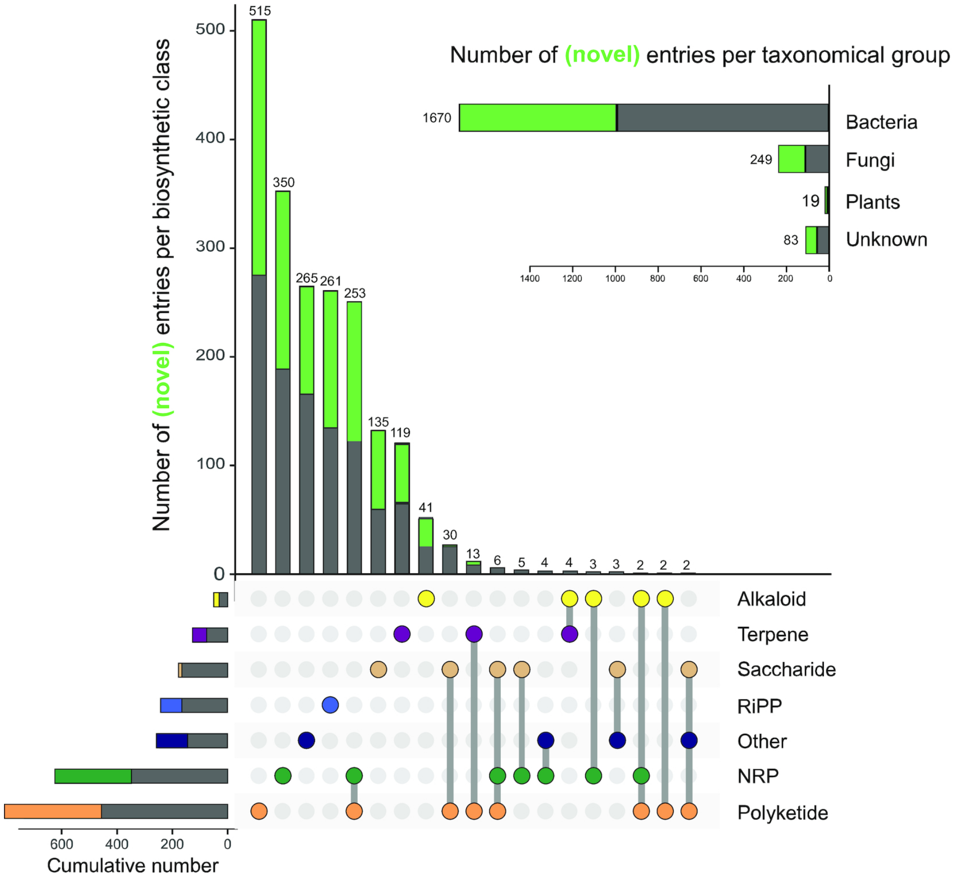


Figure 4.1. Distribution of taxonomic kingdoms and biosynthetic classes for all BGCs present in and added to MIBiG 2.0. Statistics are taken after the restructuring effort, and include retired entries. New entries are depicted in light green. Only (hybrid) classes comprising more than one BGC entry are listed in the figure. The intersection diagram is generated using the UpSetR tool [10].

Here, we provide an updated MIBiG version 2.0, which has been significantly expanded through the addition of 851 new entries over the past five years (Figure 4.1). Moreover, we performed extensive re-annotation of the entire database, increasing the overall data quality by improving the data schema, by adding hundreds of literature references and chemical structures and by providing cross-links to recently emerged databases of chemical structures and analytical data. Finally, we added useful functionalities to the online repository to make it more user-friendly, by enabling fast filtering based on compound names, taxonomic identifiers or biosynthetic classes, and facilitating the building of Boolean queries.

4.2. Methods and Implementation

4.2.1. Manual curation of entries

Since its inception in 2015, MIBiG has provided an online submission form for adding new entries. To submit a new entry, a user starts by requesting a MIBiG accession number. This is done through submitting the product name(s) and the sequence information of the BGC, preferably in the form of a set of coordinates corresponding to the BGC's position within an NCBI Genbank accession. After the request is approved by MIBiG staff, the workflow subsequently provides an extended entry form where users can input more detailed information. This crowdsourcing, open-for-all approach has garnered 140 new entries since 2015, with contributions coming from various experts all over the world.

Because not all newly characterized BGCs are submitted to the database, we actively complemented this crowdsourcing approach by periodically organizing in-house "Annotathons", where multiple scientists sat together for an entire day to work on MIBiG curation (Supplementary Table S1). This has yielded 702 new entries, and annotation quality improvements for over 600 BGCs.

More recently, we have introduced an additional MIBiG curation process into the classroom environment with the help of a comprehensive and very specific set of guidelines for the students [11,12]. By giving one task to multiple students to work on independently, and later on having an expert (the teacher) to combine and validate the results, we have generated an additional 10 high quality BGC entries, for actinomycin, carbapanem, daptomycin, ebelactone, lipstatin, nocardicin A, obaflourin, oxazolomycin, salinosporamide and tabtoxin. Scaling up this process in the future may allow the annotations of many more important entries, which have remained incomplete, because, e.g. the scientists who have worked on the pathway are no longer active in the field.

4.2.2. Data quality improvements

The MIBiG specification needs to capture the architectural and enzymatic variety present in currently described BGCs, and needs to stay flexible enough to also accommodate future discovery of even more diverse clusters and metabolites. In the initial MIBiG release in 2015, we relied only on the cluster submission form to aid annotators in creating valid entries. Now, we also adopted the JSON schema description and validation technology (https://json-schema.org) that was recently made available, which enables us to embed validation and dependency rules into the schema. This can then be processed programmatically via libraries implemented in almost all popular programming languages.

After implementing the JSON schema updates, we performed a thorough data quality assessment of the entire repository, fixing empty or mistyped information in the data, removing duplicate entries, adding and correcting structural information, adding new entries, and retiring entries we deemed of insufficient quality, e.g., when the sequence assembly does not cover the full DNA sequences of the cluster region, effectively removing spatial context from the BGC data (Supplementary Table S2).

Finally, additional cross-links have been established with the Natural Products Atlas (https://www.npatlas.org/) and the GNPS spectral library [13]. This enables users to acquire information about specialized metabolites with structures similar to those found in MIBiG, and to identify mass spectra linked to a specific molecule of interest. These additions further complement the already existing links with PubChem [14] and other compound databases. Connections were made according to compound names and structures matching between the annotated BGCs and the chemical databases.

4.2.3. The new database architecture

Previously stored in a collection of static HTML pages, the MIBiG data has now been migrated into a relational database. This setup allows users to query the metadata, using either a simple search form or an interactive query builder that assists in building more complex queries. A REST-like web API (https://github.com/mibig-secmet/mibig-api/) handles access to an underlying PostgreSQL (https://www.postgresql.org/) database. A single-page web application written in AngularJS (https://angularjs.org/) runs the user interface allowing users to browse a repository overview, view statistics about the clusters in the database, or run metadata queries. The individual BGC pages are generated using a customised antiSMASH5 module that sideloads a MIBiG annotation file (in JSON format). Annotations generated by antiSMASH are also produced alongside the manually curated MIBiG information.

4.3. Results and Discussion

4.3.1. BGC diversity

The MIBiG repository version 2.0 encompasses 2,021 manually curated BGCs with known functions, which is a 73% increase from the original 1,170. Categorically, there are seven structure-based classes: “Alkaloid”, “Nonribosomal Peptide (NRP)”, “Polyketide”, “Ribosomally synthesised and Post-translationally modified Peptide (RiPP)”, “Saccharide”, “Terpene”, and “Other”. These classes may overlap, as in the case of Polyketide-NRP hybrids such as Rapamycin (BGC0001040) and Bleomycin (BGC0000963). The “Other” category includes cyclitols like cetoniacytone A (BGC0000283), indolocarbazoles like rebeccamycin (BGC0000821) and phosphonates like fosfomycin (BGC0000938). MIBiG is currently mostly populated with entries of the Polyketide (825 BGCs) and NRP (627 BGCs) classes. Hybrids of these classes are also prominently featured. Proportionally, the new entries also contain a lot of Polyketides and NRPs, together comprising more than half (59%) of the batch. Taxonomically, BGCs in MIBiG have mostly bacterial or fungal origins (in particular, the genus *Streptomyces* is the most prominent with 568 BGCs, followed by *Aspergillus* at 79 and *Pseudomonas* at 61), with only nineteen coming from plants.

4.3.2. Annotation completeness

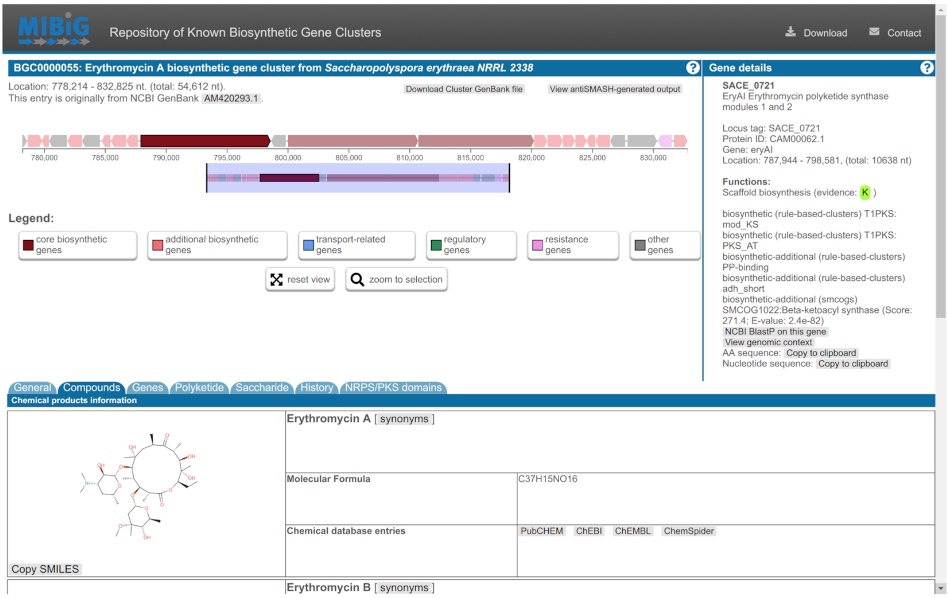
BGCs in MIBiG start with a “minimal” annotation, meaning that it consists only of locus information (Genbank accession and coordinates of the cluster), a compound name, and at least one reference publication. Detailed information such as compound structures (stored as a SMILES string), class-specific attributes (e.g. Polyketide synthase (PKS) modules), are usually, but not always, present. Prior to the schema restructuring, there were 2,021 BGCs, of which 770 did not have any chemical structure of their product(s) associated with them, and 500 had missing or incomplete properties. With the results of all manual re-curation efforts compiled into the dataset, we have incorporated new structure information for 220 BGCs, solved most of the issues with incomplete properties, and retired some BGCs of low annotation quality (Supplementary Table S2). (These retired entries are still available for download.) An overview of the updates is shown in Table 4.1 below.

Table 4.1. Annotation completeness of BGCs in MIBiG 2.0 before and after the restructuring effort.

	Before	**After**
Entries without structure information	**770**	**550**
Entries with incomplete properties	**500**	**18**
• No reference publication	148	11
• Values unknown to the schema	235	0
• Others	158	7
Retired entries		**105**
• Duplicate BGC		11
• Poor sequence quality		70
• Poor annotation quality		24

4.3.3. A new online repository

The overall design of the old repository has been thoroughly refreshed. Rows in the “Repository” page can now be filtered and sorted based on annotation metadata, such as species names or biosynthetic classes. The BGC page itself takes advantage of the modernized, well-organized look of antiSMASH5 [1]. Annotation data are now organized into their own category tabs, e.g. “General”, “Compounds”, “History”, “Polyketide”, “NRPS” and so on (Figure 4.2). Some new functionalities were also introduced to the main page. “Statistics” displays a real-time overview of the database, such as compound class distribution, taxonomy, and annotation completeness. “Search” provides users the ability to build complex queries based on MIBiG metadata, for example “find all complete RiPP BGCs from the genus *Streptomyces*”.


MIBiG
Repository of Known Biosynthetic Gene Clusters
Download
Contact

BGC0000055: Erythromycin A biosynthetic gene cluster from *Saccharopolyspora erythraea NRRL 2338*

Location: 778,214 - 832,825 nt. (total: 54,612 nt).
This entry is originally from NCBI GenBank AM420293.1 .

Download Cluster GenBank file
View antiSMASH-generated output
780,000
785,000
790,000
795,000
800,000
805,000
810,000
815,000
820,000
825,000
830,000
Legend:
🟥 core biosynthetic genes
🟥 additional biosynthetic genes
🟦 transport-related genes
🟩 regulatory genes
🟪 resistance genes
⬛ other genes
reset view
zoom to selection

Gene details

SACE_0721
EryAI Erythromycin polyketide synthase modules 1 and 2

Locus tag: SACE_0721
Protein ID: CAM00062.1
Gene: eryAI
Location: 787,944 - 798,581, (total: 10638 nt)

Functions:
Scaffold biosynthesis (evidence: K)

biosynthetic (rule-based-clusters) T1PKS: mod_KS
biosynthetic (rule-based-clusters) T1PKS: PKS_AT
biosynthetic-additional (rule-based-clusters) PP-binding
biosynthetic-additional (rule-based-clusters) adh_short
biosynthetic-additional (smcogs) SMCOG1022:Beta-ketoacyl synthase (Score: 271.4; E-value: 2.4e-82)
NCBI BlastP on this gene
View genomic context
AA sequence: Copy to clipboard
Nucleotide sequence: Copy to clipboard

Chemical products information

Erythromycin A [synonyms]

Molecular Formula	C37H15NO16
Chemical database entries	PubCHEM ChEBI ChEMBL ChemSpider

Erythromycin B [synonyms]

Figure 4.2. The new per-BGC overview page. The locus overview (top-left) section allows panning, zooming, or highlighting specific genes, for which the information would be displayed in the gene details (top-right) section. In the lower section, the “Compounds” tab is currently selected, showing all compound-related information of the BGC, such as chemical structure, molecular formula, or linked databases. Other data is linked to other specific tabs.

4.4. Data Availability

The MIBiG Repository is available at https://mibig.secondarymetabolites.org/. There is no access restriction for academic or commercial use of the repository and its data. The source code components, JSON-formatted data standard, and SQL schema for the MIBiG Repository are available on GitHub (https://github.com/mibig-secmet) under an OSI-approved Open Source license.

Acknowledgements

We thank all individual submitters who have contributed by submitting entries to MIBiG since 2015.

Funding

Netherlands Organization for Scientific Research (NWO) VENI Grant [863.15.002 to M.H.M.]; National Science Foundation (NSF) CAREER Award [CHE-1652424 to L.K.C.]; UK Biotechnology and Biological Sciences Research Council (BBSRC) Nottingham-Rothamsted Doctoral Training Partnership (grant number BB/M008770/1) and the Lawes Agricultural Trust (to G.L.); NSF GRF

[00039202]/NWO GROW fellowship (to S.L.R.); Graduate School for Experimental Plant Sciences (EPS); Novo Nordisk Foundation [NNF10CC1016517, NNF16OC0021746 to T.W.]; Netherlands eScience Center (NLeSC) Accelerating Scientific Discoveries Grant [ASDI.2017.030 to J.J.J.v.d.H.]; NSERC Discovery (to R.G.L.); National Institutes of Health [U41-AT008718 to R.G.L.]. Funding for open access charge: Wageningen University.

Conflicts of Interests Statement

M.H.M. is a member of the Scientific Advisory Board of Hexagon Bio and co-founder of Design Pharmaceuticals.

Supplementary Material

All supporting information is available online at https://bit.ly/3s0FH7O.

References

1. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87.
2. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
3. Hadjithomas M, Chen I-MA, Chu K, Huang J, Ratner A, Palaniappan K, et al. IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. Nucleic Acids Res. 2017;45: D560–D565.
4. Blin K, Pascal Andreu V, de Los Santos ELC, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2019;47: D625–D630.
5. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 2013;41: D402–7.
6. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11: 625–631.
7. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature. 2018;558: 440–444.
8. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560: 233–237.
9. Eng CH, Backman TWH, Bailey CB, Magnan C, García Martín H, Katz L, et al. ClusterCAD: a computational platform for type I modular polyketide synthase design. Nucleic Acids Res. 2018;46: D509–D515.
10. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33: 2938–2940.
11. Epstein SC, Charkoudian LK, Medema MH. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. Stand Genomic Sci. 2018;13: 16.
12. Li YF, Tsai KJS, Harvey CJB, Li JJ, Ary BE, Berlew EE, et al. Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. Fungal Genet Biol. 2016;89: 18–28.
13. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34: 828–837.
14. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47: D1102–D1109.

Chapter 5

Lessons for the Development and Curation of Microbial Natural Product Databases

Natural product (NP) discovery has always been an interdisciplinary effort involving the integration of multiple fields of study. Complementary to BGC databases like MIBiG and AntiSMASH-DB, the last decade (2010-2020) was also marked by the emergence and increasing usage of metabolomics and chemical structures/properties databases of NPs. In this chapter, we provide a brief review of relevant databases for microbial NP discovery from this era. Being involved in developing and maintaining some of the mentioned databases ourselves, we also discuss the challenges and opportunities of biological database curation and integration as we embrace the new age of "big omics".

Published as:

van Santen, J. A., Kautsar, S. A., Medema, M. H., & Linington, R. G. (2020).
Microbial natural product databases: moving forward in the multi-omics era.
Natural Product Reports.

5.1. Introduction

Information management remains a central limitation in natural products science. Access to comprehensive, structured, freely available repositories containing key data allows researchers to determine what has been found to date, understand how previous discoveries relate to new findings, and identify how new results fit into the broader picture of natural products diversity and biosynthesis. In this review we will present the current landscape of databases for microbial natural products science, and discuss how to address the challenges and limitations facing the field as we move towards the implementation of large, comprehensive, integrated data architectures for natural products data and metadata.

5.1.1. A brief history of natural products data management

Although we now take for granted the rapid, facile access to electronic data on natural products, this is a relatively recent development (Figure 5.1). Prior to the 1990s, there were essentially no online scientific databases containing information on natural products. Instead, most data management strategies involved the laborious transcription of key data from print journals to index cards for use in individual laboratories. It cannot be overstated how much this lack of access to comprehensive, ordered datasets has negatively impacted our field. Asking senior researchers about historical data management approaches yields a litany of stories describing painful days spent chasing information through the print literature. These stories include such historical curiosities as punch cards, 8″ floppy discs, photocopier accounts, suitcase sized laptops, and early mainframe computers.

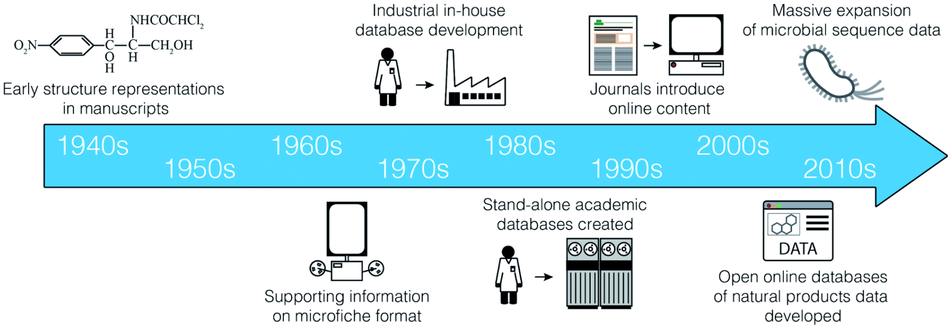

NHCOCHCl2
O2N
CH2OH
Early structure representations in manuscripts
Industrial in-house database development
Journals introduce online content
Massive expansion of microbial sequence data
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Supporting information on microfiche format
Stand-alone academic databases created
DATA
Open online databases of natural products data developed

Figure 5.1. Timeline of data distribution methods for natural products.

During this period, numerous print reference books were maintained that collated key data from the scientific literature. Of particular note were the Chemical Abstracts series, the Ring Systems Handbook [1], Fungal Metabolites volumes 1 and 2 [2,3], the Handbook of Antibiotic Compounds (volumes 1–14) [4] and the Encyclopedia of Antibiotics [5].

Searching through such compendia was inherently slow, and instances of rediscovery were common. To reduce redundant effort by individual researchers, many organizations began to develop their own in-house data collections. A representative example of this type of resource is the system developed by the pharmaceutical company Lederle Laboratories beginning in the early 1960s, as described by Dr Guy Carter:

“Lederle Laboratories maintained its own database, dubbed the Antibiotic Properties file, of which we were very proud. The database consisted of a series of three-ring binders, arranged in alphabetical order, holding a single page of information on each antibiotic including structure (if known, and a surprising number were not), biological spectrum and any other bio data, like cytotoxicity, and chemical properties that were known, like elemental analysis, mw and most importantly a UV spectrum - frequently xeroxed from the original paper and pasted on the form. The database was maintained by the Lederle library staff, and was compiled by Lederle retirees, who were hired to review the literature for new compounds - quite a system!”

In the 1980s, several important electronic resources began to emerge. CAS and Beilstein began developing the large-scale literature databases that have become Scifinder and Reaxys. Initially these tools had very strict fee-for-search models that often limited the number of searches that researchers could perform in a given month. Gradually, this evolved to the institution subscription model we know today. In the area of natural products, two academic efforts are of particular note. Professor Hartmut Laatsch created AntiBase, a database of microbial natural products, while Professors John Blunt and Murray Munro created MarinLit, a database of articles on marine natural products. Both resources were originally available on CD-ROM by paying an annual subscription to the developers to support development costs.

Commercial publishers were also developing electronic databases. For example, CRC Press began to publish the Dictionary of Natural Products which also came with a CD-ROM containing a basic search engine. These various electronic resources developed incrementally over the following decades, and remain the reference tools of choice for many natural products research groups around the world today.

5.1.2. A new age in natural products discovery

The early 2010s were marked by the emergence of new tools that made data-centric methods accessible to the ‘average’ natural products scientist; one without a dedicated training in programming or computer science. Examples of such tools include NaPDoS [6] and eSNaPD [7] for assessing the biosynthetic diversity of microbial strains, FuSiOn [8] for the de novo prediction of compound modes of action, and iSNAP [9] for the dereplication of non-ribosomal peptides from mass spectrometry data.

One tool that had a significant impact on the adoption of new data technologies was antiSMASH [10]. First released in 2011 [11], antiSMASH provided a simple, freely accessible web interface for the identification of biosynthetic gene clusters (BGCs) from genomic sequence data. The natural products community quickly recognized the power that such analyses could bring to many aspects of their research programs, and antiSMASH became a mainstay tool for many natural product programs. Instead of requiring subject experts to scan raw sequence data by hand, antiSMASH offered users a straightforward mechanism to generate initial automated annotations, which could then be prioritized for further investigation. The accessibility and power of this new resource set the tone for natural product tool development, and generated an immediate demand for new tools that would provide the same level of functionality in other areas of natural products.

5.1.3. Data storage, dissemination and collaboration

The exponential growth in omics research and so called “Big Data” is self-evident. The world's data volume has grown from about 1.5 zettabytes (ZB, 1021) in 2009 to a projected 44 ZB by 2020 [12]. Current models suggest that the global data volume will reach 175 ZB by 2025 [13].

In this age of internet and digital information, there is an increasing need to store and share not only raw experimental data but also analysis results, processed data, research protocols, knowledge materials and scientific findings. Gone are the days where scientists spent days scouring the library for answers and waiting for the next delivery of printed journals to keep track of what was happening in their field. Nowadays, people can disseminate, query, and even collaborate on research data with others around the globe in real time and in a large-scale fashion (e.g., crowdsource efforts). In this modern approach to science, databases play an essential role in ensuring that the data being generated are stored, processed, presented and shared in the most effective means.

To enable effective data storage and collaboration, databases should adhere to FAIR (findable, accessible, interoperable and reusable) principles in their implementation [14]. This is particularly important for the inclusion of researchers from developing nations, where subscription cost for commercial tools can present an insurmountable barrier to access. Many companies provide mechanisms for reduced cost or free journal access to researchers from selected countries, but for low-to-middle income countries that are not included, data access remains a significant barrier to scientific development. This barrier can be significantly reduced by creating high-quality FAIR-compliant resources.

5.2. Databases for microbial natural products research

5.2.1. Chemical structure and properties databases

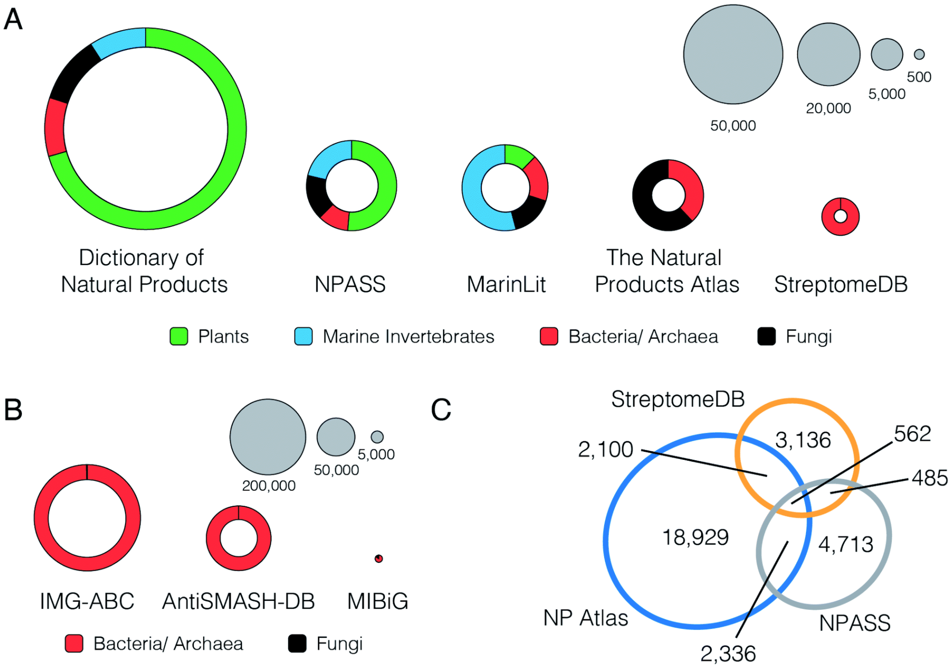


Figure 5.2. (A) Distribution of compound source types in selected natural products databases. (B) Distribution of biosynthetic gene cluster source types in selected biosynthetic gene cluster databases. (C) Overlap of microbial natural product InChIKey structure representations between open access databases. Microbial database overlap was calculated using the unique sets of the InChIKey connectivity hashes from each database. This decreases the compound count in each database because sets of configurational isomers are reduced to single flat structures: NP Atlas 25,523 to 23,927, NPASS 8,729 to 8,096, and StreptomeDB 7,125 to 6,283. The Proportional Venn Diagram was created using eulerAPE v3 [15].

The current landscape for natural product structural databases is highly fragmented. A recent comprehensive review by Sorokina and Steinbeck [16] lists an astonishing 122 resources for natural product structures developed since the year 2000. This list includes both commercial and non-commercial repositories, covering a wide range of source organisms and geographic locations. However, despite the breadth of natural product databases available, the options for microbial natural product scientists are surprisingly limited. From the 122 resources, only fifty permit access to the full set of structures. Of these, eleven contain entries for bacterial natural products, and only three (NPASS, StreptomeDB and the Natural Products Atlas) permit filtering by taxonomic origin to extract only the microbially-derived compounds. These three resources therefore currently represent the best freely available sources of information on microbial natural products structures Figure 5.2.

NPASS [17] (**http://bidd.group/NPASS/**) is a recently developed natural products database (2018) designed to provide both source organisms and biological activities for natural products. It contains partial coverage of the chemical space of natural products from several taxonomic sources, including plants, invertebrates and microorganisms. In total it contains 35,032 compounds, of which approximately 9,000 are microbial in origin.

StreptomeDB [18] (**http://www.pharmbioinf.uni-freiburg.de/streptomedb3/**) is a targeted database that focuses exclusively on the bacterial genus Streptomyces. Recently updated in 2020, it contains 7,125 compounds with source organism information, as well as some bioactivity and spectral data.

The Natural Products Atlas [19] (**http://https://www.npatlas.org/**) is a new resource (2019) designed to provide comprehensive coverage of all microbially-derived natural product structures. It currently contains 25,523 compounds (v2019_12) and is under active development. It features bi-directional links to two other natural products resources; the MIBiG database of biosynthetic gene clusters and the GNPS database of natural products mass spectra.

In addition to open source databases, a number of high-quality commercial platforms are available. Of these, the Dictionary of Natural Products (DNP), MarinLit and AntiBase are the most well established, although AntiBase was last updated in 2014. All three of these databases are large (>30,000 compounds) and contain rich metadata. They have broad coverage of the published literature and are generally very accurate. However, they have high annual subscription costs and do not permit bulk export of structural data or other information to external applications. This limits their utility to individual searches and precludes their integration with other natural products-based data resources.

DNP (**http://dnp.chemnetbase.com/**) contains over 290,000 entries (accessed Feb. 2020) and includes natural products from all major source organism groups, as well as physicochemical and biological data. The database is continually updated through an extensive process of manual curation by subject experts, ensuring high data quality standards. However, spot checks on the dataset based on compound names suggest that coverage is not universal, even for some well-known compound classes (e.g., abyssomicins).

MarinLit (**http://pubs.rsc.org/marinlit/**) is a literature database of marine natural products, including structures, taxonomy, and reports on total synthesis for 35,015 compounds (accessed Feb. 2020). It includes compounds from invertebrates and algae, as well as 8,082 compounds from marine-derived microorganisms. Impressively, this database is updated almost daily, making it the most contemporary resource in this area.

The Dictionary of Antibiotics and Related Substances [20] is a reference text of over 2,000 pages listing all known naturally occurring antibiotic substances (>10,000). It was recently updated (2013) from the original edition from the 1980s, and now includes many entries from the BMIC database, which was maintained for many years by Dr Janos Berdy and was the foundational database for the Handbook of Antibiotic Compounds. It is accompanied by a searchable CD-ROM.

There also exist numerous natural products databases from biotech and pharmaceutical companies, as discussed before. Unfortunately, many of these are difficult, if not impossible, to obtain. Most are not under active development, and are archived in only physical formats, or in legacy database structures. Despite willingness from some companies to release these data to the wider community, access can be precluded by practical challenges such as completing liability release documentation; a task of typically low priority for legal departments.

Finally, it is worth mentioning the natural products coverage of the two largest chemical literature databases; Scifinder and Reaxys. Both of these platforms include the majority of compounds from the natural products literature. However, neither is particularly well suited to natural products-based queries beyond simple structure searches. Scifinder does not include any flags identifying compounds as natural products, making it impossible to separate natural products from synthetic compounds. Reaxys does include the term "Isolated from Natural Source" but many known natural products are not annotated with this flag, meaning that searches performed using this filter are not comprehensive.

5.2.2. Biosynthetic gene cluster databases

As the rate of BGC discovery began to accelerate in the early 2000s, the biosynthesis community faced many of the same challenges that had been encountered by the natural products structure elucidation community thirty years earlier. In particular, information about BGC discovery was becoming scattered across the scientific literature, or stored in a less structured manner in genomic databases such as NCBI GenBank. As with structure-based discovery, this limited the possibilities for cross-linking between resources and prevented programmable access to exploit the knowledge within. To address this issue, several databases of BGC data have been developed.

ClusterMine360 [21] **(http://clustermine360.ca):** Made available in 2013, ClusterMine360 was one of the first platforms to venture in to the task of cataloguing the information on experimentally validated BGCs with known products. Focusing on the Nonribosomal Peptide (NRP) and Polyketide (PK) classes, it contains 300 BGCs linked to their chemical products. While initially prepared for continuous expansion via user-submitted annotations, it seems that the total number of BGCs covered by the database has not increased significantly since its initial release.

DoBISCUIT [21] **(http://https://www.nite.go.jp/en/nbrc/genome/dobiscuit.html):** Released around the same time as ClusterMine360, DoBISCUIT published an initial collection of 72 known PK BGCs. Unfortunately, the database is no longer accessible, although its main page is still active and shows a final log of 108 BGCs recorded on 27 December 2016.

MIBiG Repository [22] (**http://https://mibig.secondarymetabolites.org**)**:** In 2015, a coordinated effort of more than 150 natural product scientists resulted in the publication of the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) data standard and repository for known and experimentally characterized BGCs. Holding information on more than a thousand of characterized BGCs, MIBiG was quickly adopted by the community as a central reference database for BGC data. Notably, antiSMASH automatically compares each detected BGC to all reference gene clusters from MIBiG. Four years after the initial release, in 2019, a second iteration of both the database and schema was announced, highlighting an accumulated total of 2,021 BGC entries and a major overhaul of its online repository infrastructure. MIBiG contains only BGCs which have been experimentally verified to be responsible for the production of one or more known natural products. MIBiG entries are also subject to extensive manual curation and annotation by both the developers and the scientific community, further increasing the information content and data quality in this repository.

IMG-ABC [23] (**http://https://img.jgi.doe.gov/cgi-bin/abc/main.cgi**)**:** Taking advantage of the Joint Genome Institute (JGI)'s extensive bacterial genomic platform, IMG/M, the IMG-ABC sets out to be the most comprehensive and feature-rich database of known (indirectly sourced from MIBiG) and computationally predicted bacterial BGCs. Prior to IMG-ABC v5, the database comprised a total of more than one million BGCs predicted using both antiSMASH and the ClusterFinder [24] algorithm. The latter approach has since been dropped in favour of the more stringent but more "high-confidence" BGC class detection of antiSMASH 5. This has resulted in a drop of total BGCs provided by IMG-ABC, with 410,558 BGCs available as of 29 June 2020.

An important detail to note is that, due to the JGI's Data Usage policy (https://jgi.doe.gov/user-programs/pmo-overview/policies/), it is not advisable to do bulk-analysis and publication of IMG-ABC's data as some of the genomes may still be under embargo. In the future, we recommend that IMG/M (and IMG-ABC) should follow the footsteps of their fungal genome database counterpart, MycoCosm [25] (https://mycocosm.jgi.doe.gov) to provide a simple filtering of embargoed genomes, thus enabling a "safe" bulk-download and analysis of their data.

AntiSMASH Database [26] (**http://https://antismashdb.secondarymetabolites.org**)**:** The antiSMASH database (antiSMASH-DB) was initially released in 2016 by the same team who developed antiSMASH to act as a central repository for pre-computed antiSMASH runs. In contrast to the IMG-ABC, antiSMASH-DB aims to provide a limited, dereplicated list of putative BGCs sourced from the highest quality bacterial genomes. For sets of highly similar genomes (e.g., thousands of *Escherichia coli* genomes with only a few single nucleotide polymorphisms), representatives have been picked instead of providing results for all strains individually. One key reason to do this is to provide a seamless integration with antiSMASH via its "ClusterBlast" module, which performs a sequence comparison of each detected BGC with those in the database. Following its second release in 2018, antiSMASH-DB harbours a total of 152,106 BGCs precalculated from 24,776 bacterial genomes (of which 32,548 BGCs were derived from 6,200 complete genomes) from the NCBI RefSeq database [27]. The upcoming third release will include BGCs from high-quality fungal genomes as well.

5.2.3. Databases for metabolomics and analytical chemistry

A number of resources for the sharing and analysis of metabolomics data have arisen in the last decade. Many of these resources focus around the FAIR sharing of data to enable more productive natural products discovery, and are not limited to the scope of microbial natural products science.

The Global Natural Products Social molecular network [28] **(**http://https://gnps.ucsd.edu/**):** The GNPS system is an ecosystem for sharing and analyzing tandem mass spectrometry data. It is built on the MassIVE platform, and features an impressive suite of internally connected tools. It also provides functionality for complete data lifecycle management, from data acquisition through to publication. One of the most popular features is molecular networking, which enables the visualization relationships between spectra from MS/MS experiments. Data submitted for analysis in GNPS are organized into datasets, which can either be kept private or made public. To date there are 1,413 public datasets available online (accessed 24 February 2020). In addition, GNPS houses a number of public MS/MS spectral libraries, containing 74,130 annotated spectra.

MetaboLights [29] **(http://https://www.ebi.ac.uk/metabolights/):** MetaboLights is a database run by EMBL-EBI that was originally created in 2012, and overhauled in 2019. It is a database for metabolomics data with capabilities for storing and reporting on a large variety of data types, including NMR, GC/MS, LC/MS, as well as metabolite structures, their reference spectra, and biological roles. MetaboLights is the recommended repository for metabolomics data for a number of journals based on the FAIRsharing initiative (https://fairsharing.org/biodbcore-000168/).

5.2.4. NMR metabolomics

A recent comprehensive review by McAlpine et al. [30] established the state of NMR dereplication with respect to the field of natural products. The review demonstrates that there remains an urgent need for a comprehensive and open data exchange of NMR data for natural products. Following publication of this review, the National Center for Complementary and Integrative Health and the Office of Dietary Supplements at the NIH in the US initiated a call for proposals to develop such a resource [31]. This call resulted in the establishment in 2020 of the Natural Products Magnetic Resonance Database (NP-MRD; www.np-mrd.org) which aims to create an open access repository of experimental and calculated spectra for natural products structures.

In addition to this new initiative there are a number of current databases and tools which have addressed this problem with both experimental and predicted NMR spectra.

NAPROC-13 [32] **(http://c13.usal.es/):** NAPROC-13 is a database which contains 13C NMR spectra for over 6,000 natural product compounds. The database has a web interface allowing for rapid identification of compounds present in complex mixtures, as well as providing structural information useful for novel structure elucidation.

NMRshiftDB [33] **(https://xn--nmrshidb-vs49b.nmr.uni-koeln.de/):** NMRshiftDB contains many similar features to NAPROC-13 as well as NMR from other nuclei. However, it is not exclusive to natural products chemistry.

Biological Magnetic Resonance Data bank [34] **(http://www.bmrb.wisc.edu/):** BMRB contains a wide variety of experimental and simulated NMR data from proteins, peptides, nucleic acids, and other biomolecules. BMRB is not exclusive to microbial natural products, and also contains data from all realms of natural products and metabolomics. BMRB also maintains a library of NMR pulse sequences and computational software for biomolecular NMR.

Human Metabolome Database [35] **(http://https://hmdb.ca/):** HMDB is an open-access database which provides detailed information about metabolites found in the human body, thus including those essential to the human microbiome. Many metabolites also contain experimental 1D and 2D NMR spectra, freely available for download.

CH-NMR-NP [36] **(http://https://www.j-resonance.com/en/nmrdb/):** CH-NMR-NP is a database hosted by JEOL of NMR data compiled from a list of journals from 2000 to 2014. It contains 1H and 13C NMR data from approximately 35,500 natural products and is not exclusive to microbial natural products. CH-NMR-NP is searchable online and permits download of the NMR data in the JEOL Delta data format on a compound-by-compound basis.

5.3. Database curation and usage

5.3.1. Practical challenges for database users

Surprisingly, it remains very difficult to compare data between resources in this area. Chemical structure and compound name are the common terms connecting many of these databases. In principle it should be possible to associate data from one resource (e.g., BGC) with data from another (e.g., NMR or MS data) via the chemical structure. In practice however, there is no agreed upon standardization method for chemical structures which provides a unique, machine readable structural representation without information loss. For example, several SMILES strings are possible for a single structure, standard InChI representations do not retain information on preferred tautomers, and MOL files are large blocks of text that are unwieldy to store in most database formats. These issues mean that databases typically align poorly by structure without significant additional manual curation.

Compound names are similarly challenging. Small changes in punctuation, the inclusion and encoding of special characters, or the absence of trivial names for many compounds in the literature all contribute to poor overlap between resources. This is further complicated by the assignment of new synonyms for existing compounds and, occasionally, the erroneous assignment of the same name to multiple structures. To add further complication, some compound classes receive several different parent names, often in an attempt to increase the visibility of new discoveries. Conversely, some researchers use the same parent name for all compounds isolated from a given organism, regardless of structural relatedness. Both of these issues complicate the grouping of related structures based on trivial names.

Some resources have invested substantial effort in improving interoperability. For example, the NPAtlas and MIBiG teams have manually reviewed every entry in the MIBiG database and identified the appropriate Natural Products Atlas entry in each case. These two resources now include bi-directional links between data pages, and offer exportable tables that list links between primary keys in each platform. Similar links have been set up with the GNPS platform.

Investing similar effort to align other key resources by structure could have a significant impact on the development of new cross-discipline discovery tools. An example of an effective cross-referencing system is provided by UniChem [37], a system set up by the EMBL-EBI to connect chemical structures across multiple databases by assigning a UniChem identifier to each unique chemical structure, and linking this identifier to all the databases affiliated with the UniChem system.

5.3.2. Practical challenges for database creation and management

The current publishing model is not well suited to large-scale database creation and maintenance. Each journal has its own format and data requirements, and no journals produce standardized, machine readable files containing key primary data (Figure 5.3). Rather, these data are often provided as supplementary materials in a wide variety of formats. Deposition of data to public resources (e.g., depositing biosynthetic gene clusters with NCBI) is valuable, but accession numbers must still be extracted manually from the methods or data availability sections of the papers, slowing the rate of data curation.

For chemical structures, the situation is even more difficult. Most authors do not deposit new structures to public databases (e.g., PubChem [38] or ChEBI [39]), meaning that structures start as computerized representations (e.g., ChemDraw files) are reproduced by journals as flat images in PDFs, and must then be manually re-entered in machine readable formats. This medieval approach to information dissemination is a significant barrier to data integration efforts, and one that the community must urgently address. The American Chemical Society style guide includes a clear summary of many of the challenges surrounding machine interpretation of printed structures [40].

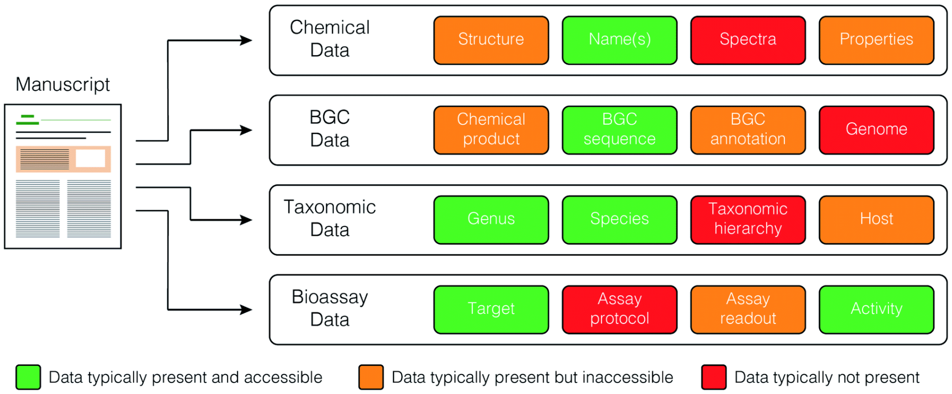


Figure 5.3. Data types and their relative accessibility from published articles in the primary scientific literature.

We propose that editors require a SMILES string in the manuscript for every new compound, as an additional component of the experimental data section. Although this is not a substitute for a separate structured data file (e.g., MDL SDF or structured JSON), it is easy to implement and would improve the digitalization of natural products research results by increasing structure availability and reducing error rates caused by manual re-entry of compound structures. Initiatives of some journals, such as Nature Chemical Biology [41], to collect such data and automatically submit all published structures to the PubChem database in a computer-readable format show that this is feasible.

For BGCs the problem is sometimes even worse as, unlike chemical structures, digital representations of BGC sequences cannot be reconstructed from images in a paper. Hence, deposition of the data to a public repository is absolutely required in order to assess a scientific paper on its merits, and to reproduce and leverage these results. The fact that many journals, even highly regarded ones such as the Journal of the American Chemical Society, regularly publish papers on BGCs without the sequence being made available anywhere is highly problematic. As is the case for proteins, we feel that it is imperative that accession numbers to GenBank entries containing the BGC are explicitly mentioned in the paper. When a BGC is characterized from a genome sequence previously published by another research group, authors should refer to the accession number of that genome and the coordinates of the BGC within it, or at least provide locus tags of the genes or accession numbers of the encoded proteins, to allow readers and database developers to find the underlying data.

Ideally, every database should relate each data point to the appropriate reference from which these data were derived. This would allow users to evaluate data more carefully than aggregated datasets where data provenance is unknown. Fortunately, the digital object identifier (DOI) system provides a unique identifier for journal articles that is easily converted to a hyperlink to each article and provides a simple method for storing article information. Frustratingly however, some publishers have not assigned DOIs to their legacy article collections. Because DOIs are not universally assigned, database systems must therefore handle both DOIs and full reference data (journal, volume, issue, pages). With the advent of e-journals that use non-standard citation formats, this has quickly become a complicated and error prone process. We therefore present a second recommendation that publishers review their legacy holdings and, where appropriate, assign DOIs to these back catalogues. This simple action would have a significant impact on the information content and interoperability of separate natural product-based data resources.

One final and often overlooked point is the cost of running and maintaining a database. Servers, IT staff, and continued software development are often forgotten in planning the longevity of data tools. Furthermore, a database may reach the end of its life due to funding or being superseded by another platform. Currently when this happens, data is often simply lost. One simple and effective solution is to store versioned releases of data dumps on a free scientific data storage solutions such as Zenodo (run by CERN and OpenAIRE, https://zenodo.org/) or GigaDB (run by the GigaScience journal, http://gigadb.org/). Otherwise, standard steps can be followed to archive a database [42]. Doing so can prevent the relegation of data to the annals of lost and forgotten databases and is best practice for FAIR data.

5.3.3. Curating microbial natural products data in 2020

Curating natural products data from the primary literature remains a predominantly manual process. It requires three main steps; identification of articles pertaining to microbial natural products discovery, extraction of structures, gene clusters and other data from each article, and organization of these data into a structured format. The most challenging of these is the identification of relevant articles. Traditionally, more than 50% of all microbial natural products discoveries were published in either the Journal of Antibiotics or the Journal of Natural Products. However, as natural products research has broadened in scope, the number of venues for reporting natural products discovery has increased. This creates challenges for data curation. Manual inspection of titles and abstracts for all published articles is now an impossibly large task. Instead, curation efforts must rely on either targeted curation of key journals, or text mining strategies using keywords to find relevant articles from public data sources such as PubMed. Both of these approaches have limitations that impact the coverage of curation efforts. Focus on a targeted list of journals can exclude reports in peripherally related areas (e.g., marine chemical ecology or microbiome studies) while text mining approaches are likely to miss core articles and are susceptible to bias depending on the algorithm(s) used for filtering. Authors can assist with this effort by ensuring that the discovery of new natural products or BGCs is prominently described in the abstract. In most cases, curators do not have bulk access to the full text versions of articles, meaning that the title and abstract are the only information available for article prioritization. A clear statement describing new compound or BGC discovery in the abstract is therefore the most effective method to ensure that new data are included in curation efforts.

5.3.4. Community contributions

A second route to data curation is through investigator-initiated submissions directly to databases. This approach has many clear advantages. It makes curation a distributed effort, rather than relying on a small number of volunteers. This in turn improves both coverage and accuracy, because the original authors are providing the key data directly. It reduces effort because these data (e.g., structures) are already in an appropriate electronic format, and reduces error rates by eliminating instances where curators incorrectly interpret data from original articles.

There are however a number of disadvantages to the community contribution model. Databases without control over data insertion can quickly become corrupted through either accidental or malicious behavior. This may often be unintentional, as it is easy to misinterpret a step in a submission form and input the wrong data. In addition, submissions from external users may not conform to the defined scope of the database. Without appropriate care, the contents of the database can quickly become heterogeneous, making it difficult or impossible to perform meaningful analyses on the entire dataset.

To address these challenges, most platforms include a secondary curation step, where external submissions are reviewed by subject experts for appropriateness and completeness. This approach is much faster than de novo literature searching, as the core data have already been submitted in an appropriate format. To make sure that submitted data are as unambiguous as possible, a clear ontology detailing the options for each data field is required, as well as clear instructions and tutorials for submission [43]. From our experience with the Natural Products Atlas and MIBiG, approximately 50% of community submissions are accepted “as is”, with a further 35% requiring format or content corrections, and 15% being rejected as outside the scope of the database.

Currently, the Natural Products Atlas, MIBiG, MetaboLights and GNPS are four of the only natural products resources that accept external submissions. This is likely in part due to low demand, because of “submission fatigue” from the ever-increasing list of requirements placed on corresponding authors. Initial submissions now require extensive information about authors and grants, and accepted articles must often be separately deposited in open repositories to satisfy funding agencies. To add to this, sequence data must typically be deposited in an open repository (e.g., NCBI) and crystal structures deposited with the Protein Data Bank or the Cambridge Structural Database. Understandably, uptake for voluntary submission of additional data is low. However, the power provided to the scientific community offered by the accumulation of data in these repositories cannot be overstated. It is up to the natural products field to lead the way in data deposition, and to develop new strategies that improve data coverage in these areas without increasing the burden on lead investigators. There are clear incentives for researchers to do so, including increased visibility and citation rates for their science, as well as the ability to see and use these data when navigating publicly available data resources.

5.4. Integration and interoperability between databases

5.4.1. Multi-omics and meta-analysis driven microbial natural products discovery

This area of natural products science is still in its infancy, but a number of important discoveries have already been enabled by the availability of comprehensive, well-structured datasets.

Global analyses performed with natural product databases. Several groups have performed recent meta-analyses on natural products science using natural product databases. Pye et al. [44] investigated the rate of novel compound discovery as a function of time and source organism type using a combination of commercial and in-house databases. They showed that, while the absolute number of novel scaffolds being discovered each year remains roughly constant, the number of derivative compounds being reported has increased dramatically over the past thirty years; currently, less than 10% of new marine and microbial compounds can be considered novel scaffolds.

Pascolutti et al. [45] used the Dictionary of Natural Products (DNP) to identify small, “fragment-like” natural products, and evaluate their physicochemical properties. They demonstrated that a subset of structures was representative of a large percentage of the total motif diversity in this sample set, and suggested that these molecules could form the foundation for future fragment-based screening libraries.

O'Hagan and Kell [46] took this premise one step further to ask which combination of 96, 384, 1,152 or 1,920 compounds would best represent the chemical space in Nature. Using a combination of the now-defunct Universal Natural Products Database [47] and DNP they were able to identify libraries that covered up to 30% of overall chemical space, and to propose a high coverage library made up entirely of commercially available natural products.

Global analyses have also been performed for BGCs, such as the study by Cimermancic et al. [24] in 2014, which surveyed the biosynthetic landscape across 1,154 sequenced bacterial and archaeal genomes, revealing widely distributed BGC classes of unknown function. Since then, the size of genomic databases has grown by orders of magnitude, however. As an example, NCBI RefSeq now holds more than 190,000 bacterial genomes compared to ±29,000 in late 2014, not to mention the rising availability of metagenome-assembled genome (MAG) sequences [48–51]. These newly available genomic data provide exciting opportunities to assess, for example, which taxonomic groups encode the richest natural product biosynthetic diversity and should therefore be targeted for discovery efforts, or how biosynthetic diversity is governed by species phylogeny versus ecology [52].

New uses for structure databases. The availability of curated structure databases has enabled the development of a number of exciting extensions to existing analytical platforms. Reher et al. [53] recently published a new version of the Small Molecule Accurate Recognition Technology platform, termed SMART 2.0. This tool uses neural networks to match HSQC NMR spectra of unknown compounds against a database of known compounds. Using this approach, the SMART 2.0 algorithm predicts the identities of compound classes for unknown molecules directly from a single NMR spectrum. In this new release, the authors included calculated HSQC spectra based on structures from several natural products databases. This dramatically increased the number of reference spectra, from 2054 in the original report to >53,000 in this new version.

In the area of mass spectrometry, a number of tools have been developed for the prediction of MS/MS fragmentation patterns [54–57]. These approaches provide a powerful new discovery modality for natural products researchers by providing an alternative to the need for validated synthetic standards for all compounds. For example, the latest version of the CFM-ID platform, CFM-ID 3.0 [56], includes a large reference library of pre-calculated spectra, as well as online and local options for calculating spectra for bespoke compound libraries. Similarly, the new release of the SIRIUS platform (SIRIUS 4) [57] incorporates the CSI:FingerID platform [58] and predicts the most likely structure for signals from mass spectrometry data, based on comparison with a database of known structures. These complement additional tools, such as MS2LDA [59] and the associated MotifDB [60], which provide annotation of metabolite substructures based on motifs found across databases of tandem mass spectra. The availability of both compound databases and tools like CFM-ID and SIRIUS therefore enables the creation of targeted annotation libraries based on specific parameters relevant to a given study (taxonomic origin, compound class, etc.).

New uses for BGC databases. One of the most obvious uses of BGC databases is in the process of dereplication: identifying whether BGCs detected in a set of (meta)genome sequences are likely to encode known biosynthetic pathways or not. For example, Crits-Christoph et al. [61] used the MIBiG database to show that >90% of BGCs they identified in metagenome-assembled genomes from uncultivated *Acidobacteria*, *Verrucomicobia*, *Gemmatimonadetes*, and *Rokubacteria* were likely to encode novel pathways. This process of dereplication can now also be automated for large genomic datasets using the BiG-SCAPE algorithm [62]. BiG-SCAPE computes sequence similarity networks from user-specified antiSMASH results together with all MIBiG database BGCs and reconstructs gene cluster families (GCFs), from which one can assess which BGCs are similar to a known BGC from MIBiG and which are not.

Another clear use case of BGC databases is to annotate functions in, for example, microbiome studies and using these annotations to infer ecological interactions. For example, Bahram et al. [63] used a set of MIBiG entries linked to products with proven antimicrobial functions to assess whether fungal antibiotic production potential is associated with the frequency of bacterial antibiotic resistance genes across topsoil metagenomes.

Furthermore, people have been using BGC databases like antiSMASH-DB to identify BGCs that contain specific combinations of genes of interest. For example, Krause et al. [64] performed pattern matching to chart the occurrence and diversity of PapR2-like regulators (SARP-type DNA-binding proteins with potential as generic activators for silent BGCs) within antiSMASH-DB, which revealed its widespread distribution across Actinobacterial genomes.

Another straightforward use of a BGC database is to chart the biosynthetic diversity of organisms within a larger taxonomic group [65]. Databases such as antiSMASH-DB make these analyses straightforward, by providing ready to use, pre-calculated BGC data and metadata (e.g., on their taxonomic origins) that can be accessed via an Application Programming Interface (API).

Finally, BGC databases also have potential to function as a “parts catalogue” for pathway engineering using synthetic biology. For example, the ClusterCAD software [66] allows users to design new modular polyketide synthase assembly lines by sourcing polyketide BGCs and polyketide synthase modules from MIBiG, and providing a graphical interface to mix and match these to build novel polyketide structures of interest. In principle, this type of computer-aided design could be expanded in various ways, e.g., by sourcing and searching any BGC from publicly available data in IMG/ABC or the antiSMASH database, or by, for example, including searches for genes encoding tailoring enzymes.

Examples of data integration between databases. There are very few examples of natural products discoveries made directly through the integration of multiple databases. This is no doubt due to the poor interoperability between most current resources, and the weak standardization of core data (structure representation, taxonomy, etc.). Some innovative research has been powered by combining chemical structure data with BGC data. For example, the GRAPE-GARLIC software pipeline [67] used retrobiosynthesis on an in-house database of chemical structures to reconstruct their monomer composition, which was then matched to monomers computationally predicted from BGC sequences found in public sequence databases. Similarly, integrating BGC data with metabolomics data has led to a range of approaches to (semi-)automatically link molecules to the genes involved in their biosynthesis based on pattern matching strategies [68–70]. There is clearly a vast opportunity for the development of new tools in these areas, and we look forward to seeing what the next decade will bring.

5.4.2. Enabling interoperability between databases

Natural products databases span a wide range of subject areas (structures, BGCs, geographic origin, taxonomic origin etc.). However, because the field is very large and data curation is slow, most databases are designed with narrow scope. This has led to a proliferation of small databases with partial overlap in terms of content, and no standardization of included fields.

A number of technologies exist which could facilitate the exchange of data between databases. In particular, the advent of the specifications for the Semantic Web (or Web 3.0) by the World Wide Web Consortium (W3C, https://www.w3.org/standards/semanticweb/) would greatly facilitate data interchange. These technologies include Resource Description Framework, Web Ontology Language, and JSON-LD, amongst many others. Implementing tools like this affords structured and linked datasets and is currently driving a change in how data is handled on the internet. Practically, these technologies make data machine-readable and are currently leveraged heavily by the web's largest driving forces, including Google and Amazon. Unfortunately, we have yet to see these technologies realized in the field of natural products. This is due in large part to the depth of technical knowledge required to implement these requirements.

A simpler approach is the development of web APIs with well-defined schemas for existing online tools. APIs can deliver data in JSON or XML format, permitting real-time extraction of information from different resources, and eliminating the need for the duplicate storage of key data. Replication of the same data in different repositories is a basic “no-no” in database science, because of the challenges associated with ensuring that both copies are always correctly synchronized.

Creating APIs not only enables the faster development of front-end tools such as data summary dashboards or detailed data pages, but it also provides informaticians with methods to more easily access and interrogate data. This in turn reduces the barrier to access to ask new questions in the field, and catalyses the exploration and development of new ideas.

To be interoperable, databases require at least one unique field that is the same in each dataset (the “primary key”). Realistically, chemical structures are the only practical option as the primary key between natural products databases. To be useful, structures must therefore be entered consistently in all cases. Database creators must decide how to handle a large number of complicated situations including: entering racemates as one compound or two, including or excluding salt forms, handling atropisomers and metal complexes, managing partial and missing configurations, identifying and updating structures that have been corrected in subsequent studies, etc.

An ideal scenario would be to have a central, comprehensive database of all natural product structures to which other resources could refer. This would vastly increase the speed of database creation (by eliminating the need to curate the structure component) and would automatically align all of these resources (via the central structure ID). Sadly, no such database currently exists. In the absence of such a resource, database managers are encouraged to cooperatively define compound standardization strategies, and to manually review and align structural data between resources. This unglamorous task receives little recognition in the community, meaning that it is a low priority for most academic research groups. Until the natural products community develops guidelines and standards for data curation, this situation will likely persist, which presents a considerable threat that the value and opportunity offered by comparing datasets from different subject areas will be lost.

5.5. Future perspective

Data-centric approaches have fundamentally altered the landscape in many areas of natural science. For example, from the laborious early determination of protein crystal structures in the 1960s, protein biochemistry has evolved to a sophisticated field where even non-experts can perform large-scale, automated docking studies of virtual libraries against almost any biological target. Similarly, the longstanding effort to create KEGG as an encyclopaedia of gene function [71] is enabling the development of tools for the automated annotation of gene function across genomes and metagenomes (e.g., BlastKOALA and GhostKOALA [72]).

Natural products science has yet to take full advantage of this changing landscape of scientific discovery. Many discovery programs remain focused on manual methods, without effectively leveraging prior knowledge in the field. This is evidenced by high rates of compound rediscovery and the heterologous expression of “unusual” BGCs that turn out to produce well-known compound classes. While this cannot always be avoided, better data integration of chemical structure data, genomic data and metabolomic data has a clear potential to improve prioritization of research efforts.

The opportunities offered by developing new data-driven discovery methods are clear. However, it is unreasonable to expect that researchers involved in tool development will also create the basal datasets required to power these tools. Instead, we must commit resources to the creation of large, well-structured repositories of key information, and must develop a culture where data deposition of new results is a standard and expected part of the discovery workflow. If we can accomplish these goals, the return on this investment will be felt powerfully in every corner of natural products science.

Acknowledgements

We thank Drs G. Carter, C. Pearce, J. Gloer, M. Balunas, S. Singh, J. Blunt and D. Newman for helpful discussions, and Dr H. Potter for providing statistics on MarinLit content.

Funding

Funding for this work was provided by NIH grants U41-AT008718 and U24-AT010811 (RGL) and the Graduate School for Experimental Plant Sciences, The Netherlands (SAK).

Conflicts of Interest Statement

MHM is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio.

References

1. Schulz H, Georgy U. From CA to CAS online. 1994. doi:10.1007/978-3-642-78663-1
2. Turner WB, Others. Fungal metabolites. Fungal metabolites. 1971. Available: https://www.cabdirect.org/cabdirect/abstract/19731307247
3. Rosazza JP. Fungal metabolites. Vol. II. By W. B. Turner and D. C. Aldridge. Academic Press, 111 Fifth Avenue, New York, NY 10003. 1983. 631 pp. 16 × 23.5 cm. Price: $80.00. Journal of Pharmaceutical Sciences. 1984. p. 1878. doi:10.1002/jps.2600731270
4. Bérdy J. CRC Handbook of Antibiotic Compounds. CRC Press; 1980.
5. Gräfe U. John S. glasby, encyclopedia of antibiotics (third edition), 515 S. chichester-New

York-Brisbane-Toronto-Singapore 1993. John Wiley & sons. £ 99.95. ISBN: 0471–92922-0. J Basic Microbiol. 1995;35: 32–32.

6. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One. 2012;7: e34064.
7. Reddy BVB, Milshteyn A, Charlop-Powers Z, Brady SF. eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. Chem Biol. 2014;21: 1023–1033.
8. Potts MB, Kim HS, Fisher KW, Hu Y, Carrasco YP, Bulut GB, et al. Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. Sci Signal. 2013;6: ra90.
9. Ibrahim A, Yang L, Johnston C, Liu X, Ma B, Magarvey NA. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc Natl Acad Sci U S A. 2012;109: 19196–19201.
10. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87.
11. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: W339–46.
12. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform. 2014;2: e1.
13. IDC report: The digitization of the world from edge to core. [cited 24 Jan 2021]. Available: https://resources.moredirect.com/white-papers/idc-report-the-digitization-of-the-world-from-edge-to-core
14. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.
15. Micallef L, Rodgers P. eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. PLoS One. 2014;9: e101717.
16. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminform. 2020;12: 20.
17. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. Nucleic Acids Res. 2018;46: D1217–D1222.
18. Klementz D, Döring K, Lucas X, Telukunta KK, Erxleben A, Deubel D, et al. StreptomeDB 2.0--an extended resource of natural products produced by streptomycetes. Nucleic Acids Res. 2016;44: D509–14.
19. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019;5: 1824–1833.
20. Bycroft BW, Payne DJ. Dictionary of Antibiotics and Related Substances: with CD-ROM, Second Edition. CRC Press; 2013.
21. Conway KR, Boddy CN. ClusterMine360: a database of microbial PKS/NRPS biosynthesis. Nucleic Acids Res. 2013;41: D402–7.
22. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2020;48: D454–D458.
23. Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res. 2020;48: D422–D430.
24. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
25. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42: D699–704.
26. Blin K, Pascal Andreu V, de Los Santos ELC, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2019;47: D625–D630.
27. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional

annotation. Nucleic Acids Res. 2016;44: D733–45.

28. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34: 828–837.
29. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48: D440–D444.
30. McAlpine JB, Chen S-N, Kutateladze A, MacMillan JB, Appendino G, Barison A, et al. The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. Nat Prod Rep. 2019;36: 35–107.
31. Sorkin BC, Betz JM, Hopp DC. Toward FAIRness and a User-Friendly Repository for Supporting NMR Data. Org Lett. 2020;22: 2867.
32. López-Pérez JL, Therón R, del Olmo E, Díaz D. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. Bioinformatics. 2007;23: 3256–3257.
33. Steinbeck C, Kuhn S. NMRShiftDB -- compound identification and structure elucidation support through a free community-built web database. Phytochemistry. 2004;65: 2711–2717.
34. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res. 2008;36: D402–8.
35. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46: D608–D617.
36. Asakura K. A NMR spectral database of natural products "CH-NMR-NP." J Synth Org Chem Jpn. 2015;73: 1247–1252.
37. Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, et al. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform. 2013;5: 3.
38. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47: D1102–D1109.
39. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res. 2016;44: D1214–9.
40. Banik GM, Baysinger G, Kamat PV, Pienta NJ, American Chemical Society. The ACS Guide to Scholarly Communication. American Chemical Society; 2020.
41. A new look for chemical information. Nat Chem Biol. 2007;3: 297.
42. Olson JE. Database Archiving: How to Keep Lots of Data for a Very Long Time. Morgan Kaufmann; 2010.
43. Epstein SC, Charkoudian LK, Medema MH. A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences. Stand Genomic Sci. 2018;13: 16.
44. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A. 2017;114: 5601–5606.
45. Pascolutti M, Campitelli M, Nguyen B, Pham N, Gorse A-D, Quinn RJ. Capturing nature's diversity. PLoS One. 2015;10: e0120942.
46. O'Hagan S, Kell DB. Analysing and Navigating Natural Products Space for Generating Small, Diverse, But Representative Chemical Libraries. Biotechnol J. 2018;13. doi:10.1002/biot.201700503
47. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. PLoS One. 2013;8: e62839.
48. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5: 170203.
49. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2: 1533–1542.
50. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37: 953–961.
51. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021;39: 105–114.
52. Hoffmann T, Krug D, Bozkurt N, Duddela S, Jansen R, Garcia R, et al. Correlating chemical

diversity with taxonomic distance for discovery of natural products in myxobacteria. Nat Commun. 2018;9: 803.

53. Reher R, Kim HW, Zhang C, Mao HH, Wang M, Nothias L-F, et al. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. J Am Chem Soc. 2020;142: 4114–4120.
54. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform. 2016;8: 3.
55. Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018;9: 4035.
56. Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, Arndt D, et al. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. Metabolites. 2019;9. doi:10.3390/metabo9040072
57. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019;16: 299–302.
58. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A. 2015;112: 12580–12585.
59. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. Proc Natl Acad Sci U S A. 2016;113: 13738–13743.
60. Rogers S, Ong CW, Wandy J, Ernst M, Ridder L, van der Hooft JJJ. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. Faraday Discuss. 2019;218: 284–302.
61. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature. 2018;558: 440–444.
62. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020;16: 60–68.
63. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560: 233–237.
64. Krause J, Handayani I, Blin K, Kulik A, Mast Y. Disclosing the Potential of the SARP-Type Regulator PapR2 for the Activation of Antibiotic Gene Clusters in Streptomycetes. Front Microbiol. 2020;11: 225.
65. Gregory K, Salvador LA, Akbar S, Adaikpoh BI, Stevens DC. Survey of Biosynthetic Gene Clusters from Sequenced Myxobacteria Reveals Unexplored Biosynthetic Potential. Microorganisms. 2019;7. doi:10.3390/microorganisms7060181
66. Eng CH, Backman TWH, Bailey CB, Magnan C, García Martín H, Katz L, et al. ClusterCAD: a computational platform for type I modular polyketide synthase design. Nucleic Acids Res. 2018;46: D509–D515.
67. Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol. 2016;12: 1007–1014.
68. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
69. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. ACS Cent Sci. 2016;2: 99–108.
70. Eldjárn GH, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. Cold Spring Harbor Laboratory. 2020. p. 2020.06.12.148205. doi:10.1101/2020.06.12.148205
71. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45: D353–D361.
72. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016;428: 726–731.

Chapter 6

BiG-SCAPE: A Computational Framework to Explore Large-scale Biosynthetic Diversity

With the increasingly high-throughput genome sequencing technology, large-scale genome mining projects involving hundreds of bacterial isolates have become a common practice for NP discovery. By grouping similarly structured BGCs into families (GCFs), relationships between homologous BGCs can be captured to guide their prioritization. In this chapter, we provide a streamlined computational workflow consisting two new software tools: BiG-SCAPE, which facilitates fast and interactive sequence similarity network analysis of BGCs and GCFs; and CORASON, which elucidates phylogenetic relationships within and across these families. We validated BiG-SCAPE by correlating its output to metabolomic data across 363 actinobacterial strains and demonstrated the discovery potential of CORASON by comprehensively mapping biosynthetic diversity across a range of detoxin/rimosamide-related GCFs, culminating in the characterization of seven detoxin analogues.

Originally published as:

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A.,
Tryon, J. H., Parkinson, E. I., ... & Roeters, A. (2020).
A computational framework to explore large-scale biosynthetic diversity.
Nature Chemical Biology, 16(1), 60-68.

6.1. Introduction

Specialized microbial metabolites are key mediators of interspecies communication and competition in the environment and in the context of host microbiomes [1,2]. Their diverse chemical structures have been critical in the development of antibiotics, anticancer drugs, crop protection agents, food additives and cosmeceuticals. Although tens of thousands of natural products have been discovered in past decades, recent evidence suggests that these represent a fraction of the potential natural product chemical space yet to be discovered [3–8].

Genome mining has emerged in the past decade as a key technology to explore and exploit natural product diversity. Key to this success is the fact that genes encoding natural product biosynthetic pathways are usually clustered together on the chromosome. These biosynthetic gene clusters (BGCs) can be readily identified in a genome sequence. Moreover, in many cases, the chemical structures of their products can be predicted to a certain extent, based on the analysis and biosynthetic logic of the enzymes encoded in a BGC and their similarity to known counterparts [9].

Initially, genome mining was performed on a single-genome basis: a research group or consortium would sequence the genome of a single microbial strain and attempt to identify and characterize each of its BGCs one by one. This approach has revealed much about the metabolic capacities of model natural-product-producing organisms such as *Streptomyces coelicolor*, *Sorangium cellulosum* and *Aspergillus nidulans*, and has provided clues regarding the discovery potential [10] from corresponding genera [11–13]. Computational tools for the identification of BGCs and the prediction of their products' chemical structures, such as antiSMASH [14] and PRISM [15] have played a key role in the success of genome mining. These *in silico* approaches have been strengthened by comparative analysis of identified BGCs with biochemical reference data, such as those provided by the MIBiG (Minimum Information about a Biosynthetic Gene cluster) community effort [16].

Fueled by rapid developments in high-throughput sequencing, genome mining efforts are now expanding to large-scale, pan-genomic mining of entire bacterial genera [4,17,18], strain collections [19] and metagenomic datasets, from which thousands of metagenome-assembled genomes can be extracted at once [20–23]. Such studies pave the path toward systematic investigations of the biosynthetic potential of broad taxonomic groups of organisms, as well as entire ecosystems. These large-scale analyses easily lead to the identification of thousands of BGCs with varying degrees of mutual similarity, ranging from widely distributed homologues of gene clusters for the production of well-known molecules to rare or unique gene clusters that encode unknown enzymes and pathways.

To map and prioritize this complex biosynthetic diversity, several groups have devised methods to compare architectural relationships between BGCs in sequence similarity networks and group them into gene cluster families (GCFs), each of which contains BGCs across a range of organisms that are linked to a highly similar natural product chemotype [3,4,24,25]. Such GCFs can be matched to molecular families identified from mass spectrometry (MS) data based on observed/predicted chemical features [26–28]. Alternatively, their presence or expression can be statistically correlated to the presence of molecular families in MS data in a process termed “metabologenomics” [4,29–31]. However, current methods fail to correctly measure the similarity between complete and fragmented gene clusters (which frequently occur in metagenomes and large-scale, pan-genome, sequencing projects based on short-read technologies), do not consider the complex and multi-layered evolutionary relationships within and between GCFs, require lengthy compute times and large-scale computing facilities when processing large datasets and lack a user-friendly implementation that interacts directly with other key resources. These shortcomings preclude adoption by the broader scientific community and impede substantial advances in natural product discovery.

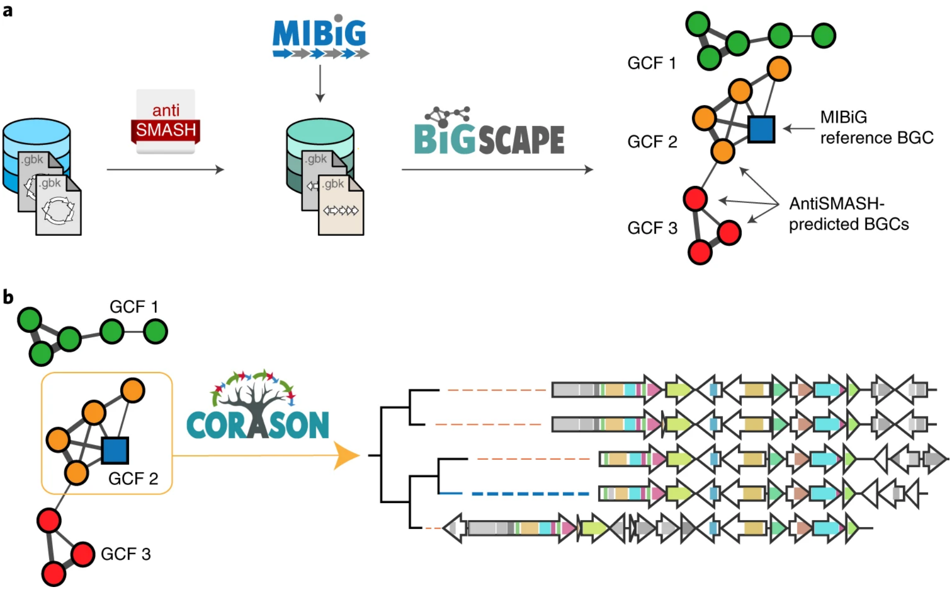


Figure 6.1. (a) The BiG-SCAPE approach analyzes a set of antiSMASH-detected BGCs to construct a similarity network and groups them into GCFs, together with MIBiG reference BGCs (indicated in blue). (b) Subsequently, CORASON-based, multi-locus, phylogenetic analysis is used to illuminate evolutionary relationships of BGCs within each GCF.

In the present study, a streamlined computational workflow is provided that tightly integrates two new software tools, BiG-SCAPE and CORASON (https://bigscape-corason.secondarymetabolites.org), with the gene cluster identification and empirical biosynthetic data comparison possible through antiSMASH and MIBiG (Figure 6.1). BiG-SCAPE facilitates rapid calculation and interactive exploration of BGC sequence similarity networks; it accounts for differences in modes of evolution between BGC classes, groups gene clusters at multiple hierarchical levels and introduces a “glocal” alignment mode to handle fragmented BGCs. CORASON employs a phylogenomic approach to elucidate evolutionary relationships between gene clusters by computing high-resolution, multi-locus phylogenies of BGCs within and across GCFs; in addition, it allows researchers to comprehensively identify all genomic contexts in which gene cassettes of interest ("subclusters" within larger BGCs) can be found. Our present study confirms that metabologenomic correlations accurately connect GCFs to mass features across metabolomic data from 363 strains. Furthermore, the power of the combined workflow is demonstrated, together with the EvoMining algorithm [7], to comprehensively map biosynthetic diversity by identifying three new families responsible for the biosynthesis of new detoxins.

6.2. Results

6.2.1. Large-scale network analysis and classification of BGCs

To provide a streamlined, scalable and user-friendly software for exploring and classifying large collections of gene clusters, BiG-SCAPE was built, written in Python and is freely available as open source software. BiG-SCAPE takes BGCs predicted by antiSMASH or annotated in MIBiG as inputs to automatically generate sequence similarity networks and assemble GCFs.

In previous studies [3,4], two sets of distance metrics had been independently developed to measure the (dis)similarity of pairs of BGCs. In BiG-SCAPE, the aim was to combine the respective strengths of both approaches. The strength of the former approach was the elegant compression of gene clusters into strings of Pfam domains [32], combined with the Jaccard index (JI) to measure domain content similarity (Figure 6.2a). However, an informative index for synteny conservation had been missing. To this end, an adjacency index (AI) was added, which measures how many pairs of adjacent domains are shared between gene clusters (see Supplementary Note 1). Also, sequence identity is an important parameter, because Pfam domains are often very broad and frequently comprise a wide range of enzyme subfamilies with different catalytic activities or substrate specificities. Yet, previous approaches suffered from extremely long compute times when including sequence identity calculations, requiring the use of supercomputers that would preclude day-to-day use. The underlying issue is that comparing large numbers of protein sequences from many BGCs is an all-versus-all problem that scales quadratically when the size of the data increases. To mitigate this, all-versus-all calculations were replaced with all-versus-profile calculations, by aligning each protein domain sequence to its profile Hidden Markov Model from Pfam using the hmmalign tool from the HMMER suite (http://hmmer.org). This leads to a marked speed increase compared with conventional, multiple-sequence alignment using MUSCLE or MAFFT, especially for large numbers of sequences. The profile-based alignment was input into the domain sequence similarity (DSS) index, which measures both Pfam domain copy number differences and sequence identity. The combination of JI, AI and DSS indices into a new combined metric constitutes a fast and informative method to calculate distances between BGCs. BiG-SCAPE obtains very similar results in a fraction of the time compared with the previously published method [4] (see **subchapter 6.4**).

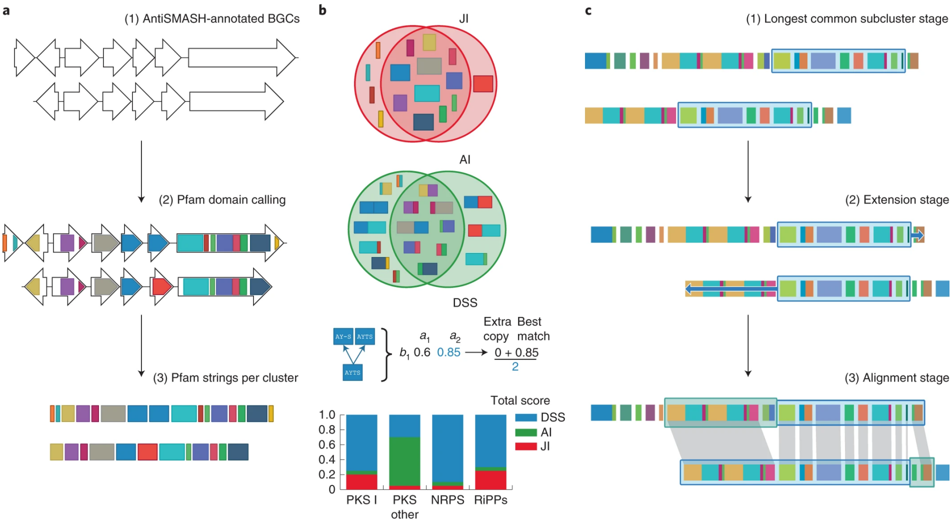


Figure 6.2. (a) Input data consist of BGC sequences directly imported from antiSMASH runs and/or MIBiG. Nucleotide sequences are translated and represented as strings of Pfam domains. (b) The three metrics that are combined in a single distance include the JI, which measures the percentage of shared types of domains, the AI, which measures the percentage of pairs of adjacent domains, and the DSS, which is a measure of sequence identity between protein domains encoded in BGC sequences. Weights of these indices have been optimized separately for different BGC classes. For simplicity, only four classes are shown. (c) In glocal mode, BiG-SCAPE starts with the longest common subcluster of genes between a pair of BGCs and attempts to extend the selection of genes for comparison using a match/mismatch penalty system.

One notable limitation of a generic distance metric is that different classes of BGCs have different evolutionary dynamics. For example, the chemical structures of aryl polyenes have been shown to remain very stable across large evolutionary timescales, whereas the amino acid sequence identity between their key biosynthetic enzymes has become less than 30–40% [3]. On the other hand, the structures of rapamycin-family polyketides exhibit major differences even when sequence identities are as high as ~80% [33]. Although there is not enough information available to construct individual metrics for each specific natural product family, specific weights of the JI, AI and DSS indices were calibrated for BGCs encoding eight different BiG-SCAPE classes (type I polyketide synthases (PKSs), other PKSs, nonribosomal peptide synthetases (NRPSs), PKS/NRPS hybrids, ribosomally synthesized and post-translationally modified peptides (RiPPs), saccharides, terpenes and others; Figure 6.2b) by choosing the weight combination that maximized the correlation between BGC and compound distances for every pair of BGCs in the same class (see **subchapter 6.4** and Supplementary Note 2). In the BiG-SCAPE output, separate networks are generated for each BiG-SCAPE class, along with an optional overall network that combines BGCs from all classes (see Supplementary Table 1).

Another problem of previous approaches for calculating distances between BGCs was how to handle comparisons between complete and partial BGCs (for example, from fragmented genome assemblies), as well as comparisons with pairs of genomically adjacent BGCs that are merged by BGC identification tools. Both global similarity (used in all previous methods) and local similarity lead to artifacts in such cases. To compare the appropriate corresponding regions between BGCs, a new glocal alignment mode was introduced, which first finds the longest common substring between the Pfam strings of a BGC pair, and then uses match/mismatch penalties to extend this alignment (Figure 6.2c and see **subchapter 6.4**). Information about whether an antiSMASH-annotated cluster is located at the edge of a contig can also be used to automatically select a third pairwise distance calculation mode, which relies on global alignment for complete clusters and glocal alignment when at least one of the BGCs in a pair is fragmented.

BGC sequence similarity networks are then generated by applying a cutoff to the distance matrix calculated by BiG-SCAPE. Subsequently, two rounds of affinity propagation clustering [34] are performed to group BGCs into GCFs, and GCFs into “gene cluster clans” (GCCs) (see **subchapter 6.4**). Although tighter (lower) cutoffs are more appropriate for grouping BGCs that produce identical compounds, looser (higher) cutoffs provide a broader perspective on related families of natural products. This process of categorization facilitates calculating metabologenomic correlations [30,35] at multiple levels.

6.2.2. Validation using large-scale metabolomics data

To verify that BiG-SCAPE can group BGCs that are known to be related, a chemical similarity network was constructed from all products of BGCs in MIBiG (see **subchapter 6.4**), and this was used to derive a curated set of 376 compounds, which were manually classified into 92 groups (for example, fourteen-membered macrolides, benzoquinone ansamycins, quinomycin antibiotics and so on; see Supplementary Dataset) and nine classes (for example, polyketides, NRPs, RiPPs and so on). Then, BiG-SCAPE was used to group the corresponding BGCs into GCFs, and good correspondence was observed between manually curated families and those predicted by BiG-SCAPE (see Supplementary Figure 1).

Arguably, the greatest value of BiG-SCAPE lies in the practical use of the predicted GCFs for discovery applications. Hence, the accuracy of correlations of BiG-SCAPE-predicted GCFs to MS ions was assessed from known natural products through metabologenomics [30]. First, a BiG-SCAPE analysis of 74,652 BGCs from 3,080 actinobacterial genomes (see **subchapter 6.4**) was performed, including 1,393 reference BGCs from MIBiG. BiG-SCAPE grouped these BGCs into a total of 17,718 GCFs and 801 GCCs using default parameters.

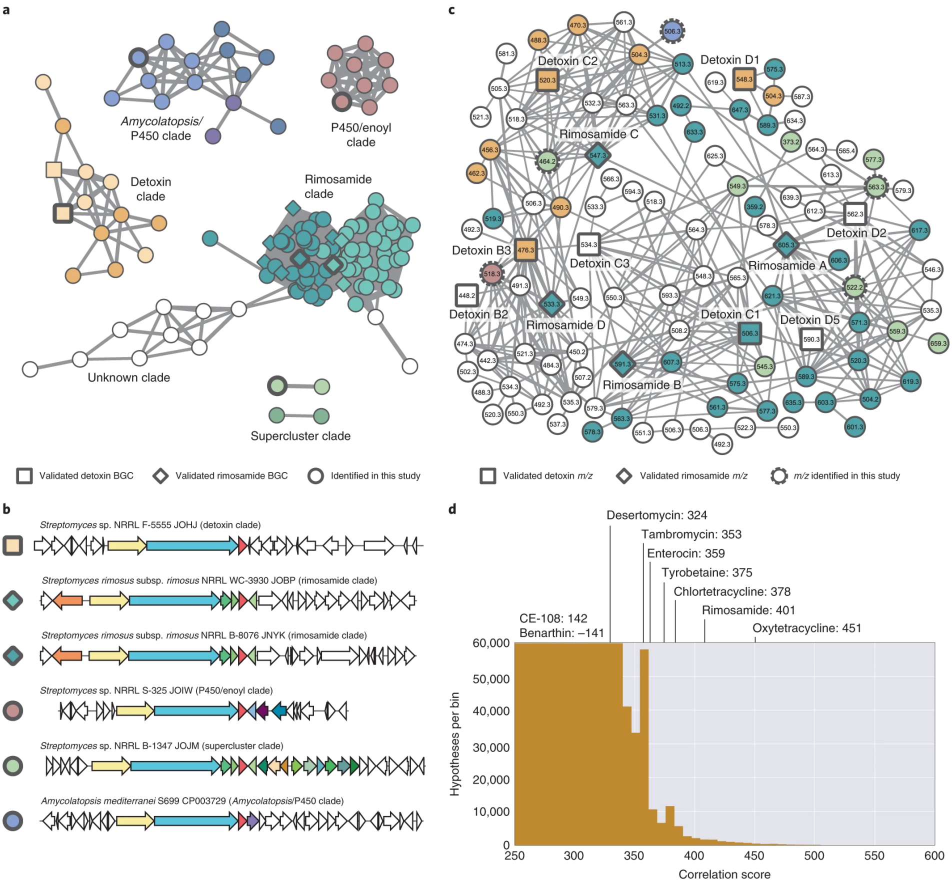


Figure 6.3. (a) Detail of a BiG-SCAPE network containing validated detoxin and rimosamide BGCs, filtered for the presence of the TauD domain. BiG-SCAPE GCF classifications include the rimosamide (turquoise shades) and detoxin (orange shades) families, as well as the *"Amycolatopsis*/P450" (violet shades), "P450/enoyl" (pink) and "supercluster" (light-green shades) families explored in the present study. (b) Validated BGCs represented by bold-outlined nodes. (c) The detoxin and rimosamide molecular family, based on MS/MS data of a 363-strain actinomycete library, is colored by BiG-SCAPE family. Known detoxin (squares) and rimosamide (diamonds) nodes have solid bold outlines, whereas putative detoxins are circular nodes and new analogues from the present study are indicated by bold, dotted outlines. (d) Histogram of all ion-GCF correlation scores resulting from the metabologenomics round run with a 0.30 glocal distance cutoff. Known ion-GCF pair correlation scores are overlaid; six of nine appear in the "tail" of the distribution, which would be indicative of a true connection. The low scoring for benarthin is due to the complicated fragmentation pattern of its BGCs (see Supplementary Figure 3).

Extracts from 363 actinomycete strains were analyzed using untargeted high-resolution liquid chromatography–tandem MS (LC–MS/MS) [4,30,35,36]. Exploration of gene cluster networks and molecular networks [37] highlighted the high diversity in both gene clusters and molecules, for example, 105 different BGCs were identified (at default <0.3 distance) related to known detoxin/rimosamide gene clusters (Figure 6.3a and 6.3b) and 110 different molecules were identified related to detoxins and rimosamides (Figure 6.3c). The GCF annotations for all 363 strains from two BiG-SCAPE modes (global and glocal) at two distance cutoffs (0.30 and 0.50) were used to generate and compare four rounds of metabologenomic correlations, using a binary scoring metric (see Supplementary Figure 2) as described previously [4,30]. BiG-SCAPE's GCF annotations were then assessed against ion production patterns. A test dataset of nine known ion signals and their characterized gene clusters (for CE-108, benarthin, desertomycin, tambromycin, enterocin, tyrobetaine, chlortetracycline, rimosamide and oxytetracycline), which were known to be present across multiple strains in the data, were manually tracked across the four correlation rounds. Based on the metabologenomic analysis of the four rounds, the glocal mode with a 0.3 distance cutoff (Figure 6.3d) was chosen as the default for BiG-SCAPE (see Supplementary Tables 2 and 3). Using these parameters, the analysis showed that at least six of these nine molecule to GCF combinations ended up in the rightmost tail of the distribution of all correlation values, which would indicate a possible/likely connection if it were used as a prediction (Figure 6.3d and see **subchapter 6.4**).

6.2.3. BGC phylogenies resolve evolutionary relationships

Genetic diversity of BGCs within GCFs is often directly related to structural differences between their molecular products, and even small chemical variations can lead to different biological activities [33]. Hence, mapping the evolutionary relationships between BGCs within and across GCFs is crucial for the discovery process. To this end, the CORASON software was introduced, written in Perl and available open source (Figure 6.4). Given a query gene inside a BGC of interest, the CORASON pipeline identifies other genomic loci that contain homologues of this gene and identifies the conserved core of these loci (Figure 6.4a). Based on this core, a multi-locus, approximately maximum-likelihood, phylogenetic tree [38] is constructed to reveal clades that may be responsible for the biosynthesis of different types of chemistry, due to the association of specific types of additional enzyme-coding genes (Figure 6.4b). This procedure can be performed for the "core" enzyme-coding genes of a BGC, but also for, for example, tailoring genes, to reveal various GCFs that would probably produce molecules with similar chemical modifications (Figure 6.4c).

CORASON is available as a downloadable software and also allows working with customizable genomic databases. A version of the CORASON algorithm, called "family-mode", was also integrated with BiG-SCAPE; this generates a multi-locus phylogeny of all BGCs within each GCF using the sequences of their common domain core.

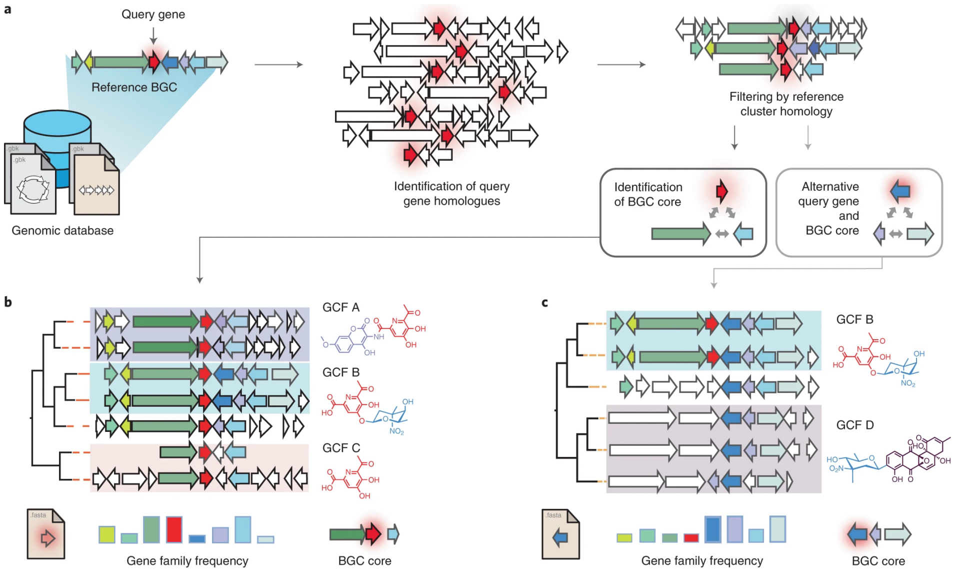
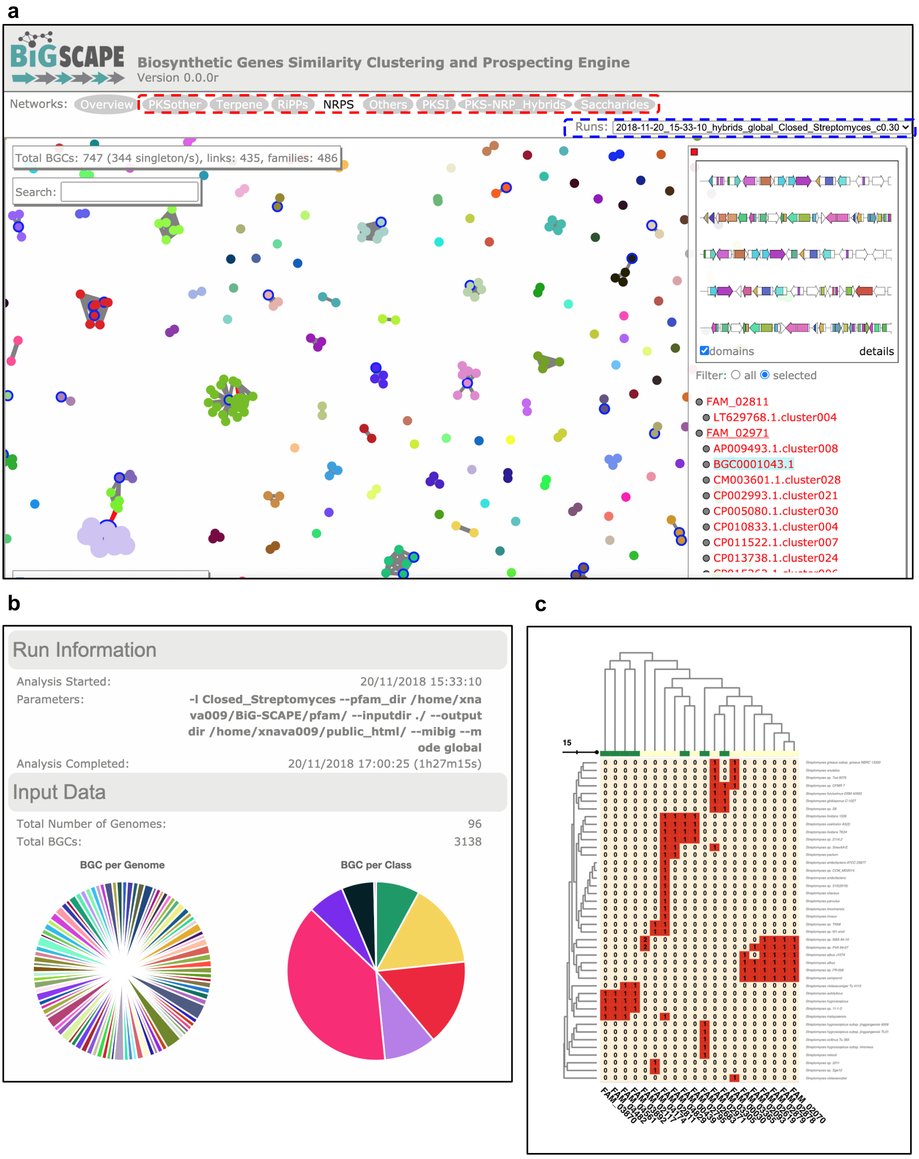

a
Query gene
Reference BGC
Genomic database
Identification of query gene homologues
Filtering by reference cluster homology
Identification of BGC core
Alternative query gene and BGC core
b
GCF A
GCF B
GCF C
Gene family frequency
BGC core
c
GCF B
GCF D
Gene family frequency
BGC core

Figure 6.4. (a) Given a query gene in a reference cluster and a custom genome database, CORASON 1- searches for query gene homologues, 2- creates a CVD by filtering out all genomic loci not related to the reference BGC, but keeping fragmented clusters and 3- identifies the CVD gene core based on multidirectional best hits. (b) Then, CORASON infers a phylogenetic tree by curation and concatenation of the CVD gene core, and calculates the frequency of occurrence for each gene family from the reference BGC. The tree will reveal clades of BGCs that may correspond to GCFs from BiG-SCAPE, and may be responsible for the production of different structural analogues of a natural product family. (c) With the same reference BGC, if a new query gene is selected from accessory enzymes rather than the current CVD core, CORASON will visualize a new phylogeny. This tree may contain clades that correspond to GCFs with diverse biosynthetic cores (of scaffold biosynthesis enzymes) that encode the same molecular modifications in different contexts.

6.2.4. An integrated workflow and interactive visualization

BiG-SCAPE and CORASON connect seamlessly with antiSMASH and MIBiG, because GenBank outputs of antiSMASH can be used directly as inputs for the workflow, and MIBiG reference data can be included in the analysis automatically. Although calculations on hundreds or thousands of genomes are too computationally intensive to provide them on a free public web server, the results of each BiG-SCAPE run are still made available in an interactive HTML visualization that enables efficient exploration of biosynthetic diversity across large datasets for nonprogrammers (Figure 6.5). For every run, BiG-SCAPE will output a self-containing folder called “html_content”, consisting of a static webpage file named “index.html”, along with a range of other system- and data-related files: a “js” folder for storing javascript libraries and modules, a “css” folder for storing web-style settings, an “img” folder for storing media-related contents, and a “network_*” folder for storing each run’s data in JSON-formatted files.


a
BiG SCAPE
Networks: Overview PKSother Terpene RiPPs NRPS Others PKSI PKS-NRP_Hybrids Saccharides
Search:
domains
details
b
BGC per Genome
BGC per Class
c
15

Biosynthetic Genes Similarity Clustering and Prospecting Engine
Version 0.0.0r

Runs: 2018-11-20_15-33-10_hybrids_global_Closed_Streptomyces_c0.30

Total BGCs: 747 (344 singleton/s), links: 435, families: 486

Filter: ○ all ● selected

- FAM_02811
  - LT629768.1.cluster004
- FAM_02971
  - AP009493.1.cluster008
  - BGC0001043.1
  - CM003601.1.cluster028
  - CP002993.1.cluster021
  - CP005080.1.cluster030
  - CP010833.1.cluster004
  - CP011522.1.cluster007
  - CP013738.1.cluster024

Run Information

Analysis Started:	20/11/2018 15:33:10
Parameters:	-l Closed_Streptomyces --pfam_dir /home/xnava009/BiG-SCAPE/pfam/ --inputdir ./ --outputdir /home/xnava009/public_html/ --mibig --mode global
Analysis Completed:	20/11/2018 17:00:25 (1h27m15s)

Input Data

Total Number of Genomes:	96
Total BGCs:	3138

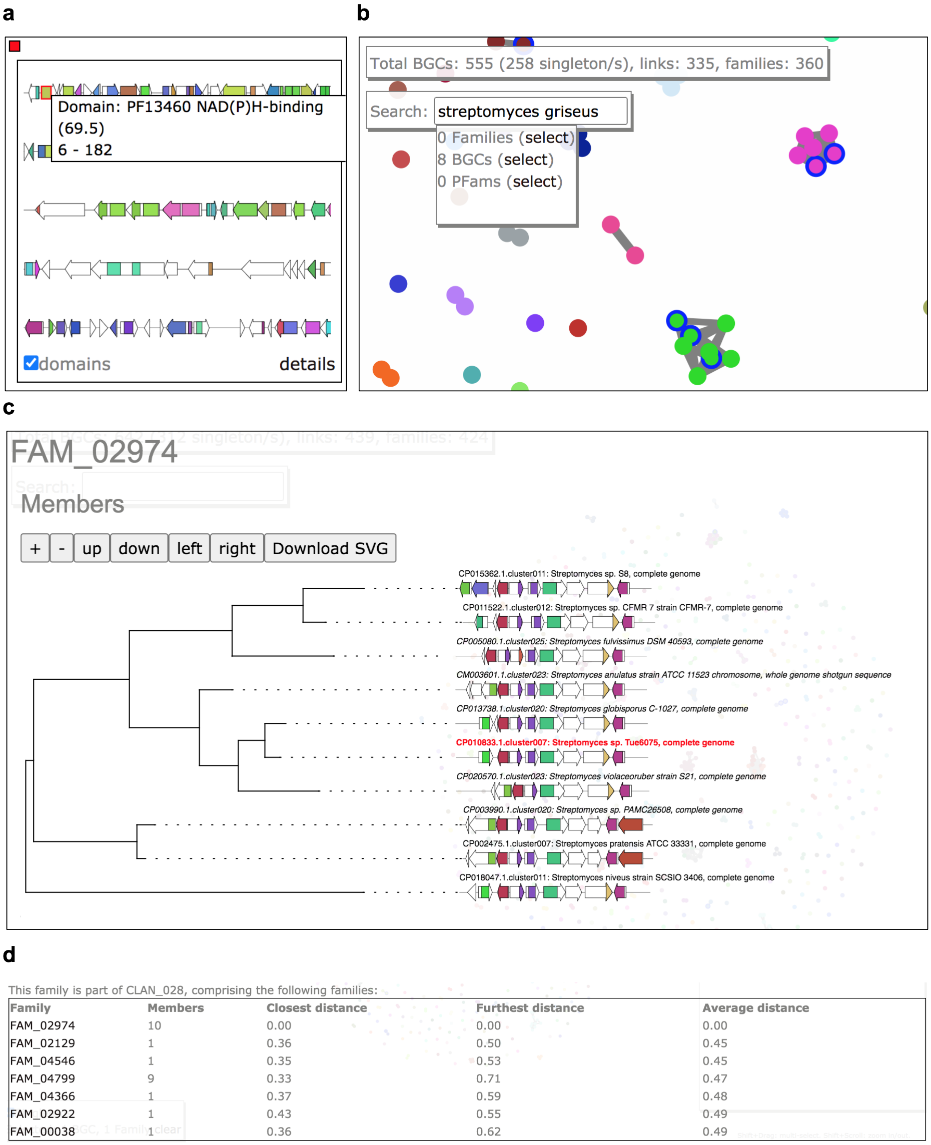
Figure 6.5. Screenshots of several features contained within BiG-SCAPE's interactive visualization module: (a) GCF network visualization showing dynamic interactions between nodes and edges, (b) general summary showing run parameters and input metadata and (c) GCF-to-genome absence/presence heatmap.

To access this interactive output visualization, users can double-click on the "index.html" file on their computer; then, the embedded software will open n any modern internet browser (Mozilla Firefox is recommended for the best compatibility). Users will then first access an overview page showing general information about the run, such as parameters and input data statistics (Figure

6.5c). Moreover, an absence/presence heatmap of GCFs across genomes (Figure 6.5b) is also shown on the right of this page. GCFs (columns) can be clustered by genomic presence or by the averaged distance matrix between the GCF's BGCs. The genomes (rows) can be clustered by either GCF presence or by supplied genome names. To support further downstream analyses, the active selection can be downloaded into a comma-separated (CSV) file by clicking on the respective "Download" button above the heatmap.

If a user performed multiple runs using the same dataset (simply by designating the same output folder for all the runs), the result for any specific run can be selected via a dropdown box at the top right corner of the web page (Figure 6.5a, dashed blue box). By default, the runs will be named according to the starting date/time, run mode (e.g., normal/hybrid/mixed, global/glocal) and the input folder name. To assign a custom identifier, users can opt to supply a unique (i.e., not being used before) run name via the `--label` parameter. By default, BiG-SCAPE will split its GCF network results according to BGC classes, and users can view each network by clicking the class-specific button at the top of the web page (figure 6.5a, red dashed box). Normally, BGCs belonging to multiple classes will be placed in their own unique category, e.g., "PKS-NRP_Hybrids", unless a –`-no_classify` parameter is supplied. To enable analysing these BGCs together with their single-class counterparts (in this case, the said PKS/NRP hybrids will be incorporated into both NRP and PKS analysis results), users can turn on the `--hybrid` parameter when starting a new run. Additionally, when the `--mix` parameter is supplied, an extra "Mixed" button will be available, which will show the full network of all BGCs in the dataset together.

In each single view, the network visualization displays BGC nodes colored by GCF in interactive sequence similarity networks, side by side with arrow visualizations of the gene clusters (Figure 6.5a). Each node in the network can be interactively selected and moved using the mouse pointer, and users can zoom in and out of the network using mouse scrolls. When hovered over with the mouse, each colored section within a gene arrow will show its annotation and Pfam details on the underlying domain (Figure 6.6a). To assist in locating specific BGCs in the network, a search can be performed using compound names (for MIBiG reference BGCs), Pfam domains of interest, or species names, with resulting matched nodes being instantly highlighted within the network (Figure 6.6b). Each GCF has its own view panel, which shows the CORASON-based, multi-locus phylogeny of the underlying BGCs (Figure 6.6c) and includes links to related families within the same GCC along with their (maximum, minimum and average) distances to the GCF (Figure 6.6d).


a
b
c
d
domains
details
0 Families (select)
8 BGCs (select)
0 PFams (select)
+
-
up
down
left
right
Download SVG
CP015362.1.cluster011: Streptomyces sp. S8, complete genome
CP011522.1.cluster012: Streptomyces sp. CFMR 7 strain CFMR-7, complete genome
CP005080.1.cluster025: Streptomyces fulvissimus DSM 40593, complete genome
CM003601.1.cluster023: Streptomyces anulatus strain ATCC 11523 chromosome, whole genome shotgun sequence
CP013738.1.cluster020: Streptomyces globisporus C-1027, complete genome
CP010833.1.cluster007: Streptomyces sp. Tue6075, complete genome
CP020570.1.cluster023: Streptomyces violaceoruber strain S21, complete genome
CP003990.1.cluster020: Streptomyces sp. PAMC26508, complete genome
CP002475.1.cluster007: Streptomyces pratensis ATCC 33331, complete genome
CP018047.1.cluster011: Streptomyces niveus strain SCSIO 3406, complete genome

Domain: PF13460 NAD(P)H-binding (69.5)
6 - 182

Total BGCs: 555 (258 singleton/s), links: 335, families: 360

Search: streptomyces griseus

FAM_02974

Members

This family is part of CLAN_028, comprising the following families:

Family	Members	Closest distance	Furthest distance	Average distance
FAM_02974	10	0.00	0.00	0.00
FAM_02129	1	0.36	0.50	0.45
FAM_04546	1	0.35	0.53	0.45
FAM_04799	9	0.33	0.71	0.47
FAM_04366	1	0.37	0.59	0.48
FAM_02922	1	0.43	0.55	0.49
FAM_00038	1	0.36	0.62	0.49

Figure 6.6. (a) An example of the hoverable arrower BGC visualization showing the information of the selected gene region. (b) Input box to search for keywords related to the stored BGC metadata; in this case eight NRP BGCs from *Streptomyces griseus* are highlighted in the network. (c) Interactive phylogenetic tree for a selected GCF, upon which the BGC used as an anchor by the CORASON algorithm is highlighted in red (additionally, when a MIBiG BGC is present, its accession number will be highlighted in blue). This GCF view also includes (d) the list of related GCFs with links to their respective page along with their distance to the selected GCF.

To demonstrate this interactive feature, an example output of an analysis with antiSMASH-predicted BGCs from 103 complete Streptomyces genomes has been provided, including as outgroups the genomes of *Catenulispora acidiphila* and *Salinispora arenicola* (see https://bigscape-corason.secondarymetabolites.org/streptomyces_example/). To connect the absence/presence map of GCFs across these genomes to species phylogeny, a high-resolution, multi-locus, whole-genome phylogeny (see Supplementary Figure 4) was inferred from the *Streptomyces* conserved-core using CORASON, and the tree was decorated with the GCF absence/presence patterns (see Supplementary Figure 5). As has been observed before in other genera such as *Salinispora* [25], this shows high conservation of some GCFs across a larger number of genomes (27 GCFs (~2%) occur across >10 genomes), combined with a large number of rare GCFs that are specific to one or a few genomes (1,564 GCFs (92%) occur across ≤3 genomes).

6.2.5. Case study: identification of new detoxin analogues

Analysis of 3,080 actinobacterial genomes revealed that detoxin and rimosamide BGCs are taxonomically widespread and architecturally diverse. Thus, the present study focused on GCFs of this class to showcase the ability of the BiG-SCAPE/CORASON workflow to analyze and map large, diverse datasets at high resolution [36] (see Supplementary Figure 6 and 7). The conserved core of detoxin and rimosamide BGCs is composed of one NRPS, one NRPS/PKS hybrid and one *tauD*-like gene. The rimosamide BGC differs from those of the detoxins by having an additional NRPS, which codes for an extension of the common detoxin/rimosamide core scaffold with isobutyrate and glycine [36].

The fact that the *tauD* gene is present across all detoxin/rimosamide-related BGCs, but relatively unique within secondary metabolism, caught our attention. The product of the *tauD* gene belongs to the Fe(II)/2-oxoglutarate-dependent hydroxylase enzyme superfamily and is named for the commonly encoded 2-oxoglutarate-dependent taurine dioxygenase (TauD) involved in the oxygenolytic release of sulfite from the amino acid taurine [39]. Interestingly, this family also includes enzymes across fungi, bacteria and plants that catalyze hydroxylations, desaturations, ring expansions and ring formations, among other chemical transformations. To date, the role of TauD in detoxin and rimosamide biosynthesis is unknown, although it has been suggested as being responsible for the proline oxidation observed in some analogues [36].

An EvoMining [7] analysis of the TauD dioxygenase protein family showed specialized metabolism-related expansions of paralogs across genera such as *Streptomyces*, *Rhodococcus*, *Frankia* and *Amycolatopsis* (see Supplementary Figure 8). One expanded clade contained fifteen tauD homologues that belonged to experimentally characterized BGCs from MIBiG v.1.3, as well as one within the rimosamide BGC (see Supplementary Table 4). Next, the genomic contexts of all tauD expansions (comprising 1,175 BGCs) were investigated, with the goal of identifying new detoxin- and rimosamide-related BGCs. The BGCs were processed by CORASON using tauD as the query gene. Although ideally the detoxin/rimosamide BGC core would be defined as also containing the NRPS and NRPS/PKS hybrid genes, herein tauD was used as the sole member of the “BGC core” to allow also for the identification of fragmented BGCs.

Gaps in the genome sequences were observed for some organisms, including *Streptomyces humi* and *Amycolatopsis vancoresmycina* (Figure 6.7). CORASON analysis revealed that the detoxin and rimosamide GCFs identified in BiG-SCAPE were part of a larger GCC related to peptide biosynthesis, which also comprised unexplored clades across the phylum Actinobacteria (Figure 6.7 and see Supplementary Figure 9). Importantly, the high-resolution organization of BGC relationships enabled by the CORASON phylogeny revealed additional BGCs that were omitted by GCF clustering in BiG-SCAPE. This is because the fragmented nature of genome assemblies or the merging of adjacent BGCs by antiSMASH made these BGCs sufficiently different to be classified into different GCFs under the cutoffs used, whereas CORASON could organize BGC relationships based on the single *tauD* gene (see Supplementary Figure 3 and 10).

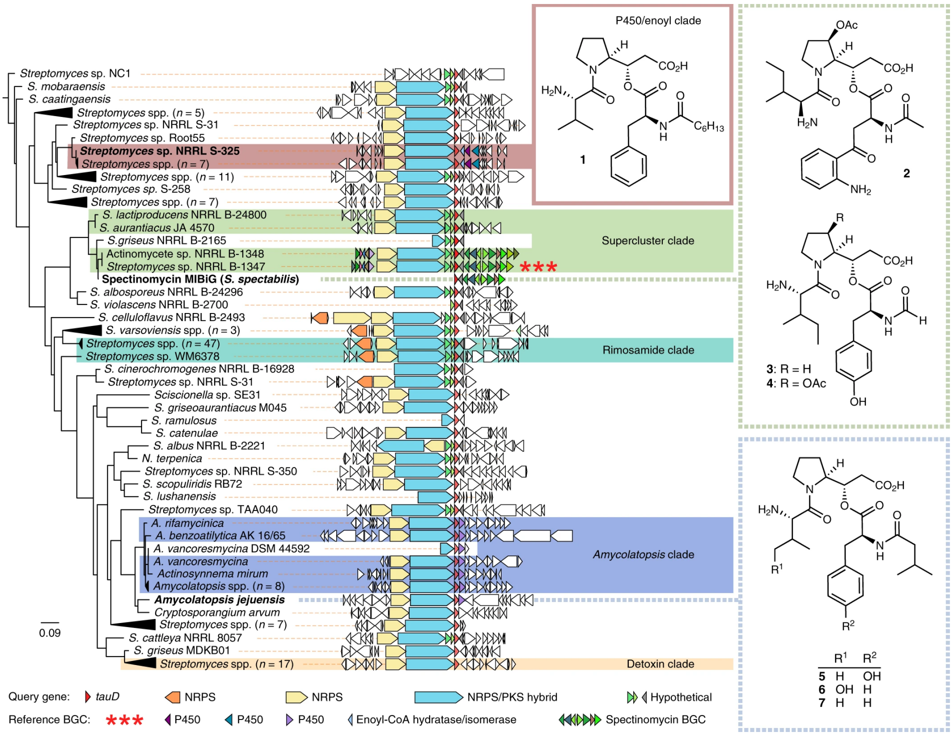

Streptomyces sp. NC1
S. mobaraensis
S. caatingaensis
Streptomyces spp. (n = 5)
Streptomyces sp. NRRL S-31
Streptomyces sp. Root55
Streptomyces sp. NRRL S-325
Streptomyces spp. (n = 7)
Streptomyces spp. (n = 11)
Streptomyces sp. S-258
Streptomyces spp. (n = 7)
S. lactiprodcens NRRL B-24800
S. aurantiacus JA 4570
S.griseus NRRL B-2165
Actinomycete sp. NRRL B-1348
Streptomyces sp. NRRL B-1347
Spectinomycin MIBiG (S. spectabilis)
S. albosporeus NRRL B-24296
S. violascens NRRL B-2700
S. celluloflavus NRRL B-2493
S. varsoviensis spp. (n = 3)
Streptomyces spp. (n = 47)
Streptomyces sp. WM6378
S. cinerochromogenes NRRL B-16928
Streptomyces sp. NRRL S-31
Sciscionella sp. SE31
S. griseoaurantiacus M045
S. ramulosus
S. catenulae
S. albus NRRL B-2221
N. terpenica
Streptomyces sp. NRRL S-350
S. scopuliridis RB72
S. lushanensis
Streptomyces sp. TAA040
A. rifamycinica
A. benzoatilytica AK 16/65
A. vancoresmycina DSM 44592
A. vancoresmycina
Actinosynnema mirum
Amycolatopsis spp. (n = 8)
Amycolatopsis jejuensis
Cryptosporangium arvum
Streptomyces spp. (n = 7)
S. cattleya NRRL 8057
S. griseus MDKB01
Streptomyces spp. (n = 17)
0.09
***
P450/enoyl clade
Supercluster clade
Rimosamide clade
Amycolatopsis clade
Detoxin clade
1
2
3: R = H
4: R = OAc
5
6
7
R¹ R²
5 H OH
6 OH H
7 H H
Query gene: tauD
NRPS
NRPS
NRPS/PKS hybrid
Hypothetical
Reference BGC: ***
P450
P450
P450
Enoyl-CoA hydratase/isomerase
Spectinomycin BGC

Figure 6.7. CORASON phylogenetic reconstruction with *tauD* as the query gene and the *Streptomyces sp.* NRRL B-1347 BGC as query cluster, rooted with *tauD* from *Streptomyces sp.* NC1. Branches of redundant and highly divergent BGCs were compressed for readability (see the uncompressed tree in Supplementary Figure 9; names are followed by their GenBank accession numbers when available). Genes not found in the reference cluster are colored based on BLAST analysis. Highlighted sections on the tree correspond to BiG-SCAPE-defined families. Bolded strain/BGC names were those investigated in the present study, with dotted lines indicating BGCs and detoxins discovered just outside the BiG-SCAPE-defined families. The representative structures for each clade illustrate the correspondence between molecular and genomic variations.

It was hypothesized that the detoxins produced from BGCs in the unexplored clades would contain new chemical variations related to the observed genetic variations. Fortunately, forty of the 152 strains harboring these BGCs were represented in the 363-strain, LC-MS/MS metabolomics dataset. Molecular networking analysis of these data (see **subchapter 6.4**) indicated the presence of eight known detoxins, four known rimosamides and 99 putatively new detoxin or rimosamide analogues (Figure 6.3c), confirming the vast chemical diversity suggested in the BiG-SCAPE/CORASON data.

There were three detoxin BGC clades identified by BiG-SCAPE within the CORASON phylogenetic tree that captured our interest. The first was named the "P450/enoyl clade" because of the presence of putative cytochrome P450 and enoyl-CoA hydratase/isomerase genes in these BGCs. Analysis of MS/MS data from extracts of *Streptomyces sp.* NRRL S-325, which has a BGC from this clade, and comparison with fragmentation patterns of known detoxins, led to the discovery of detoxin S1 (**1**) (see Supplementary Figure 11 and 12). This contained a heptanamide side chain, a unique substructure among the detoxins and rimosamides that is probably installed by the condensation domain of the NRPS, potentially following processing by the predicted enoyl-CoA hydratase/isomerase and cytochrome P450s.

The second clade of interest, termed the "supercluster clade", comprised BGCs with genes related to detoxin biosynthesis immediately adjacent to the known spectinomycin BGC [40]. This was discovered because the spectinomycin MIBiG entry (BGC0000715) clustered with them on the CORASON tree, as it contained a tauD gene at its periphery. Since the *tauD* gene is not known to be involved in spectinomycin biosynthesis, it was hypothesized that there were likely additional detoxin genes adjacent to this spectinomycin BGC in *Streptomyces spectabilis* NRRL-2792. This strain was acquired to determine whether CORASON analysis could facilitate prediction of detoxin production based solely on the presence of a single query gene. MS/MS analysis of the strain's extract revealed production of five detoxin-like natural products (Supplementary Figure 13), including detoxin N1 (**2**), detoxin N2 (**3**) and its acetoxylated analogue, detoxin N3 (**4**). Interestingly, ions with retention times and fragmentation patterns matching the latter two were also observed in extracts of *Streptomyces sp.* NRRL B-1347 from the supercluster clade, confirming the unique ability of CORASON to guide discovery by phylogenetically linking the limited NRRL-2792 sequence data to the detoxin supercluster clade. During finalization of this manuscript, the genome of NRRL-2792 was published [41], and an abbreviated CORASON analysis confirmed the presence of the detoxin BGC in a supercluster configuration with the spectinomycin BGC (see Supplementary Figure 14). LC-MS analysis of NRRL-2792 cultures supplemented with stable isotope-labeled amino acids corroborated structural predictions based on analysis of the closely related *Streptomyces sp.* NRRL B-1347 supercluster and MS/MS data (Supplementary Figure 15-20). All three new analogues were found to fully incorporate labeling from [$^{13}C_6$]isoleucine, but only d_7-proline was fully incorporated into compound **3**. Loss of one deuteron from d_7-proline in compounds **2** and **4** supported assignment of acetoxylation of the pyrrolidine ring, common in reported detoxins and rimosamides [36,42]. Structural features unique to the N-series detoxins included the incorporation of an N-formylated tyrosine in compounds **3** and **4** in place of the typical detoxin/rimosamide phenylalanine residue, which was supported by incorporation of ring-d_4-tyrosine. Compound **2** exhibited the unique incorporation of a tryptophan-derived residue at this position, made evident by retention of four deuterons when fed indole-d_5-tryptophan (see Supplementary Figure 16). Although MS data were insufficient to deconvolute this substructure, compound **2** was produced by *Streptomyces spectabilis* NRRL-2792 in sufficient abundance for isolation and structure elucidation by nuclear magnetic resonance (NMR). Various one-dimensional (1D) and two-dimensional (2D) experiments confirmed assignments from MS data analysis and established an N-acetylated kynurenine as the tryptophan-derived substructure in compound **2** (Supplementary Figure 15-16 and Supplementary Note 3 a-h).

The third detoxin clade that was targeted comprised BGCs that were almost entirely within the genus *Amycolatopsis*. This clade's BGCs also contained a unique predicted cytochrome P450 gene; hence, it was named the "*Amycolatopsis*/P450 clade". Although there were no metabolomics data for strains with BGCs in the BiG-SCAPE-defined GCF, the CORASON visualization allowed the selection of an *Amycolatopsis* strain in the present dataset with a very similar BGC (80-90% amino acid identity for the core genes) that contains a homologue of the desired P450 gene (adjacent to the *Amycolatopsis*/P450 clade). Analysis of MS/MS data from an *Amycolatopsis jejuensis* NRRL B-24427 fermentation extract revealed detoxin isomers P1 (**5**) (see Supplementary Figure 13 and 21), containing a tyrosine, P2 (**6**) (see Supplementary Figure 13 and 22), featuring phenylalanine and a hydroxylated valine, as well as detoxin P3, a closely related analogue free of hydroxylation (**7**) (see Supplementary Figure 13 and 23). Only five of the seven new detoxins described in the present study appear as nodes in the molecular network of Figure 6.3c, with the notable absence of two P-series analogues. This is because detoxin isomers P1 and P2 had a cosine similarity >0.6 and were collapsed into one node, whereas detoxin P3 was identified in fermentations following those that were a part of the original MS dataset. As before, validation of amino acid assignments, observed in MS/MS fragmentation data for detoxins P1-P3, was achieved through several metabolic feeding experiments using stable isotope-labeled amino acids (see Supplementary Figure 24-33). Detailed structural analysis for compounds **1**-**7**, including results from feeding studies using stable isotope-labeled amino acids, deconvolution of MS/MS spectra and full 1H, $_{13}$C and 2D NMR assignments for compound **2** are available in Supplementary Note 3, Supplementary Figure 15-33 and **subchapter 6.4**. Previously reported detoxins and rimosamides antagonize blasticidin-S inhibition of *Bacillus cereus*, a bioactivity that will be investigated for these analogues in follow-up studies [36,42].

The results of the present study illustrate how BiG-SCAPE can effectively identify sets of related BGCs across large numbers of genome sequences. Moreover, the use of CORASON to visualize the evolutionary diversity of gene clusters proved powerful for the discovery of new BGC clades encoding uncharted natural product chemistry. When focused toward detoxin/rimosamide discovery in "query mode", CORASON exhibited a unique ability to aid mining of a large genomic library for the discovery of seven new detoxins. Specifically, organization of gene content variation across BGCs facilitated the identification of corresponding variation in chemical structure.

6.3. Discussion

The comprehensive computational workflow introduced in the present study enables effective exploration of biosynthetic diversity across large strain collections, pan-genomes of entire bacterial or fungal genera and metagenomic datasets with thousands of metagenome-assembled genomes. The BiG-SCAPE/CORASON platform overcomes computational bottlenecks in previous approaches by enabling the assignment of GCFs with both partial and complete BGCs, accounting for class-specific differences between BGCs, incorporating sequence identity information within limited computing time and determining evolutionary relationships between and within GCFs. In addition, an interactive and intuitive user interface enables comprehensive investigation of these advanced outputs. Hence, it is anticipated that the BiG-SCAPE/CORASON platform will enhance the correlation of BGCs to metabolites, enabling metabologenomics studies at unprecedented scales.

Furthermore, the ability to perform phylogenetic analyses of large sets of complete BGCs, as well as their individual genetic components, a long-standing challenge that has remained unsolved since first posed in 2008 [43], will constitute a key technology to facilitate fundamental studies on the evolutionary origins of natural product chemical innovations. For example, phylogenies provide a stepping stone to perform detailed analyses of how gene cluster architectures evolve from their constituent independent enzymes and subclusters. A logical next step will be the unified classification of the millions of BGCs within publicly available genome sequences, and a Pfam-like database for the assignment of biosynthetic GCFs to known and unknown areas of natural product chemical diversity.

6.4. Methods

6.4.1. Dataset

A set of 2,831 actinobacterial genomes was downloaded from the National Center for Biotechnology Information (NCBI) by querying for “Whole genome shotgun sequencing project” or “Complete genome” in combination with the taxonomic identifier for Actinobacteria. The orders Propionibacteriales, Micrococcales, Corynebacteriales and Bifidobacteriales were excluded, because they contain large numbers of genomes without relevant natural product-producing capacity, except the Nocardiaceae family from Corynebacteriales (see below). To these, 249 additional draft assemblies from the Metcalf lab were added (for example, *Streptomyces* sp. B-1348; see BioProject PRJNA488366). Draft genome assemblies from this BioProject were obtained using SPAdes [44] with default options.

All files were processed using antiSMASH v.4 [14] (parameters: `--minimal`). The antiSMASH-annotated genome sequences are available from [45] To the resulting 73,260 predicted BGCs, 1,393 more were added from the MIBiG (release 1.3, August 2016, antiSMASH-analyzed versions from each entry) as reference data.

This final BGC set was then analyzed using BiG-SCAPE v.31 of the Pfam database. The “hybrids” mode, which allows BGCs with mixed annotations to be analyzed in their individual class sets (for example, a BGC annotated as lanthipeptide-t1pks will be analyzed as both an RiPP and a PKSI) was enabled. Two results sets were created (BiG-SCAPE results network files are available from [45]): one with the global mode enabled and the other with glocal mode enabled (see Figure 6.2).

6.4.2. Actinobacteria genome set

The extended set of genomes selected to be processed by antiSMASH and BiG-SCAPE was obtained by using the following query in the NCBI website on 30 January 2018 (2,891 results):

```
(“whole genome shotgun sequencing project”[title] OR “complete
genome”[title]) AND (Actinobacteria[Organism] NOT
(Propionibacteriales[Organism] OR Micrococcales[Organism] OR
Corynebacteriales[Organism] OR Bifidobacteriales[Organism]) OR
Nocardiaceae[Organism]) AND (bacteria[filter] AND biomol_genomic[PROP]
AND ddbj_embl_genbank[filter]) NOT (scaffold[title] OR plasmid[title]
OR segment[title])
```

The CORASON and EvoMining results used the same unpublished draft genomes but a reduced set of 1,668 actinobacterial genomes from an earlier query on the NCBI website, obtained on 3 February 2017, with the following query in the NCBI website (1,668 results):

```
(“whole genome shotgun sequencing project”[title] OR “complete
genome”[title]) AND (Actinobacteria[Organism] NOT
(Propionibacteriales[Organism] OR Micrococcales[Organism] OR
Corynebacteriales[Organism] OR Bifidobacteriales[Organism]) OR
Nocardiaceae[Organism]) AND (bacteria[filter] AND biomol_genomic[PROP]
AND ddbj_embl_genbank[filter]) NOT scaffold[title]
```

6.4.3. BiG-SCAPE algorithm

Alignment method comparison. To compare alignment methods for domain sequences, the regular version of BiG-SCAPE was used against a custom-prepared version of the same snapshot using MUSCLE 3.8.1551-h6bb024c_4 [46] (parameters: `–maxiters 2`) and MAFFT v.7.407 [47] (parameters: `--auto`); the three versions of the code are available from (Alignment Method Comparison). MUSCLE was parallelized using Python’s pool.map on single-core instances for each domain sequence fasta file, whereas MAFFT was parallelized on each file with its `--thread` parameter. Comparison of the final GCF calling (using BiG-SCAPE’s `--mix` parameter) indicates high agreement across the three methods (see Supplementary Figure 34), with hmmalign showing shorter runtimes as the number of BGCs in the input data increases (see Supplementary Figure 35).

Clustering algorithm optimization. The selection of the clustering algorithm was based on an initial analysis of the BGCs from the MIBiG database using BiG-SCAPE (`--hybrid` mode disabled). In this analysis, the network went through a targeted attack first to identify the most suitable cutoff for clustering algorithm evaluation. The targeted attack removes the edges above a certain cutoff value while calculating, for each iteration, the number of nodes and graph density, and identifying the connected components after removal of isolated vertices (BGCs). Network statistics such as the number of vertices/edges lost for each cutoff value, as well as the size of the connected components that emerged, were calculated during the attack.

Supplementary Figure 36 shows the dynamics and impact of the different filtering thresholds applied to the different BGC training networks, with a cutoff of 0.75 being the value that maximized the number of nodes, while minimizing the impact on the structural integrity of the network. This analysis was performed using the igraph package [48] for the network analyses and ggplot2 [49] for plotting.

Next, entropy was calculated on MIBiG networks for several clustering algorithms (see Supplementary Table 5) based on the selected cutoff of 0.75 in combination with the curated compound data (see Supplementary Dataset). Supplementary Figure 37 and 38 show the results of applying the different clustering methods to the different training networks (glocal and global), with the affinity propagation clustering method showing the most sensible results, producing clusters with low entropy and average size. All the other methods tested resulted in clusters present in the principal quadrant, indicating that these methods could not partition the data properly and lumped together vertices (large size) that encode different types of compounds (large entropy). Based on these results, affinity propagation was chosen as the clustering algorithm in BiG-SCAPE.

Input data. The input BiG-SCAPE consists of text files in GenBank format (.gbk extension) and the Pfam database [32] (already processed with hmmpress). Although BiG-SCAPE can work with files not processed by antiSMASH, it relies on antiSMASH's product prediction to separate the BGCs in their correct biosynthetic class, thereby reducing computational time. If the product annotation is unknown, missing, or several different classes are mixed, the BGC will be classified as "Other".

Algorithm overview. After selecting and filtering (for example, for a certain size, in base pairs) the input GenBank files, protein sequences are extracted. All the sequences from each file are searched for conserved domains using a user-supplied external Pfam database. Overlapping domains are filtered based on the score calculated by hmmer. The sequences of every predicted domain type are aligned using each corresponding model by hmmalign. A distance matrix is created by calculating the distance between every pair of BGC in the dataset (see overview of the algorithm in Supplementary Figure 39). For this study, v.31 of the Pfam database was used with HMMER v.3.1b2.

Distance calculation. Pairwise distance calculation is divided between three values that measure 1- the percentage of shared domain types (JI), 2- the similarity between aligned domain sequences (DSS index; domains from the same type are first matched for best similarity using Munkres algorithm, as implemented in the Scikit-Learn library [50]) and 3- the similarity of domain pair types (AI). For specific details of each index, see Supplementary Note 1.

There are two ways of selecting the domains predicted within each BGC for the calculation of distance. In the global mode, all domains are considered. For cases where the difference in size is large (due to, for example, one BGC being placed at the edge of a contig or when comparing curated BGCs with shorter gene borders), the so-called glocal mode was implemented, in which a selection of domains is used in the distance calculation. In this mode, genes in each BGC are represented as a concatenated string of Pfam domains, and each BGC in the pair is represented as a list of those domain concatenations (strandedness is not considered).

BiG-SCAPE then uses the SequenceMatcher method from Python's difflib library to find the longest match (internally called the LCS or longest common subcluster) in either orientation (including the reverse complement of the subject BGC).

To proceed to the next step, the LCS must be either three genes long or contain at least one gene marked by antiSMASH as core biosynthetic (that is, genes that encode the first step in the assembly of the metabolite's scaffold and that are used by antiSMASH as a first step in defining a biosynthetic cluster, for example, PKSs or NRPSs).

In the extension stage, the selection of domains is extended for the BGC with the least number of genes up- or downstream (up until the end of the BGC or a contig break in the genome assembly). The remaining BGC domain selection (per side, that is, both left and right) will be subjected to expansion according to the following scoring algorithm in the alignment stage: for every gene in the reference BGC, a gene with the same domain organization is searched for in the remaining BGC. If such a gene is found, the score will be added as a bonus (match=5) plus a penalty proportional to the distance from the current position (number of genes x gap penalty, where gap=−2), and the current position will be moved to the position of the matching gene. If a gene with the same domain organization is not found, the score will be decreased with a penalty (mismatch=−3). In the end, the highest-scoring extension is chosen to form the "matching" BGC region on which the similarity will be calculated.

GCF clustering. Once the distance matrix has been calculated for each BiG-SCAPE class (see Supplementary Table 1), GCF assignment is performed for every cutoff distance selected by the user (the interactive visualization of BiG-SCAPE will show the one with the largest number), with 0.3 being the default. For every cutoff, BiG-SCAPE creates a network using all distances lower than or equal to the current cutoff. The affinity propagation clustering algorithm [34] is applied to each subnetwork of connected components that emerges from this procedure. The similarity matrix for affinity propagation includes all distances between members of the subnetwork (that is, it includes those with a distance greater than the current cutoff).

GCC setting (enabled by default) will perform a second layer of clustering on the GCFs. For this, affinity propagation will be applied again, but network nodes are represented by the GCFs, defined at the cutoff level specified in the first value of the `--clan_cutoff` parameter (default 0.3). Clustering will be applied to the network of all GCFs connected by a distance lower than or equal to the GCC cutoff (second value of the `--clan_cutoff` parameter; larger distances are discarded; default 0.7). Inter-GCF distance is calculated as an average distance between the BGCs within both families. Affinity propagation parameters used are the following: damping=0.9, max_iter=1000, convergence_iter=200.

Interactive output. BiG-SCAPE's interactive output is written as a collection of plain HTML and JavaScript programs that are being generated for every run along with the corresponding data stored in JSON format. The VivaGraphJS (https://github.com/anvaka/VivaGraphJS) library is used as the basis for the animated network visualization (the drag-and-drop and zoom function). FuseJS (https://fusejs.io/) is used to implement the "fuzzy" search query function, meaning that users can type any related keywords (either Pfam domain accession, genome name, or compound name) in one search form and the algorithm will attempt to find any BGC entry related to the keywords with tolerance for small typos. A custom script built upon the SVGJS library (https://svgjs.com/docs/3.0/) is used to draw SVG-based arrower visualization, which is publicly available at https://github.com/satriaphd/arrower-js. ChartJS (https://www.chartjs.org/) is used to draw the interactive pie chart on the overview screen, whereas InchLib (https://www.openscreen.cz/software/inchlib/home/) is used to draw the GCF absence/presence heatmap in combination with the clusterfck (https://harthur.github.io/clusterfck/) library, which allows real-time hierarchical clustering for the heatmap's dendrograms. TreeLib (https://treelib.readthedocs.io/en/latest/) is used to provide the dendrogram-like phylogenetic tree of BGCs in the GCF view, and the KineticJS (https://github.com/ericdrowell/KineticJS) library is used to combine the dendrogram with the BGC arrower visualization. The complete source code for BiG-SCAPE's html visualization module can be downloaded from GitLab (https://git.wageningenur.nl/medema-group/BiG-SCAPE/-/tree/master/html_template/output/html_content).

CORASON family mode. As part of BiG-SCAPE's visual output, a CORASON-like tree is generated for every GCF page. This tree is created using the sequences of the Core Domains in the GCF. These are defined as the domain type(s) that 1- appear with the highest frequency in the GCF and 2- are detected in the central (or exemplar) cluster, defined by the affinity propagation cluster. All copies of the Core Domains in the exemplar are concatenated, as well as those from the best matching domains of the rest of the BGCs in the GCF (aligned domain sequences are used). The tree is constructed using FastTree [38] (default parameters). Visual alignment is attempted using the position of the "longest common information" from the distance calculation step (between the exemplar BGCs and each of the other clusters).

Weight optimization methods. Tuning of weights for each BiG-SCAPE class was calculated by a brute-force approach, choosing the weight combination that maximized the correlation between BGC and compound distances for every pair of BGCs in the same class in a manually curated compound group table (see Supplementary Dataset). The dataset comprised all BGCs from the MIBiG database (v1.3) that had linked compound SMILES and at least two predicted domains to filter out minimal gene cluster entries. BGC distances were calculated by moving in steps of 0.01 across the JI, DSS and original Goodman-Kruskal (GK) [51] indices and AIs, such that JI+DSS+GK+AI=1. The anchor boost parameter of DSS was allowed to change in the range one to four with steps of 0.5. For the DSS index, only the original four anchor domains were considered (condensation domain, PF00668; beta-ketoacyl synthase N-terminal domain, PF00109; beta-ketoacyl synthase C-terminal domain, PF02801 and the terpene synthase N-terminal domain, PF01397). Compound distances were calculated only once, between all BGCs in the MIBiG v1.3 that had an annotated SMILES string representing the molecule. Their pairwise distance was calculated using RDKit (Tanimoto's coefficient based on Morgan's fingerprinting, radius=4). The nine original curated compound classes were used to tune the weights of seven BiG-SCAPE classes (the terpene BiG-SCAPE class was initially included in the "others" compound class due to a low number of points and was assigned default weights: JI=0.2, DSS=0.75 and AI=0.05).

Results (see Supplementary Figure 40) indicated clear tendencies to favor different indices in each case and corroborated that the proposed AI was more informative than the original Goodman-Kruskal synteny metric used in Cimermancic et al. [3], which led to the decision to drop this index from the final distance formula (additional details in Supplementary Note 2 and Supplementary Figure 41).

Comparison with other methods. To compare BiG-SCAPE with the GCF algorithm in Doroghazi et al. [4], 11,618 GenBank files were reconstructed from data related to that study (allClusterProts.fasta file from https://www.igb.illinois.edu/labs/metcalf/gcf/search.html). These reconstructed cluster files were analyzed using antiSMASH v4, and its output was used to make a run in BiG-SCAPE. Unlike Doroghazi's method, BiG-SCAPE follows a two-step process to infer GCFs. First, it filters the resulting network using a predefined empirical cutoff distance of 0.3, and later the GCFs are identified by the affinity propagation clustering algorithm. This two-step approach partitions the natural emerging components from the filtering step, increasing the resolution of the inferred GCFs. To provide a fair comparison with GCFs inferred by that method, the natural emerged components were used after the filtering steps and the different clustering results were compared; good agreements were found between both methods (see Supplementary Figure 42 for details), although BiG-SCAPE took only a fraction of the runtime of the previously published tool (see Supplementary Table 6).

6.4.4. CORASON algorithm

CORASON inputs are a custom genomic database, a reference cluster and a query gene located within the reference cluster. The genomic database is a collection of either genomes or BGCs in GenBank format. CORASON will identify the conserved core of the reference BGC within the genomic database. Best bidirectional hits are pairs of genes that exist in two different sets of genes (genomes, metagenomes or BGCs) that are more similar to each other than with any other sequence in the set pair. In CORASON, this relationship was generalized in a stricter algorithm that considers all-versus-all comparisons between every set in the collection to remove paralogues and conserve only true orthologues. As a result, the conserved core is composed of gene families that are each guaranteed to be a best bidirectional hit across the whole collection (although they need not be contiguous). The BGC-conserved core facilitates reconstruction of the BGC evolutionary history in a multi-locus tree. The query gene assures that at least one element will be present in the conserved core and will also be used to visually align the BGC variations in the graphic output.

Identification of reference BGC variations on the genomic DB. CORASON uses BlastP, with an e-value cutoff of 0.001, to find all query gene homologues within the genomic database. The genomic contexts of the query gene homologues are expanded to ten genes on each side and stored in a temporary database. Next, protein sequences from the reference BGC, located within fewer than n genes (default: n=10) from the query gene, are blasted against the temporary database using the same e-value cutoff. Genomic context size, e-value and bit score cutoffs are user-adjustable parameters. Finally, all genomic contexts with at least two homologues (by default), including the query gene and at least one additional from the reference cluster, are kept as the cluster variation database (CVD) for further analysis.

Gene core determination. To reconstruct the phylogeny of the BGC variations, the conserved core is calculated. The core is strongly dependent on the taxonomic diversity of the organisms considered and also on the genome quality. For instance, if the BGCs are not closely related, the core may be reduced to only the query gene. A set of homologous genes is considered part of the conserved core if, and only if, they are shared among the cluster variations internal database (all BGCs) and are multidirectional best hits, that is, if they are best n-directional hits in an all-versus-all manner. Formally, for H defined as:

$$H=\{h_i \mid h_i \in \mathrm{BGC}_i \ \forall_i \in \{1,2,...,N\}\}$$

where every homologous gene h_i belongs to a set of N BGC variations, H belongs to the conserved core if, and only if,

$$h_i \text{ is } h_j \text{ best bidirectional hit } \forall_{i,j} \in \{1,2,\ldots,N\}$$

that is, when every pair of homologous genes h_i and hj within H are best bidirectional hits.

Phylogenetic reconstruction and gene cluster alignment. For each BGC, its conserved core sequences are concatenated and then aligned using MUSCLE v3.8.31. The alignments are curated using Gblocks [52] with a minimum block length of five positions, a maximum of ten contiguous nonconserved positions and considering only positions with a gap in less than 50% of the sequences in the final alignment. If the curation turns out to be empty, then the non curated alignment will be used for the tree. If the alignment itself is empty, it is recommended to reduce the score cutoff or the scope of the taxonomic diversity on the genomic database. Without the alignment, BGCs will be drawn but not sorted. Approximately maximum-likelihood phylogenetic trees are inferred using FastTree43 v.2.1.10 from the curated amino acid alignment.

BGC prioritization graphic output. CORASON produces an SVG file containing the BGC variations sorted as stated by the phylogenetic reconstruction and aligned according to the query enzyme. The Newick tree is converted to SVG by applying Newick Utilities v.1.6 [53] and each BGC is drawn with the Perl module SVG. As an additional feature to facilitate even more visual differentiation of BGC families within BGC clans, genes on each cluster are visually represented with a color gradient according to the sequence similarity to their homologous gene on the reference cluster. Other CORASON outputs include the Newick tree, the GenBank files of the BGC variations and the conserved core report.

6.4.5. Streptomyces closed genome analysis

Sequences from 103 complete *Streptomyces* genomes were retrieved from the NCBI by querying for "*Streptomyces*" and "complete genome" not "segment". Two genomes corresponding to *C. acidiphila* and *S. arenicola* (CP001700 and CP000850) were used as outgroups. These genomes were analyzed by antiSMASH v4 and the resulting gene cluster files were used as input for BiG-SCAPE. The conserved core was extracted and curated using the CORASON algorithm. The tree was constructed using FastTree with default values over a matrix of 114,051 amino acids in size, from 446 conserved gene families. The interactive report of BiG-SCAPE reports only 96 genomes, because genomic scaffolds that belong to the same genome are grouped by ORGANISM identifier, and several strains have more than one assembly project associated in NCBI with the same ORGANISM identifier.

6.4.6. Phylogenomic analysis

For the TauD expansions tree (see Supplementary Figure 8), a tauD sequence from *Escherichia coli* K12 was used as query to conduct a blast search against the reduced genomic database of 1,917 Actinobacteria genomes (e-value 0.001), followed by an EvoMining analysis and a search for recruitments on MIBiG database (e-value 0.001). Recovered tauD orthologues were aligned with MUSCLE v.3.8.31 and alignments were curated using Gblocks as described before. An unrooted approximately maximum-likelihood tree was built using FastTree. The tree was colored using Newick Utilities according to BiG-SCAPE families. The CORASON tree has as its query gene tauD from the reference cluster of the organism *Streptomyces* NRRL B-1347 (JOJM01). CORASON trees are unrooted, but this tree was posteriorly rooted with the BGC from the genome *Streptomyces sp.* NC1, because this BGC is different from all other clusters in the dimeric peptide clan; it does not share the core but only the accessory enzyme-coding genes with other BGC clan members.

6.4.7. Molecular networking methods

Cultivation of actinomycetes for MS-based metabolomics. All strains analyzed for metabolomics were grown on four media types: arginine/glycerol/salts, mannitol/soy flour, ISP medium 4 or glycerol/sucrose/beef extract/casamino acids as previously reported [4]. After ten days of growth, plates were frozen, then thawed and pressed to release spent liquid media. Media were then filtered and extracted using 30 mg Supel-Select HLB SPE cartridges (Supelco) and resuspended to a concentration of approximately 2 mg ml^{-1} in 5% acetonitrile before LC-MS analysis.

Acquisition and analysis of LC-MS metabolomics data. All LC-MS/MS analyses were performed using an Agilent 1150 HPLC coupled with a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Reversed-phase chromatography was performed at a 200 µl min^{-1} flow rate on a Phenomenex Kinetex C18 RP-HPLC column (150x2.1 mm^2 inner diameter, 2-µm particle size, 100Å pore size (1Å=0.1 nm). Mobile phase A was water with 0.1% formic acid and mobile phase B was acetonitrile with 0.1% formic acid. Mass spectral data for both MS and MS/MS were acquired using a 250-3,750 m/z scan range, a resolution of 35,000, a maximum inject time of 40 ms and an AGC target value of 1x106. The top five most intense ions in each full MS spectrum were targeted for fragmentation by higher-energy collisional dissociation at 25 eV. MS/MS data were analyzed using spectral networking as previously described [54]. Signals detected in multiple strains were determined to be the same ion if the observed accurate masses were within 4ppm and fragmentation cosine similarity scores were >0.75, yielding 5,824 ions detected in two or more strains.

LC–MS molecular networking. Molecular networking was performed as previously reported [54]. Briefly, individual MS/MS scans were extracted from each MS raw file and filtered to remove the 25% of ions with the lowest intensity. Each MS/MS scan was further processed by taking the square root of each ion's intensity and normalizing it so that the sum of all intensities in each MS/MS scan was equal to 1. Cosine similarities were calculated between all MS/MS scans, with scores ranging between 0 and 1, with a score of 1 indicating that two MS/MS scans were identical. Precursor ions were determined to be identical if they were within 0.01 m/z and their corresponding MS/MS spectra had a cosine similarity score >0.6. A visualization of the network was constructed in Cytoscape by drawing edges between scan nodes with a cosine similarity >0.6. The network was manually analyzed to identify ions related to known detoxins and rimosamides, which were found to cluster together as one molecular family, with 99 putatively new analogues, a subset of which were characterized herein as detoxins S1, N1-N3 and P1-P3.

6.4.8. Metabolic labeling of detoxins N1–N3 and P1–P3 with stable isotope-labeled amino acids

Streptomyces spectabilis Dietz NRRL-2792 (ATCC 27741) was obtained from the American Type Culture Collection (ATCC) and was grown on 60-mm solid agar medium Petri plates containing arginine/glycerol/salts medium (1l of distilled water, 15g of agar, 1g of arginine, 12. g of glycerol, 1g of potassium phosphate dibasic, 1g of sodium chloride, 0.5g of magnesium sulfate heptahydrate, 10mg of iron(II) sulfate hexahydrate, 1mg of copper(II) sulfate pentahydrate, 1mg of manganese(II) sulfate monohydrate and 1mg of zinc sulfate heptahydrate). *Amycolatopsis jejuensis* NRRL B-24427 was obtained from the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA) and was grown on solid agar medium Petri plates containing mannitol/soy flour medium (1l of distilled water, 15g of agar, 20g of D-mannitol and 20g of soy flour). For all metabolic labeling experiments, the medium was supplemented with 1ml of a 10mM solution of each stable isotope-labeled amino acid. Stable isotope-labeled amino acids used were [$^{13}C_6$]isoleucine, d_8-[^{15}N]phenylalanine, d_7-proline, 2,5,5-d_3-proline, phenyl-d_4-tyrosine, d_8-valine, 2-d_1-valine and 3-d_1-valine. After five days of incubation in the presence of stable isotope-labeled amino acids, plates were frozen overnight at −20°C, thawed and pressed to release spent liquid medium. Extracellular secondary metabolites were extracted using 30mg Supel-Select HLB SPE cartridges (Supelco) and eluted with 90% acetonitrile. Samples were dried, resuspended in 5% acetonitrile and analyzed by reversed-phase LC-MS/MS on a Q-Exactive mass spectrometer as described above. The methods used for LC-MS data acquisition on the Q-Exactive were the same, except for occasional parameter adjustments made to target major unnatural isotope ions for optimal fragmentation.

Acquisition of NMR data. All NMR experiments were performed in ^{2}H2O. ^{1}H, ^{13}C, correlation spectroscopy, heteronuclear single quantum coherence, heteronuclear multiple bond correlation and nuclear Overhauser enhancement spectroscopy spectra were obtained on a Bruker NEO spectrometer (600MHz for 1H, 150MHz for 13C) with a QCI-F cryoprobe. The 1H-1H TOCSY spectrum was obtained on a Bruker Avance III 500MHz spectrometer (500MHz for ^{1}H) equipped with a DCH CryoProbe. Chemical shifts (δ) are given in ppm and coupling constants (J) are reported in Hz. ^{1}H and ^{13}C chemical shifts were referenced to sodium formate (δH 8.44; δC 171.67). ^{1}H and ^{13}C NMR resonances of compound **2** are reported in Supplementary Note 3i.

Metabologenomic correlations. Strains with metabolomics data were referenced against the BiG-SCAPE GCF absence/presence matrices. GCFs that had representative gene clusters in two or more strains were considered correlatable and entered into the correlations dataset. The different BiG-SCAPE modes and cutoffs produced variable numbers of correlatable GCFs and thus different numbers of ion-GCF hypotheses (see Supplementary Table 2). Supplementary Figure 43 shows the full version of Figure 6.3d.

Acknowledgments

We thank the following: the ARS of the USDA for providing bacterial strains; H. Sook Ann, Z. Crispino, Y. Kim, N. Ciszek and K. Espejo for generating bacterial culture extracts; R. McClure, M. Robey and G. Miley for assistance with and contributions to metabolomic data collection methods and acquisition; and Dr. Y. Zhang and Dr. Y. Wu of the Integrated Molecular Structure Education and

Research Center (IMSERC) at Northwestern University for assistance in acquiring NMR data. Some analyses were carried out using CONABIO's computing cluster, with funds from the Secretariat of Environment and Natural Resources. We thank K. Blin for technical assistance with setting up the website on the secondarymetabolites.org domain.

Funding

The research reported in this publication was supported by the Netherlands Organization for Scientific Research (grant no. 863.15.002 to M.H.M.), the Graduate School for Experimental Plant Sciences (grant to M.H.M.); National Institutes of Health (NIH) Genome to Natural Products Network supplementary award (no. U01GM110706 to M.H.M.), CONACyT grants (grant nos. CBS2017_285746 and 2017_051TAMU to F.B.-G.; postdoctoral scholarship 263661 to J.C.N.M.; PhD scholarship 204482 to N.S.M. (who was also supported by the Innovation Secretary of Guanajuato)), the National Cancer Institute of the NIH (award no. F32CA221327 to M.W.M.), the National Institute of General Medical Sciences (award no. F32GM120999 to E.I.P.), the São Paulo Research Foundation (FAPESP, grant no. 17/08038-8 to L.T.D.C.), the National Center for Complementary and Integrative Health of the NIH (award no. R01AT009143 to R.J.T. and N.L.K.) and Warwick Integrative Synthetic Biology Centre, a UK Synthetic Biology Research grant from the Biotechnology and Biological Sciences Research Council and Engineering and Physical Sciences Research Council (grant no. BB/M017982/1 to E.L.C.D.L.S). This work made use of the IMSERC at Northwestern University, which has received support from the NIH (grant nos. 1S10OD012016-01/1S10RR019071-01A1), the State of Illinois and the International Institute for Nanotechnology. A.F.-G. received funding from the European Union's Horizon 2020 research and innovation program (Blue Growth: Unlocking the Potential of Seas and Oceans; grant agreement no. 634486).

Conflicts of Interest Statement

M.H.M. is on the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals. N.L.K., W.W.M. and R.J.T. are on the board of directors of MicroMGx, and A.W.G. is chief scientific officer at MicroMGx.

Data Availability

Genomes used in this study include assemblies from the sequencing project deposited in NCBI BioProject PRJNA488366, in Sequence Read Archive runs with accession numbers SRX4638772 to SRX4639021. AntiSMASH, BiG-SCAPE and CORASON results for all genome assemblies, along with raw files of phylogenetic trees, are available from [45]. Fully annotated nucleotide sequences for the BGCs for detoxin S1, detoxins N2-N3 and detoxins P1-P3 have been deposited in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under accession numbers BK010707, BK010852 and BK010851, respectively, and in MIBiG under accession numbers BGC0001840, BGC0001878 and BGC0001841, respectively. All raw MS data files for strains producing one or more of the nine compounds used for correlation analysis have been submitted to MassIVE under accession number MSV000083738. Raw MS data files and isolated MS/MS scan files for all newly identified detoxin analogues have been uploaded to MassIVE with accession number MSV000083648, and MS/MS data for other strains are available upon request.

Code Availability

An overview of both BiG-SCAPE and CORASON can be found at https://bigscape-corason.secondarymetabolites.org, BiG-SCAPE project at https://git.wur.nl/medema-group/BiG-SCAPE and CORASON project at https://github.com/nselem/corason.

Supplementary Material

All supplementary information can be found online at https://go.nature.com/3os5jbl.

References

1. Traxler MF, Kolter R. Natural products in soil microbe interactions and evolution. Nat Prod Rep. 2015;32: 956–970.
2. Davies J. Specialized microbial metabolites: functions and origins. J Antibiot . 2013;66: 361–364.
3. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
4. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
5. Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol. 2016;12: 1007–1014.
6. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A. 2017;114: 5601–5606.
7. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biol Evol. 2016;8: 1906–1916.
8. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics. 2017;33: 3202–3210.
9. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nat Chem Biol. 2015;11: 639–648.
10. Katz L, Baltz RH. Natural product discovery: past, present, and future. J Ind Microbiol Biotechnol. 2016;43: 155–176.
11. Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature. 2002;417: 141–147.
12. Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO, et al. Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol. 2007;25: 1281–1289.
13. Bergmann S, Schümann J, Scherlach K, Lange C, Brakhage AA, Hertweck C. Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. Nat Chem Biol. 2007;3: 213–217.
14. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45: W36–W41.
15. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of

natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017;45: W49–W54.

16. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11: 625–631.
17. Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. Nat Microbiol. 2017;2: 17044.
18. Tobias NJ, Wolff H, Djahanschiri B, Grundmann F, Kronenwerth M, Shi Y-M, et al. Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. Nat Microbiol. 2017;2: 1676–1685.
19. Grubbs KJ, Bleich RM, Santa Maria KC, Allen SE, Farag S, AgBiome Team, et al. Large-Scale Bioinformatics Analysis of Bacillus Genomes Uncovers Conserved Roles of Natural Products in Bacterial Physiology. mSystems. 2017;2. doi:10.1128/mSystems.00040-17
20. Freeman MF, Gurgui C, Helf MJ, Morinaka BI, Uria AR, Oldham NJ, et al. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science. 2012;338: 387–390.
21. Agarwal V, Blanton JM, Podell S, Taton A, Schorn MA, Busch J, et al. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. Nat Chem Biol. 2017;13: 537–543.
22. Owen JG, Charlop-Powers Z, Smith AG, Ternei MA, Calle PY, Reddy BVB, et al. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. Proc Natl Acad Sci U S A. 2015;112: 4221–4226.
23. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2: 1533–1542.
24. Leao T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. Proc Natl Acad Sci U S A. 2017;114: 3198–3203.
25. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc Natl Acad Sci U S A. 2014;111: E1130–9.
26. Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comput Biol. 2014;10: e1003822.
27. Mohimani H, Liu W-T, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA. NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. J Nat Prod. 2014;77: 1902–1909.
28. Mohimani H, Kersten RD, Liu W-T, Wang M, Purvine SO, Wu S, et al. Automated genome mining of ribosomal peptide natural products. ACS Chem Biol. 2014;9: 1545–1551.
29. Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, Zhao X, et al. MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci U S A. 2013;110: E2611–20.
30. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. ACS Cent Sci. 2016;2: 99–108.
31. Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol. 2015;22: 460–471.
32. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40: D290–301.
33. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput Biol. 2014;10: e1004016.
34. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315: 972–976.
35. Parkinson EI, Tryon JH, Goering AW, Ju K-S, McClure RA, Kemball JD, et al. Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. ACS Chem Biol. 2018;13: 1029–1037.
36. McClure RA, Goering AW, Ju K-S, Baccile JA, Schroeder FC, Metcalf WW, et al. Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using

Metabolite/Gene Cluster Correlations. ACS Chem Biol. 2016;11: 3452–3460.

37. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A. 2012;109: E1743–52.

38. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26: 1641–1650.

39. Hausinger RP. FeII/alpha-ketoglutarate-dependent hydroxylases and related enzymes. Crit Rev Biochem Mol Biol. 2004;39: 21–68.

40. Kim K-R, Kim T-J, Suh J-W. The gene cluster for spectinomycin biosynthesis and the aminoglycoside-resistance function of spcM in Streptomyces spectabilis. Curr Microbiol. 2008;57: 371–374.

41. Sinha A, Phillips-Salemka S, Niraula T-A, Short KA, Niraula NP. The complete genomic sequence of Streptomyces spectabilis NRRL-2792 and identification of secondary metabolite biosynthetic gene clusters. J Ind Microbiol Biotechnol. 2019;46: 1217–1223.

42. Ogita T, Seto H, Otake N, Yonehara H. Studies on Detoxin complex, the selective antagonists of Blasticidin S. The structures of minor congeners of the detoxin complex. Agricultural and Biological Chemistry. 1981. pp. 2605–2611. doi:10.1271/bbb1961.45.2605

43. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci U S A. 2008;105: 4601–4608.

44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–477.

45. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. Genomic data for “A computational framework to explore large-scale biosynthetic diversity.” 2018. doi:10.5281/zenodo.1532752

46. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

47. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.

48. Csardi G, Nepusz T, Others. The igraph software package for complex network research. InterJournal, complex systems. 2006;1695: 1–9.

49. Wickham H, Chang W, Others. ggplot2: An implementation of the Grammar of Graphics. R package version 0 7, URL: http://CRAN R-project org/package= ggplot2. 2008;3.

50. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12: 2825–2830.

51. Lin K, Zhu L, Zhang D-Y. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics. 2006;22: 2081–2086.

52. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17: 540–552.

53. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26: 1669–1670.

54. Henke MT, Soukup AA, Goering AW, McClure RA, Thomson RJ, Keller NP, et al. New Aspercryptins, Lipopeptide Natural Products, Revealed by HDAC Inhibition in Aspergillus nidulans. ACS Chem Biol. 2016;11: 2117–2123.

Chapter 7

BiG-SLiCE: A Highly-scalable Genomic Tool Charts a “Global” Map of Natural Product Diversity

The wealth of genomic data collected over the years provides an invaluable resource for natural product discovery. As many microbial secondary metabolic pathways rarely produce (enough) compounds to facilitate analytical studies, genome-based analysis of their BGCs can serve as a proxy to chart an organism’s true biosynthetic potential. In this chapter, I introduce BiG-SLiCE, a new bioinformatics tool that allowed a simultaneous homology analysis of over 1.2 million BGCs predicted from more than 209 thousand microbial genomes available in public databases. From this analysis, I drew a “global” map of natural product diversity, revealing the breadth of unexplored chemical space in the microbial world. With its ability to perform a rapid calculation of BGC families, BiG-SLiCE can be used to find distant relatives of known BGCs and to answer strategic questions such as: “which taxa harbor the greatest potential for novel discovery?”

Published as:

Kautsar, S. A., Van Der Hooft, J. J., de Ridder, D., & Medema, M. H. (2021).
BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million
biosynthetic gene clusters. *GigaScience*, 10(1), giaa154.

7.1. Introduction

The microbial world is teeming with diverse microorganisms competing and collaborating for survival. A major theme in these microbial interactions is the use of bioactive compounds from secondary metabolism. Some of these compounds have long been exploited by humans for their medicinal, antifungal, and antibacterial effects [1]. Some others found their use in agriculture [2], wastewater treatment [3], and everyday products such as detergents and cleaning products [4]. A recent report by the World Health Organization highlights the need to explore novel chemistry from nature amid the increasing problems caused by antimicrobial-resistant (AMR) bacteria [5]. It was previously estimated that there might be billions of microbial species living on earth [6,7] and even from the heavily mined genus of *Streptomyces*, novel discoveries continue to be made [8–13]. Tapping into this vast space of natural product diversity will increase chances to achieve future medicinal breakthroughs. More fundamentally, by learning about microbes and the compounds they produce, we can gain knowledge about mechanisms of interaction within microbiomes, enabling us to study how their microbial composition is associated with human health and disease [14] or to learn about the symbiotic relationships between soil microbes and their plant host [15].

One promising way to reveal this knowledge is to leverage the power of large-scale omics. Metabolomics provides a complete snapshot of metabolites produced by microbes at a given time, while transcriptomics and proteomics provide insight into metabolic pathways and their regulation [16–18]. On the other hand, genomics allows the rapid profiling of an organism's metabolic potential via the computational prediction of biosynthetic gene clusters (BGCs) [19–21]. Previous studies [22–29] show that grouping BGCs with similar architecture (i.e. sharing a similar set of homologous core genes) into gene cluster families (GCFs) can yield useful insights into the chemical diversity of the analyzed strains, and can support linking BGCs to their products via the emerging technique of metabologenomics [23,27]. BGCs responsible for the production of retimycin A [29], tambromycin [27], tyrobetaines [30] and several detoxin-rimosamide analogs [22] have been elucidated via this approach. GCFs have also been used as functional markers in human health studies [31,32] and to study soil suppressiveness against fungal pathogens [33]. This gradual shift from a gene-centric approach in functional metagenomics to a gene cluster-centric one is likely to be stimulated further with the increasing accessibility of long sequencing reads that easily span tens to hundreds of kbp (kilobase pairs) in size [34], effectively covering the full span of a typical microbial BGC within a single read.

Given their direct relationship to the catalytic enzymes, and subsequently, the compounds produced from their encoded pathways, BGCs (and, by extension, GCFs) can serve as a proxy to explore the chemical space of microbial secondary metabolism. By cataloging all the GCFs in sequenced microbial genomes, one can obtain an overview of the existing chemical diversity and gain insights into what future lead discovery efforts should prioritize. For example, one could focus on species harboring the most potential novelty, or on identifying natural variants of a known antibiotic-producing BGC. For such global analyses, the clustering algorithm to group BGCs into GCFs needs to be able to work with massive volumes of data. While a trend of increasing input capacity can be observed for the past 5 years (from 11,000-33,000 analyzed BGCs in 2014 [23] to 73,260 in 2019 [22]), it is still dwarfed by the total amount of data currently available. As of 27 March 2020, antiSMASH-DB [35] and IMG-ABC [36], the two largest BGC databases, jointly comprise 565,096 BGCs predicted from 85,221 bacterial genomes. This number will increase even more if we account for genomes and metagenomes not covered by these databases. For example, assuming they hold similar average numbers of BGCs, the ~180,000 bacterial genomes in the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/) may yield more than a million BGCs when processed with tools like antiSMASH [19] (or other BGC prediction tools like PRISM [37], EvoMining [38], ClusterFinder [28] and DeepBGC [39]).

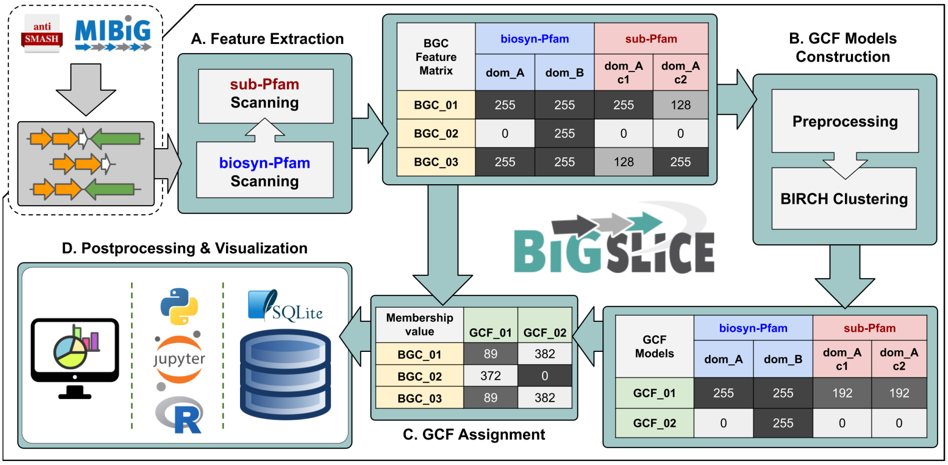
To handle a dataset this large, even the currently fastest tool (one tool we previously developed, BiG-SCAPE [22]) will require an estimated 37,000 hours of runtime on a 36-core CPU (see Results and Discussion), which is impractical if not impossible. A major bottleneck is the expensive pairwise BGC comparison used to construct similarity networks and perform clustering analysis, leading to quadratic time complexity ($O(n^2)$, where n is the total number of BGCs). Thus, there is an urgent need for an alternative method that better scales with the available genomic data, which will grow even further as the cost and performance of Next Generation Sequencing (NGS) technology continue to improve and get democratized [40]. Here, we introduce BiG-SLiCE (Biosynthetic Genes Super-Linear Clustering Engine), which projects BGCs into Euclidean space to enable the usage of a partitional clustering algorithm running in a near-linear (~O(n)) time complexity. Using this approach facilitates analysing large datasets of BGCs orders of magnitude faster, finally allowing truly global GCF analyses on all available microbial genomes.

7.2. Methods and Implementation

The BiG-SLiCE workflow starts at the vectorization (feature extraction) step (Figure 7.1A), converting input BGCs into vectors of numerical features based on the absence/presence and bitscores of hits obtained from querying BGC gene sequences against a library of curated profile hidden Markov models (pHMMs). Those features are then processed by a super-linear clustering algorithm (Figure 7.1B), resulting in a set of centroid feature vectors representing the GCF models. All BGCs in the dataset are finally queried back against those models (Figure 7.1C), outputting a list of GCF membership values for each BGC. In the end, an interactive visualization output is produced, which enables users to explore the analyzed data (Figure 7.1D).

7.2.1. BGC feature extraction

In BiG-SCAPE, the shared occurrence and synteny (order) of Pfam [42] domains is measured for each pair of BGCs, along with the sequence similarity of homologous core genes, in order to construct a pairwise-distance network and define GCFs in this network using the Affinity Propagation algorithm [43]. While this hierarchical approach enables a very sensitive measurement of the relationships between BGCs and provides networks that can be interactively explored, it leads to a quadratic runtime complexity that does not allow application beyond a few tens of thousands of BGCs. To enable more efficient calculation of GCFs via partitional, near-linear time complexity clustering algorithms such as K-means [44] or BIRCH [45], we need to transform BGCs into numerical feature vectors (commonly known as quantization or vectorization).



BGC Feature Matrix	biosyn-Pfam		sub-Pfam	
	dom_A	dom_B	dom_A c1	dom_A c2
BGC_01	255	255	255	128
BGC_02	0	255	0	0
BGC_03	255	255	128	255

GCF Models	biosyn-Pfam		sub-Pfam	
	dom_A	dom_B	dom_A c1	dom_A c2
GCF_01	255	255	192	192
GCF_02	0	255	0	0

Membership value	GCF_01	GCF_02
BGC_01	89	382
BGC_02	372	0
BGC_03	89	382

Figure 7.1. An overview of BiG-SLiCE's GCF analysis workflow. Taking an input of region/cluster GenBank files from antiSMASH and MIBiG, (A) BiG-SLiCE converts BGCs into numerical feature vectors, which are used to (B) construct the GCF models (cluster centroids) and (C) calculate BGC-to-GCF membership values. Processed data and results are all stored in a file-based SQL database (using SQLite3 [41]), which can then be used to perform further analysis (via external scripts) or to visualize the result in a user-interactive application (D).

Several approaches have been previously developed to perform multi-protein vectorization by adapting the Word2Vec [46] natural language processing algorithm. ProtVec applies an n-mer amino acid residue embedding to model sequence identity as a continuous multi-dimensional vector which can be used for protein family classification. While theoretically we can aggregate features from multiple proteins to generate a BGC vector, its applicability might be limited, as the extended total sequence length will lower the vector feature's discriminative power. More recently, the Pfam2Vec approach that treats Pfam domain hits as tokens has been implemented and used to encode genome content [47], assign putative functions to unknown Pfam domains [48], and predict new BGC classes [39]. However, the construction of these models typically involves extensive hyperparameter tuning [47,49], which together with the less directly interpretable nature of the embedded vectors complicates the clustering (i.e., threshold assignment) problem.

For BiG-SLiCE, we therefore chose to take the more simple and straightforward approach of directly constructing a domain absence/presence matrix for each BGC, which is reminiscent of the Jaccard Index (JI) component of the BiG-SCAPE algorithm. To improve the information content of the original JI index, we semi-manually curated two sets of feature models: 1) the biosynthesis-specific domains (biosynthetic-Pfam) and 2) the clade-specific signature domain fingerprints (sub-Pfam).

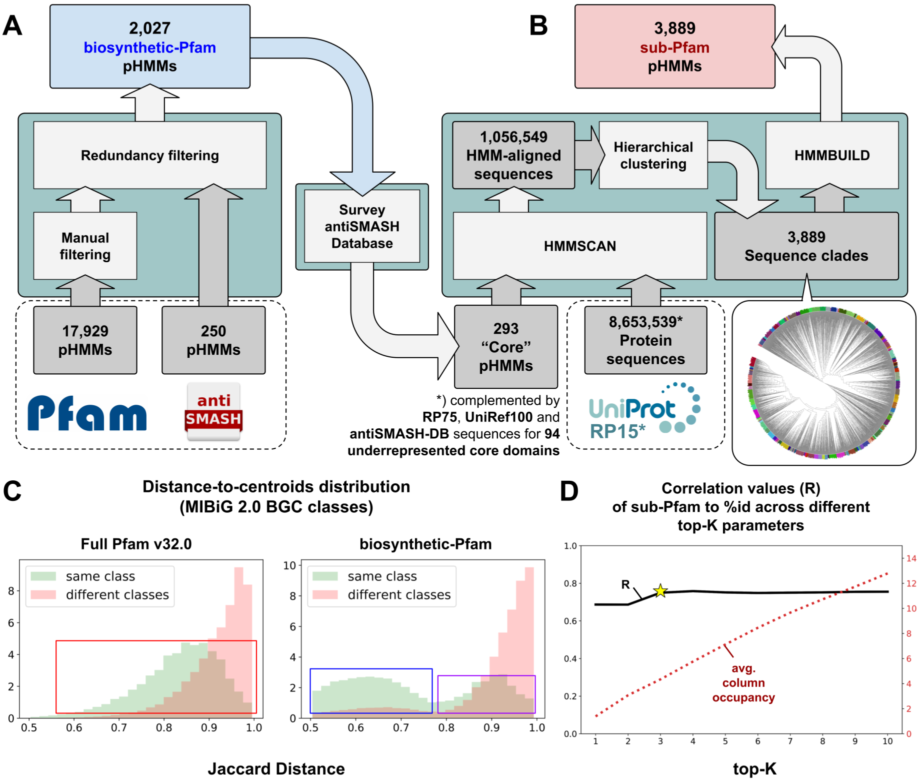


Figure 7.2. (A) Construction of biosynthetic-Pfam features and (B) Sub-level Pfam (sub-Pfam) features. (C) Effect of Pfam model filtering on the discriminatory power of domain-presence Jaccard distance (JI index in BiG-SCAPE) measurements to separate MIBiG v2.0 generic classes (Polyketide, NRP, RiPP, Alkaloid, Terpene, Saccharide, Other). It is shown that the filtering strategy will produce more clearly separated within-class distances (blue box) than the full Pfam counterparts (red box). The second mode at the right side of the biosynthetic-Pfam same-class distribution (purple box) largely stems from hybrid BGCs, containing signature domains of two or more distinct classes (i.e. NRPS-PKS, PKS-Terpene-Saccharide, etc.). (D) Pearson correlation values between protein sequence similarity (%-identity) and the corresponding sub-Pfam-based scoring in all AMP-binding domains (3,419 sequences, 879 BGCs) from the MIBiG v2.0 dataset across different top-K settings. Better correspondence (avg. R=0.75) is shown starting at top-K=3 (BiG-SLiCE's default) onwards. The larger the top-K values, the more columns occupied (dashed red line) by the BGC's composite sub-Pfam features as opposed to the biosynthetic-Pfam features, which can be thought of as a way to "tune" the core domain's feature weight (akin to BiG-SCAPE's anchor boost setting).

Feature set 1: biosynthetic domain absence/presence matrix (biosynthetic-Pfam). Domain hits (retrieved using hmmscan [50] with the gathering threshold) obtained for a reduced list of Pfam version 32 [42] pHMM models (Figure 7.2A) were used to construct a boolean (here represented by values of 0 or 255) feature matrix for every BGC. This list was constructed by filtering all Pfam domains for biosynthetically related protein families using the combination of ECDomainMiner

[51] (which allows us to filter for domain related to enzymatic functions) and manual filtering based on each domain's full description (Supplementary Table 1). This filtering was done to reduce the influence of non-biosynthetic domains, i.e., from genes that may be important for a BGC to function but are not directly responsible for generating structural variation of the produced metabolites (such as transporter enzymes and regulators). A library of 250 pHMM models from antiSMASH [19] was also included, as they harbor many curated biosynthetic domains not covered by the Pfam database alone. Altogether, this combination of 2,027 "biosynthetic-Pfam" models shows an increased selectivity compared to the full Pfam database when used to separate BGCs according to the chemical class of their predicted products (Figure 7.2C).

Feature set 2: signature domain fingerprinting (sub-Pfam). While the biosynthetic-Pfam models work well to capture the pattern of BGC diversity across generic chemical classes, they are not sensitive enough to cover the more granular level of the inter-class diversity. BGCs of the same class typically share a limited set of "core" enzymes that determines the end product's scaffold based on the combination of their specificity and/or copy number variation. For example, the compound's scaffold produced by a Type-I Polyketide BGC is largely driven by the specificity of its (often multiple) Acyltransferase (AT) and Ketosynthase (KS) domains [52]. To cover this sequence-level protein diversity, we constructed alignments of 9,451,490 representative protein sequences in the RP15 database (Release 2020_01) [53] to our pre-selected 293 core biosynthetic domain pHMMs (Supplementary Table 2). We performed hierarchical clustering analysis to group similar aligned sequences into clades, then built sub-level protein family pHMMs from the sequences of each clade (Figure 7.2B). This approach resulted in a distinct set of 3,889 sub-level Pfam (sub-Pfam) models (10-100 clades per core domain). For each aligned core domain in a BGC, an hmmscan search is performed using the specific sub-Pfam models, of which the hits are then ranked according to their bitscores. A set number of top hits (top-K) is then used to assign descending values of the corresponding feature in the matrix - for example, if a domain A has top-3 hits of A-c15, A-c3, and A-c2, its ranked feature values could be A-c15=255, A-c3=170 (255 x ⅔), and A-c2=85 (255 x ⅓). When a BGC has multiple hits on the same sub-Pfam column, the maximum value for that column will be taken. Using this ranked normalization scoring strategy for building the numerical feature representation of each core gene, we show that the sub-Pfams can together act as a proxy for sequence-level protein diversity (Figure 7.2D).

7.2.2. GCF models construction

To efficiently group BGC features into GCFs, BiG-SLiCE uses a clustering method based on the python scikit-learn [54] implementation of the BIRCH [45] algorithm. When using gene cluster GBK files from antiSMASH v4.2 or higher (the version in which the attribute `on_contig_edge` was implemented to indicate which BGCs lie on the edge of a contig and may therefore be incomplete), users can opt to build the GCF features only from non-fragmented BGCs (using `--complete` parameter). Then, a distance sampling test will be performed to ascertain a default threshold value *T* for the clustering algorithm, unless a value is directly supplied by users via the `--threshold` parameter.

The former is done by taking the average X^{th}-percentile (default X=1) of Euclidean pairwise distances between 100x1000 randomly sampled features from the input data. Afterwards, a flat-tree BIRCH (*branching_factor* >= *n_samples*) [55] clustering method is used to incrementally scan BGC features and build the GCF centroids. Then, a global cluster assignment is performed to match all input BGCs with the top-*N* (default *N*=3) scoring GCFs per BGC along with their membership scores. By considering multiple GCFs at once, users will be able to judge the confidence level of each BGC-to-GCF assignment. This is useful, for example, when determining the context of a fragmented BGC, where (low) membership scores might be distributed almost equally across different best-matching GCF models. Furthermore, by performing feature extraction on a set of newly sequenced (putative) BGCs, users can immediately match them with previously calculated GCF models (using the `--query` mode of BiG-SLiCE) and retrieve information on their characteristics and potential novelty.

7.2.3. Comparison against manually curated GCFs

In order to judge the quality of results produced by its heuristic-based algorithm, we compared BiG-SLiCE clustering against 92 manually curated groups of MIBiG v1.3 BGCs provided in the original BiG-SCAPE paper. Several different threshold parameters *T* were tested {300, 600, 900, 1,200, 1,500} and corresponding results were compared to the reference groups. We calculated the *V-score* [56] of each run, which measures both the homogeneity (whether cluster members share the same target class) and completeness (whether members from a single target group are assigned into exactly one cluster) of a clustering result when matched to a manually defined target reference (Figure 3A), and plot it alongside the difference of GCF counts (*⊿GCF*) between the two. We found that BiG-SLiCE produces a generally agreeable result at the selected example threshold (*T*=1,100 with *V-score*=0.81), but is not able to capture the "perfect" clustering denoted by the reference groups (Figure 7.3B). This stems from the fact that the manual categorization of the 92 compound groups does not always translate into the groups sharing a similar distance distribution in the BGC space, making it impossible to set a single clustering threshold that reproduces the membership assignment. BiG-SCAPE seems able to handle this issue better (*V-score*=0.91, see Supplementary Figure 1) due to its Affinity Propagation [43] based clustering algorithm that allows finding non-convex clusters, as opposed to the spherical partitioning approach of BIRCH, which is one of the main trade-offs for its hyper-scalability. However, BiG-SLiCE accurately captures the underlying biosynthetic signal that connects the genomic space of BGCs and the chemical space of their products, as demonstrated by the bimodal distribution of intra- vs inter-group distances of the BGCs (Figure 7.3C) and the visualized feature heatmap of the most challenging groups (Figure 7.3D).

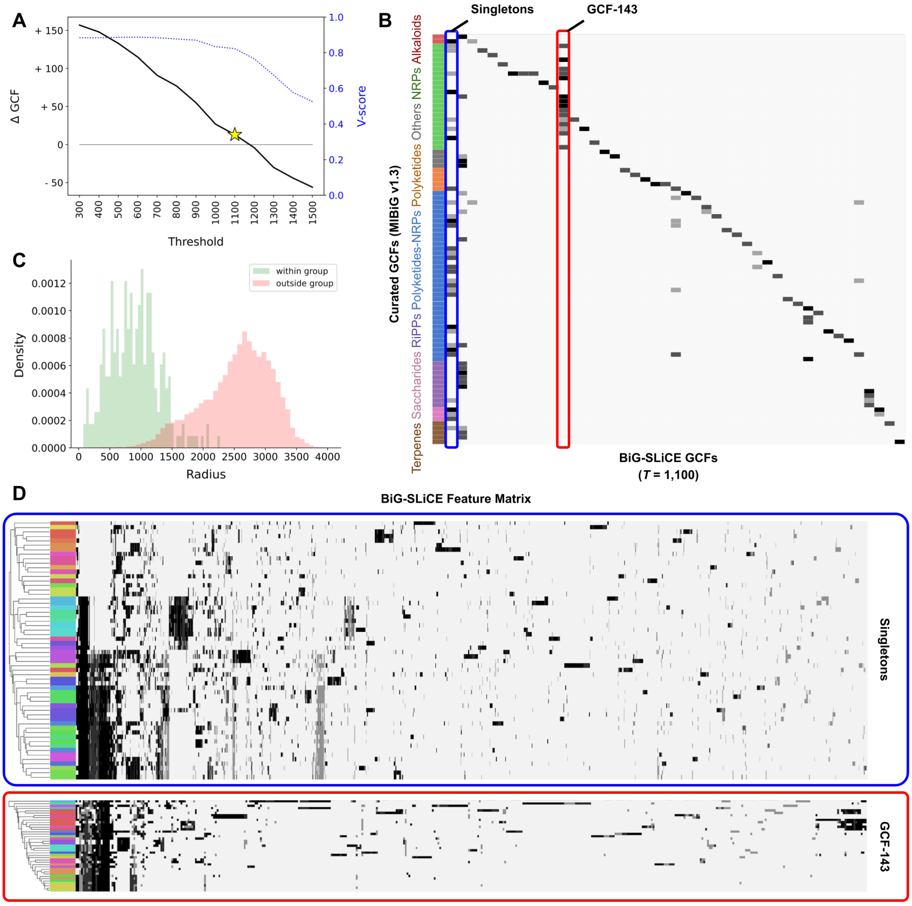
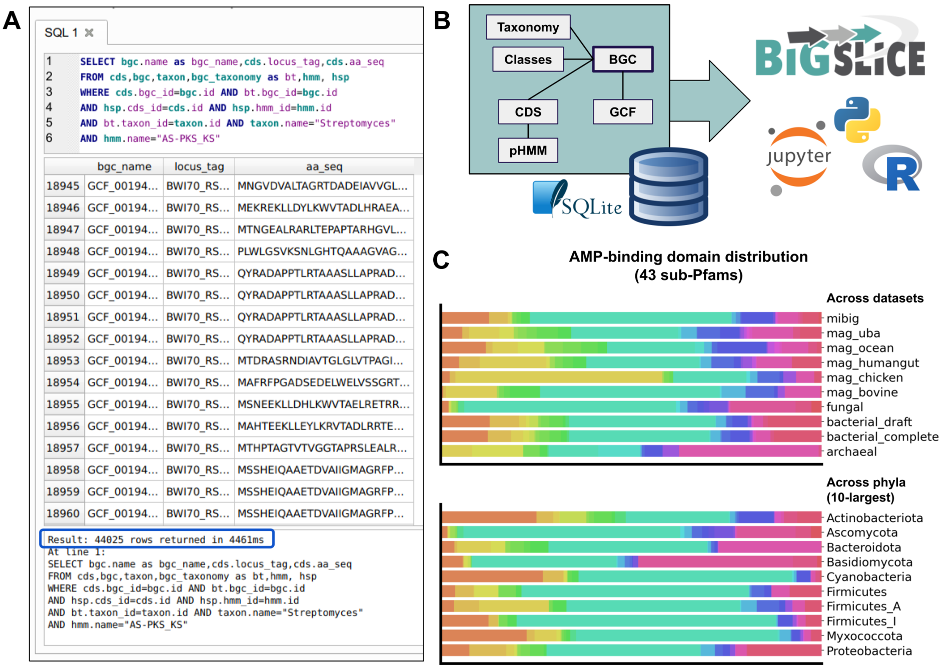


Figure 7.3. (A) BiG-SLiCE analysis results for a range of threshold values, as measured by the difference of GCF counts (*ΔGCF*) and the level of clustering agreement (*V-score* of 1.0 for perfect clustering) compared to MIBiG curated groups. A single threshold result with the lowest *ΔGCF* while maintaining a *V-score*>0.8 (*T*=1,100) was selected as an example for further analysis in this figure. (B) Confusion matrix of BiG-SLiCE clusters vs curated GCFs. To help in visualization, all singletons of the BiG-SLiCE result (58 GCFs) were collapsed into a single column (leftmost column, highlighted in blue box), showing together BGCs requiring a more lenient threshold (*T*>1,100) to match the curated information. Conversely, another column, GCF-143 (red box), highlights the need for a stricter threshold (*T*<1,100) to obtain a more fine-grained clustering for some parts of sequence space. (C) BGC-to-centroid distance value (i.e. radius) distribution of within and between group pairs in the curated dataset. The centroid of each curated group was calculated by averaging the feature vectors of all BGCs assigned to it. (D) Feature heatmap of the collapsed singleton group and GCF-143 (colored bars on the left indicate manually curated groups) showing that the underlying pattern captured by BiG-SLiCE features tends to agree with the manually curated information, i.e., rows with the same color tend to be located near each other.

7.2.4. SQL-based data storage enables extensive functionality

A typical BiG-SLiCE run produces a large amount of useful information on top of the GCF membership for each BGC. Taxonomic metadata, information on chemical compound classes and protein annotations are commonly included in the antiSMASH-generated BGC genbank files. To integrate that information and provide a truly comprehensive analysis output, a structured approach to data storage and processing is required. The architecture of BiG-SLiCE is centered around the use of a relational SQL database schema (Supplementary Figure 2) implemented as a file-based SQLite data store. Processed input (including all metadata), supporting data and clustering results are systematically stored in database tables.



```sql
SELECT bgc.name as bgc_name,cds.locus_tag,cds.aa_seq
FROM cds,bgc,taxon,bgc_taxonomy as bt,hmm, hsp
WHERE cds.bgc_id=bgc.id AND bt.bgc_id=bgc.id
AND hsp.cds_id=cds.id AND hsp.hmm_id=hmm.id
AND bt.taxon_id=taxon.id AND taxon.name="Streptomyces"
AND hmm.name="AS-PKS_KS"
```

	bgc_name	locus_tag	aa_seq
18945	GCF_00194...	BWI70_RS...	MNGVDVALTAGRTDADEIAVVGL...
18946	GCF_00194...	BWI70_RS...	MEKREKLLDYLKWVTADLHRAEA...
18947	GCF_00194...	BWI70_RS...	MTNGEALRARLTEPAPTARHGVL...
18948	GCF_00194...	BWI70_RS...	PLWLGSVKSNLGHTQAAAGVAG...
18949	GCF_00194...	BWI70_RS...	QYRADAPPTLRTAAASLLAPRAD...
18950	GCF_00194...	BWI70_RS...	QYRADAPPTLRTAAASLLAPRAD...
18951	GCF_00194...	BWI70_RS...	QYRADAPPTLRTAAASLLAPRAD...
18952	GCF_00194...	BWI70_RS...	QYRADAPPTLRTAAASLLAPRAD...
18953	GCF_00194...	BWI70_RS...	MTDRASRNDIAVTGLGLVTPAGI...
18954	GCF_00194...	BWI70_RS...	MAFRFPGADSEDELWELVSSGRT...
18955	GCF_00194...	BWI70_RS...	MSNEEKLLDHLKWVTAELRETRR...
18956	GCF_00194...	BWI70_RS...	MAHTEEKLLEYLKRVTADLRRTE...
18957	GCF_00194...	BWI70_RS...	MTHPTAGTVTVGGTAPRSLEALR...
18958	GCF_00194...	BWI70_RS...	MSSHEIQAAETDVAIIGMAGRFP...
18959	GCF_00194...	BWI70_RS...	MSSHEIQAAETDVAIIGMAGRFP...
18960	GCF_00194...	BWI70_RS...	MSSHEIQAAETDVAIIGMAGRFP...

```
Result: 44025 rows returned in 4461ms
At line 1:
SELECT bgc.name as bgc_name,cds.locus_tag,cds.aa_seq
FROM cds,bgc,taxon,bgc_taxonomy as bt,hmm, hsp
WHERE cds.bgc_id=bgc.id AND bt.bgc_id=bgc.id
AND hsp.cds_id=cds.id AND hsp.hmm_id=hmm.id
AND bt.taxon_id=taxon.id AND taxon.name="Streptomyces"
AND hmm.name="AS-PKS_KS"
```

Figure 7.4. (A) An example SQL query for all protein sequences harboring at least one Ketosynthase (AS-PKS_KS) domain from streptomycete BGCs. Here, the search performed against the total of ~29 million CDSes and >101 million domain hits in the database was completed in under five seconds, returning 44,025 CDS that satisfy the criteria. (B) A cartoon illustration on how the interconnected SQL tables holding various BGC-related information can be leveraged by downstream analyses, e.g., using programs and notebooks written in Python and R. (C) An example downstream analysis using the data on sub-Pfam hits to chart the diversity of AMP-binding domains across datasets and across phyla. Here, each colored bar represents the distribution of a specific sub-Pfam clade across the sampled dataset/phylum. Each analysis including the SQL query took around 55 seconds to complete. A script to perform such analyses (which can also be used to investigate other biosynthetic domains) and generate the plots can be found in the “figure_4” folder of the Supplementary Data.

Using this setup, it is possible to build complex queries and perform all sorts of analyses even beyond the scope of GCF reconstruction. For example, one can use the preprocessed SQL database as a personal “data management” solution for custom BGC collections, enabling a fast search and query of specific protein sequences based on taxonomy and domain contents (Figure 7.4A). Furthermore, this structured information about BGCs, their homology (GCF membership), taxonomy, biosynthetic classes, and protein domain hits can also be combined with a bioinformatics pipeline or analytical scripts written in Python or R (both of which have native support for SQLite) (Figure 7.4B) to perform even more complex analyses, for example to study the diversity of biosynthetic domains across samples and across taxonomy (Figure 7.4C). As a matter of fact, all analyses performed in this study heavily benefitted from, and relied on the data-wrangling convenience provided by BiG-SLiCE's SQLite database.

Finally, as previously demonstrated by the success of antiSMASH and BiG-SCAPE, one way in which regular end users can really benefit from a tool is when they are provided with an interactive and easy-to-use output visualization as a way to explore the data and analysis results. BiG-SLiCE offers this functionality by combining the portability of SQLite database with a mini web application written using Python's Flask library [57]. This allowed us to implement a feature-rich visualization "software" that can be deployed and run with minimal amount of installation effort on a user's personal computer. While this feature is currently at a prototype stage, offering simple functionalities such as browsing and viewing the processed BGCs and GCFs, we plan to continue to improve and implement more advanced features along the way, such as searching and filtering for specific BGCs/GCFs of interest, generating phylogenomic alignments of BGCs [22,58], or even incorporating additional useful information such as the presence/absence of antibiotic-resistant genes [59] and regulatory domains [60] within the BGCs.

Results and Discussion

In order to show how BiG-SLiCE could be applied to large datasets that capture the full diversity of BGCs from cultured and uncultured microbes, we decided to collect a merged dataset of publicly available microbial genomes and metagenome-assembled genomes (MAGs). We then predicted their BGCs using antiSMASH v5.1.1, filtering out contigs <5,000bp (`--minlength 5000`) and used the respective taxonomy options wherever applicable (`--taxon bacteria` for bacterial and archaeal genomes, and `--taxon fungi` for fungal ones).

7.3.1. Collecting a near-comprehensive dataset of publicly available BGCs

We downloaded 19,169 complete and chromosome-level bacterial NCBI RefSeq genomes up to 27 March 2020, 12:15PM CET. To capture the extensive strain-level diversity within the bacterial kingdom, 162,352 draft RefSeq genomes were also downloaded and processed, resulting in a total number of 1,060,594 BGCs when combined. For fungi and archaea, we downloaded 5,939 and 1,162 genomes from NCBI Genbank with "Refseq-like" filters turned on, resulting in 123,939 fungal and 2,578 archaeal BGCs, respectively (all NCBI query scripts used for this data collection step are available in Supplementary Text 1).

Table 7.1. Numbers of genomes and BGCs in all datasets included for the large scale diversity analysis. Numbers inside brackets indicate the total number of genomes assigned to each kingdom based on the subsequent taxonomy analysis. “Others” category includes the kingdom of Archaea, Viridiplantae (from MIBiG dataset) and unassigned taxa. A complete list of all genome accessions and their BGC counts can be seen in Supplementary Table 3.

Dataset Name	Study	Counts (Genomes, BGCs)					
		Bacterial		Fungal		Others	
RefSeq complete bacteria	-	19,169 (19,166)	101,531	0 (0)	0	0 (3)	0
RefSeq draft bacteria	-	162,352 (162,297)	959,061	0 (0)	0	0 (55)	346
GenBank fungi	-	0 (0)	0	5,939 (5,905)	123,816	0 (34)	123
GenBank archaea	-	0 (1)	2	0 (0)	0	1,162 (1,161)	2,109
Parks 2017 (Uncultivated Bacteria and Archaea MAGs)	[61]	7,280 (7,280)	15,829	0 (0)	0	623 (623)	756
Tully 2018 (TARA ocean MAGs)	[62]	2,283 (2,326)	4,829	0 (0)	0	344 (301)	518
Almeida 2019 (Unified Human Gut MAGs)	[63]	4,616 (4,616)	4,766	0 (0)	0	28 (28)	25
Stewart 2019 (Cow’s rumen MAGs)	[64]	4,815 (4,815)	8,380	0 (0)	0	126 (126)	589
Glendinning 2020 (Chicken’s caecum MAGs)	[65]	469 (469)	481	0 (0)	0	0 (0)	0
MIBiG v2.0	[66]	0 (0)	1,594	0 (0)	276	0 (0)	40
Total		200,984 (200,970)	1,096,473	5,939 (5,905)	124,092	2,283 (2,331)	4,506

Furthermore, we collected and processed 20,584 MAGs from previously published studies [61–65], resulting in a total of 36,173 BGCs. This list was arbitrarily selected from available studies describing the construction of large-scale MAG assemblies from different environments at the time of data collection. Although this list was in no way comprehensive (for example, there are many other notable recent publications [67–78] not covered by this initial effort, not to mention the huge number of shotgun metagenomic studies publishing only contig-level assembly of unassigned bins), the ~20K MAGs presented here may already give us a glimpse on the untapped biosynthetic diversity of uncultured microbes. Finally, we incorporated all 1,910 entries from MIBiG v2.0 [66] as a reference set of known and experimentally verified BGCs. In total, a final count of 1,225,071 BGCs were predicted from 209,206 genomes and MAGs, as shown in Table 7.1 above.

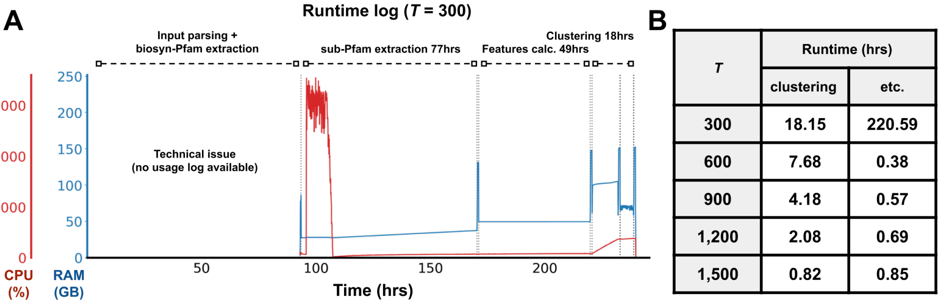
7.3.2. Improving the taxonomy assignment of genomes

Before performing any taxonomy-related diversity analysis, we ensured that all included genomes were correctly assigned to their respective taxa. Several studies pointed out that there might be a potentially widespread misclassification of bacterial genomes within the NCBI database [79–81]. To avoid this issue, we chose to use the taxonomy derived from the GTDB (Genome Taxonomy Database), which were posited to be more phylogenomically accurate than that of NCBI [82]. We queried all bacterial and archaeal NCBI genome accessions through the GTDB API (version 04-RS89, https://gtdb.ecogenomic.org/api/) to fetch their taxonomy information, resulting in 123,245 taxonomy-assigned genomes. For the remaining genomes, i.e. those from metagenomic studies and more recent NCBI genomes not yet covered by the API, we used the GTDB toolkit [82], a bioinformatics pipeline that integrates several tools [50,83–87], to infer their taxonomy based on their genomic marker composition. This further assigned taxonomy information to another 79,964 genomes. Original NCBI taxonomy information was retained for all fungal genomes and MIBiG BGCs (a list of all GTDB and NCBI-assigned taxonomy per genome is available in Supplementary Table 3).

7.3.3. Large-scale Homology Analysis of 1.2 Million BGCs

We then performed BiG-SLiCE clustering analyses over the merged datasets using a 36-core, 252GB RAM shared computing server facility. Taking advantage of the antiSMASH5-enabled annotation of fragmented BGCs (clusters residing on contig edges), the `--complete-only` parameter was used for the clustering phase, using 802,287 (65%) non-fragmented BGCs from the input data to build the GCF models. This ensures that the variation in the models is derived from actual BGC diversity and not due to technical gene losses (from contig splits). Later on, the full input datasets were queried back against the GCF models, in order to map the fragmented BGCs onto their corresponding GCFs based on the calculated membership values d. For this analysis, we arbitrarily categorize GCF-to-BGC relationships into “core” ($d<=T$), “putative” ($T<d<=2T$), or “orphan” ($d>2T$)

on a best-hit basis (parameter `--n_ranks=1`). Five different threshold values (*T*={300, 600, 900, 1,200, 1,500}) were tested, producing a decreasing number of GCF models (more BGCs per GCF) as *T* gets bigger (more lenient) (Supplementary table 4). The first run (*T*=300) which carries the full workflow load (from features extraction to membership assignment) was finished in ~240 hours (ten days), or >150x faster than the estimated runtime of BiG-SCAPE (Supplementary Figure 3). A large chunk of this runtime is spent at the feature extraction step, which includes the I/O heavy hmmscan and non-parallelizable SQL inserts (Figure 7.5A). Subsequent runs (*T*=600-1,500) reused the precalculated features, taking only an average of around four hours runtime for each run (Figure 7.5B).



T	Runtime (hrs)	
	clustering	etc.
300	18.15	220.59
600	7.68	0.38
900	4.18	0.57
1,200	2.08	0.69
1,500	0.82	0.85

Figure 7.5. (A) Runtime breakdown of the full run (*T*=300) on a 36-core CPU, 262GB RAM server. Due to some technical issues, no usage log is available for steps prior to the sub-Pfam extraction. CPU usage log shows that most of the time, BiG-SLiCE only uses one CPU core, giving a room for further improvement e.g. via SQL parallelization. Spikes on the RAM usage (peak ~150GB) came from the periodic "dumping" of the in-memory database (used in order to speed up runtime) into an SQLite db file. (B) Runtime comparison between multiple runs, with *T*=300 bearing the full load of performing input processing and features extraction. Here, runtimes are separately shown for both the clustering (GCF models construction plus membership assignment) and other steps (input parsing, hmmscanning and features extraction).

7.3.4. Charting a global map of BGC diversity

Each GCF in the global clustering analysis result represents a functional niche captured from a group of BGCs sharing a similar biosynthetic make-up. To enable the visualization of this biosynthetic diversity, we partitioned the 121,299 centroid features of the GCFs produced by the *T*=300 run into 500 GCF "bins" using K-Means (via sci-kit's library, with *K*=500 and a random, but reproducible initialization step; see the reproduction script included in Supplementary Data for details). Another round of membership assignment was performed to match the full set of 1.2M BGC features into the resulting 500 GCF bin centroids. Those centroids were also subjected to an average-linkage agglomerative clustering analysis (sci-kit implementation, euclidean distance). The produced hierarchical tree object was then converted to a newick file (using a custom script provided in the Supplementary Data) and plotted via the iTOL web server (https://itol.embl.de/) [88]. By annotating this tree with various types of quantitative information (Supplementary Table 5), the resulting phylogram pictures a generic, "bird-eye view" on the entire set of 1.2 million BGCs (Figure 7.6).

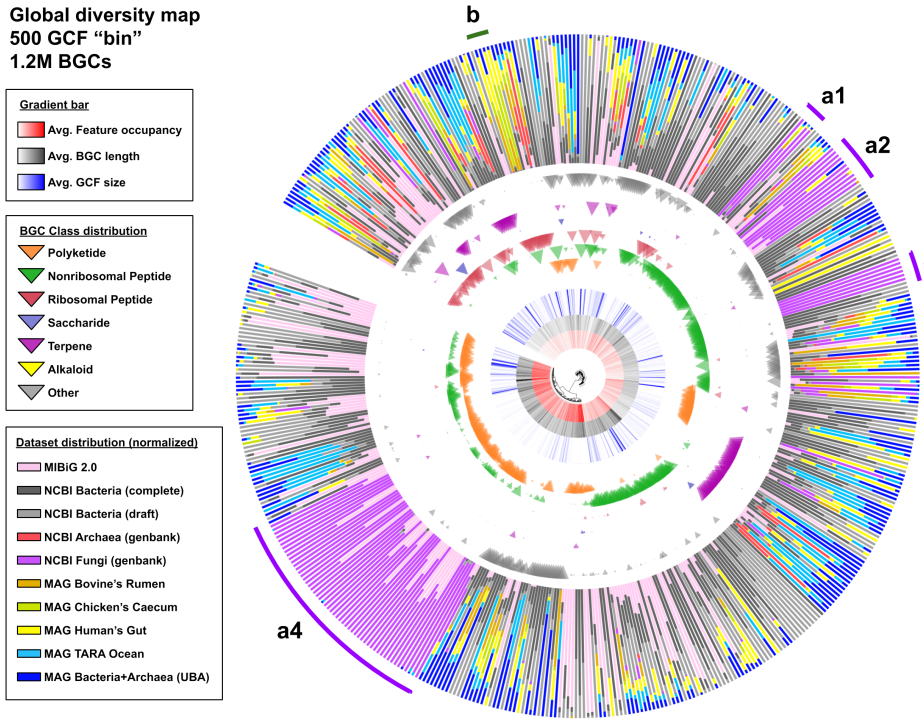


Figure 7.6. A phylogram created via the hierarchical clustering analysis of 500 GCF bins. The phylogram was rooted on a null (all zeros) dummy feature matrix. For each node, the raw dataset distribution values (Supplementary Table 5) were double-normalized, first against the number of BGCs each dataset has in total, giving the fraction values, then against all fraction values of other datasets in the bin. Furthermore, some notably interesting clades were manually highlighted (a1-a4, b) for follow-up discussion (see main text).

An important thing to note is that due to the non-deterministic nature of K-means, the number of BGCs that goes into each bin depends a lot on the randomly placed initial centroids (for example, there are 21 bins made up of a single BGC (Supplementary Table 5), which can happen when the randomly placed initial centroid hits an outlier/singleton in the dataset). This is analogous to taking a two-dimensional satellite picture of the earth from a specific coordinate, looking down at a specific angle. There are an infinite number of ways to take a picture, giving a different perspective and snapshot of an object each time, but the inherent three-dimensional structure of the object will always remain constant. While the map shown in Figure 7.6 can give us insights into the major "landmarks" formed by the larger groups of BGCs, it will not show all the nooks and crannies to be explored from the entire dataset (which could be explored using more fine-grained tools such as BiG-SCAPE).

The very first thing that we can notice from the phylogram is how fungal BGCs (purple bars, "a1" to "a4") have quite distinct features that discriminate them from the rest of the (mostly bacterial) datasets. Clades "a1" to "a3" contain mostly NRP (99.93%) BGCs: 20,398 from "a1", 18,770 from "a2" and 8,606 from "a3". Clade “a1” shares its 9,402 fungal BGCs with 10,972 bacterial (67.56% came from *Pseudomonas*) and 13 archaeal ones. This clade includes two simple NRP-encoding fungal BGCs from MIBiG dataset, encoding the biosynthesis of the proteasome inhibitor fellutamide B [89] (BGC0001399) and aspergillic acid [90] (BGC0001516) from *Aspergillus* (and on the bacterial side: four MIBiG BGCs including another simple proteasome inhibitor livipeptin [91,92] encoded by BGC0001168 from *Streptomyces lividans*). Clade “a2” contains a major part (50 out of 61) of known non-hybrid fungal NRP BGCs in MIBiG, and shares the clade with 85 bacterial NRPs. Last but not least, clade “a3” almost exclusively (except for 1 beta-lactam BGC from *Mycobacterium gordonae* and 10 BGCs from unknown taxa) consists of uncharacterized fungal NRPs. A closer look at this clade leads to an interesting observation in terms of shared features / domains. We found that no domain (even at biosynthetic-Pfam level) is shared by more than 70% of the BGCs, except from a few sub-Pfams: AS-NAD_binding_4-c7 (91.92%), AS-AMP-binding-c6 (98.84%) and Epimerase-c26 (99.03%). These domains are often contained in one protein-coding gene, sometimes with an extra ACP (AS-PP-binding) domain (found in 75.34% of the BGCs). This clade therefore seems to contain mostly proteins related to α-aminoadipate reductases, which have been previously inferred to have an evolutionary origin prior to, or early in, the evolution of fungi [93]. Detailed results and reproducible scripts for analyses from this and subsequent paragraphs can be found in the “figure_6+sup_table_5” folder of the Supplementary Data.

At the opposite side of the phylogram, 42,716 out of 43,840 (97.43%) BGCs from clade ”a4” are of the Type-I Polyketide (T1-PKS) subclass, and as many as 7,811 of them are “true” PK/NRP hybrids (determined by the presence of Acyltransferase, Ketosynthase, AMP-binding and Condensation domains together in the BGC). This clade shows an enrichment of AS-PKS_AT-c7 (95.1%) and ketoacyl-synt-c8 (95.94%) sub-Pfam domains possibly linked to the iterative mechanism almost exclusively attributed to fungal PKSes [94]. Interestingly, 2,255 BGCs from this clade have bacterial origins (966 *Mycobacterium*, 438 *Streptomyces*, 851 others), which might possibly be connected to a group of non-canonical, iterative T1-PKSes from bacteria [95–97]. However, no bacterial BGC from MIBiG, including those of known iterative type [98,99], falls into this clade.

We can also see a narrow but distinct clade “b” highly represented by RiPP BGCs from the “gut” metagenome datasets (bovine’s rumen, chicken’s caecum, human gut). Aside from the 2,546 (17.88% of the three datasets total) MAG-derived BGCs, this clade also contains 4,254 BGCs from the NCBI bacterial RefSeq genomes (0.40% of the dataset’s total) and is populated by BGCs from various kinds of firmicutes (99.32% of the clade’s total). Looking closer at the BGC classes gives away an important clue: 99.68% of the BGCs belong to the sactipeptide RiPP subclass as annotated by antiSMASH, and seems to encode a group of RiPPs known as SCIFF (Six-Cysteine in Forty-Five) peptides [100] (recently proposed to be reclassified as ranthipeptides [101]), as 100% of those RiPPs have the signature TIGR03973 precursor domain (along with >99% occurrence of Radical_SAM and the iron-sulfur binding Fer4_12 domains). It is largely unknown why this particular class of BGCs are highly represented in the gut microbiomes, except for the fact that they can only be found in typical resident microbes of those environments (80.52% of BGCs came from *Clostridia*). Recently, a series of analyses performed by Chen et al. [102] in solventogenic *Clostridia* suggested that these RiPPs might play a role in the quorum sensing system and in controlling cell metabolism of such organisms.

Next, by looking at how the pink (innermost) bar is spread all across the phylogram, we can infer that despite holding no more than 2,000 entries presently, the BGCs in the MIBiG database are actually diverse enough to cover much of the general diversity of BGCs. However, we also need to be aware of the fact that most of the detection rules in antiSMASH were almost directly derived from the knowledge of experimentally characterized BGCs that are also present in MIBiG. This means that the 1.2 million BGCs we captured from those 209 thousand genomes are all evolutionarily related, although distantly, to at least one MIBiG BGC. To go beyond these canonical pathways, several unsupervised but "lower-confidence" alternative algorithms [39] have been developed that can potentially complement antiSMASH to cover more exotic areas of biosynthetic space.

Finally, this visualization suggests that several aspects can still be improved upon this first version of the BiG-SLiCE clustering algorithm. The three innermost gradient bars of the phylogram show the variation in the length of BGCs, extracted features, and the size of GCFs. By looking at them, it is quite apparent that there is a distinct separation between two major groups of GCF bins: a high feature counts group (more intense red bars) consisting mostly of domain-rich Polyketide (and some nonribosomal peptide / NRP) BGCs, and a low feature counts group (less intense red bars) consisting a large majority of NRP BGCs along with most Terpene and RiPP BGCs (Supplementary Figure 4A). This causes a large dichotomy in GCF sizes (Supplementary Figure 4B) due to the limitation of the single-threshold clustering method of BIRCH as described before. While, generally, the number of extracted features depends a lot on the length of a BGC (longer BGCs may contain more genes and domains), this is not always the case. For example, there may be a great degree of copy number variation between biosynthetic domains (e.g. in some NRP BGCs) that is not captured by BiG-SLiCE (Supplementary Figure 4C), as it only looks at absence/presence patterns of (sub-)Pfam features. Additionally, the pHMM models of BiG-SLiCE may fail to capture the diversity of certain tailoring domains. Conversely, there are also cases where the structure of the end products depends largely on the residue-level variability of particular proteins, such as for the large majority of RiPP BGCs, in which biochemical variation is largely governed by the sequences of precursor peptides (Supplementary Figure 4D). Thus, one way to optimize BiG-SLiCE clustering in the future is to try and balance the average feature counts across BGC (sub)classes, i.e. by surveying and including the missed neighboring domains, by putting more emphasis on core domain specificity (more columns for subpfam models) of a manually selected set of enzymes, and/or by taking into account copy number variation of domains (e.g. counting the actual number of biosynthetic-pfam hits rather than using a boolean absence/presence value). Alternatively, large BiG-SLiCE GCFs can be analyzed in more detail using BiG-SCAPE or using protein sequence similarity networks [103] (which can, for example, be very powerful for analyzing RiPP precursor peptide variation [104–106]).

7.3.5. *Measuring the "hidden iceberg" of microbial secondary metabolism*

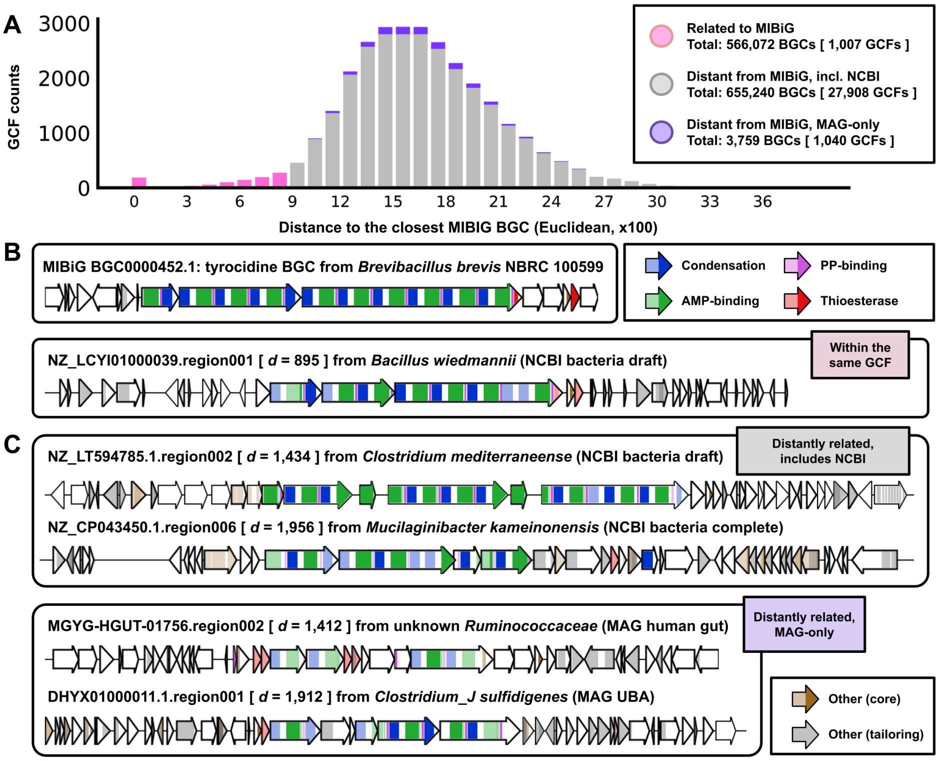


Figure 7.7. (A) Histogram of Euclidean distances (x-axis) of GCF models to their closest BGC from the MIBiG 2.0 dataset. Here, all GCFs having *d*<=900 were denoted as "Related to MIBiG" and "Distant from MIBiG" if otherwise, particularly highlighting those coming only from the MAG datasets. (B) Selected anecdotal example of a MIBiG BGC and one of the farthest (*d*=895) BGCs from the same GCF, which does not encode a biosynthetically equivalent pathway. Colored sections of the arrows represent biosynthetic domains captured by BiG-SLiCE, where darker colors represent putative core domain homologues (as measured by the sub-pfam signature) shared between the MIBiG BGC and its distant relatives. (C) Example BGCs from GCFs having a distant best-hit to the tyrocidine BGC as shown by their generally high *d* values (1,412-1,956) to the MIBiG BGC in question.

Only limited numbers of studies have considered global measurements of biosynthetic potential across taxa, or comparisons between cultivated and uncultivated bacteria [23,107,108]. To demonstrate how BiG-SLiCE could be used in such studies to quantify unexplored biosynthetic potentials, we took the 29,955 GCFs calculated from *T*=900 and measured the distance of every GCF model against their closest MIBiG BGC features (Supplementary Table 6), then plotted a histogram from the data (Figure 7.7A).

Indeed, it is immediately clear from Figure 7.7A that the great majority (96.63%) of GCFs remain uncharacterized (distantly related to any MIBiG BGC), representing a huge iceberg of unknown secondary metabolism hidden under the surface represented by the MIBiG database. Of these 28,948 GCFs, 1,040 can only be found in MAG datasets, representing unique BGCs from uncultured and unculturable microbes. However, care should be taken not to accept the numbers at face value, as there are still a lot of factors yet to be considered. On the one hand, while we previously showed that the 1,910 BGCs in MIBiG have good diversity coverage across biosynthetic classes, the database is not entirely comprehensive in capturing all experimentally characterized BGCs to date. On the other hand, the arbitrary threshold used to define the relationship (*T*=900) might be too lenient in some cases, as shown by an NRP BGC seemingly unrelated to the tyrocidine BGC being put together in the same GCF (Figure 7.7B). This also means that many BGCs with very low feature count would be lumped together in a large GCF with some MIBiG ones, contributing to an overestimated number (566,072 BGCs, or 46.2% of total input) of BGCs “related to MIBiG BGCs”. Combined with the fact that the analysis only includes what antiSMASH covers, we argue that the actual number of BGCs encoding distinct secondary metabolic pathways unrelated to known ones is likely to be even bigger.

7.3.6. Exploring biosynthetic potential across taxonomy

One of the potential use cases of BiG-SLICE is the systematic exploration of biosynthetic potential across taxonomy, which may provide detailed insight to direct discovery efforts. Having the species information of 209,206 genomes at hand, we sought to showcase how such an application could work by calculating the total number of GCFs within species having four or more strain-level genomes from our datasets (a total of 3,181 species from 1,043 genera) (Supplementary Table 7). To get a rough idea on the alpha diversity of GCFs within each species, we used the result of two threshold parameters, *T*=300 and *T*=900, and counted the numbers of GCFs per species across the two runs (Figure 7.8A). In this scenario, three Firmicutes (*Bacillus velezensis*, *Bacillus thuringiensis*, *Streptococcus pneumoniae*) and five Proteobacteria (*Escherichia flexneri*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Escherichia coli*, *Burkholderia ubonensis*) dropped out of the top thirty list of richest species when going from the stringent threshold to the more lenient one. This suggests that the perceived GCF richness in those species was largely confounded by the effect of (multiple) gene insertions/deletions near BGCs (in flanking regions included by antiSMASH) rather than the actual recruitment of new BGCs (i.e. via lateral gene transfer [109–111]).

Four *Streptomyces* species made it into the selected list of nineteen species that consistently ranked top thirty in both runs (Figure 7.8B] despite having relatively few genomes (24 to 78) in the dataset, confirming their status as prolific producers of natural products: 75-80% of approved antibiotics are sourced from this genus alone [1,112]. More detailed analysis of the set of species that have precisely four genomes in the dataset (723 species from 486 genera) (Supplementary table 7) showed that 26 species (104 genomes) from this “run-of-the-mill” drug discovery genus harbor an average number of 36.69 unique GCFs (at *T*=900) per species, putting it first among other bacteria, followed by *Saccharopolyspora* (36 GCFs from 1 species), *Nocardia* (avg. 30 GCFs from 2 species), and *Amycolatopsis* (avg. 29 GCFs from 3 species).

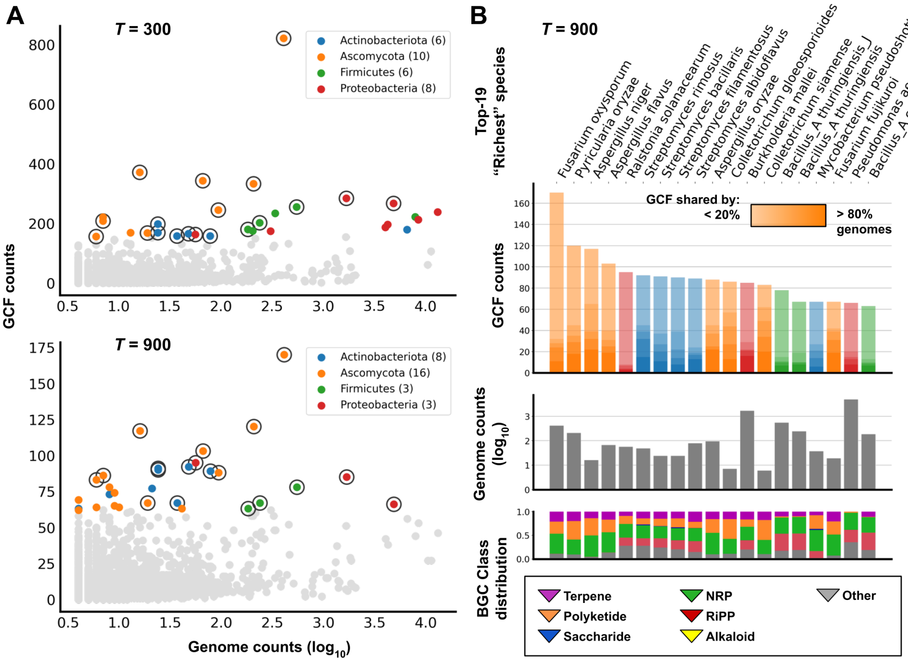


Figure 7.8. (A) Distribution of GCF counts across species having four or more genomes in the dataset. Two plots showing results at the most stringent (*T*=300) and a fairly lenient (*T*=900) threshold, each highlighting thirty species with the highest GCF counts (colored dots). Nineteen species present in the top thirty of both thresholds are marked with black circle. (B) Detailed view of the top-19 species, taking GCFs from the *T*=900 result. Gradients from the colored bars (GCF counts) represent the extent to which a GCF is shared between all genomes in a species (in 20%-wide steps) (Supplementary Table 7). Additionally, the total distribution of BGC classes per species is also measured (Supplementary Table 8).

The rest of the bacterial species (1 actinobacterium, 3 firmicutes and 3 proteobacteria) that made it into the top nineteen are mainly composed of pathogens that have had many of their genomes sequenced (183 to 4,838 genomes) within the NCBI database, which contributes greatly to their elevated GCF richness measure. However, two species from the list showed numbers that deviate from this observation. *Mycobacterium pseudoshottsii*, a slow-growing fish pathogen originally isolated from striped bass (*Morone saxatilis*) during mycobacterial outbreak in Chesapeake Bay [113] harbors a total of 67 unique GCFs within its 37 genomes. This makes the species distinct compared to the rest in the genus: *Mycobacterium avium* which harbors 58 GCFs from 197 genomes followed by *Mycobacterium tuberculosis* with 56 GCFs from 6,606 genomes. However, a closer look shows that the majority (35 out of 37) of the GTDB-Tk assigned genomes from this species actually belong to the closely related *Mycobacterium marinum* and *Mycobacterium ulcerans* in NCBI, which might explain the group's observed higher total GCF diversity. These accessions are now included and are assigned correctly in the newer version of GTDB R05-RS95 (and the accompanying GTDB-Tk version 1.3.0).

Ralstonia solanacearum (also known as *Pseudomonas solanacearum*), the final pathogenic species from the bacterial list actually made it into the top-5 (first place among bacteria) with 95 GCFs derived from its 56 genomes. A striking observation from this species data is how little overlap occurred between the BGCs from different strains: 87 out of the 95 (91.5%) GCFs are shared only between less than 20% of strain genomes, meaning that every 11 strains may harbor ~17 unique BGCs that cannot be found in any other strain of the species. Not much can be said about the potential natural products that can be mined from this diversity (two hybrid NRP/Polyketide compounds, an antimycoplasma micacodin [114] and a fungi-colonizing agent ralsolamycin [115] from a tomato-associated strain GMI1000, were deposited in MIBiG under accessions BGC0001014 and BGC0001363/1754), but several comparative genomic analyses [116,117] have linked this highly divergent metabolic capacities with their unusual ability to attack a vast range of plant species [118].

Finally, fungal secondary metabolism presents an enigma in the space of natural product and drug discovery: although some of the most important drugs came from fungi, such as cyclosporine, penicillins and lovastatin, they arguably remain underexplored when compared to the bacteria. Indeed, there are only 88 entries from *Aspergillus* as opposed to 636 from *Streptomyces* in MIBiG 2.0. Similarly, there are around 2,000 streptomycete genomes in NCBI GenBank compared to ~400 from *Aspergillus*. This phenomenon might be attributed to the general difficulty of working with filamentous fungi, due to, e.g., their relatively complex genomes. Nevertheless, many fungal species managed to place themselves onto the list of species with the richest GCF repertoires. As many as 32 ascomycota from seventeen different genera were part of the top-100 ranked species in the *T*=900 list, and despite its lower genome count (410) compared to, e.g., the bacterial pathogen *Pseudomonas aeruginosa* (4,858), *Fusarium oxysporum* managed to top the chart with 821 unique GCFs. Similarly, three *Aspergillus* species have a genome-to-GCF ratio similar to, or in some cases higher than the *Streptomyces* species on the list. As fungi and bacteria seem to frequently compete with each other in the wild [119], it may be logical to increase the search for new antibacterial compounds from this nemesis of bacteria, complementary to bacterial genome mining.

7.4. Conclusions and Future Perspectives

Here, we demonstrated that with BiG-SLiCE, we finally have the means to generate and exploit a truly global map of secondary metabolic diversity, which can provide insights for both fundamental (studying the diversity and evolution of microbial secondary metabolism) and practical (drug and novel compound discovery) purposes. To draw more solid biological conclusions from this kind of analysis, the issue of uneven feature coverage needs to be addressed (leading to some BGCs being more granularly clustered than others at any given threshold) and a more robust approach needs to be designed for choosing a threshold for clustering. For that reason, we currently focused our support for outputs from curation-based tools and databases such as antiSMASH and MIBiG, allowing us to fine-tune BiG-SLiCE's clustering algorithm on well-known and experimentally-validated BGC classes. In the future, we envision that the tool could also incorporate BGCs from other sources, particularly those coming from semi-supervised tools like ClusterFinder [28], EvoMining [58], and DeepBGC [39].

Furthermore, the sub-Pfam approach that we introduced here could have potential use beyond GCF construction. By using it in place of the more generic Pfam models, we can apply a Pfam2Vec analysis, the corpus being a dataset of computationally identified BGCs, to find biosynthetically relevant pairs of co-evolving genes that can be associated to specific chemical moieties [120]. With its improved sensitivity, we can also use sub-Pfam to survey putative antimicrobial resistant gene families across the >1.2 million BGCs in BiG-SLiCE, potentially revealing a wide array of potential antibiotic-producing BGCs using what can be thought as a global target-directed genome mining approach [59,121].

One important topic that has not been discussed extensively is how we can deal with fragmented BGCs. This is especially important when considering incorporation of more MAGs and shotgun metagenomic data in future analyses. Although the fuzzy membership approach provides a way for an objective (manual) inspection of BGC placement, an automatic but statistically-informed placement strategy still needs to be developed (as opposed to taking only the best hit coupled with some arbitrary thresholds as done here). Additionally, implementing a vector-based counterpart of BiG-SCAPE's “glocal” comparison, which matches only the aligned fraction of a complete BGC against a fragmented one (e.g. by only calculating the euclidean distance of shared columns) might help to dampen the effect of the variable feature size each GCF had.

While this first version of the software constitutes a big leap in scalability of BGC analyses, a long road is still ahead. We invite the community to help improve BiG-SLiCE by sending feedback and using it to investigate the many specific questions that they have which were impossible or highly impractical to answer before. Finally, while a similar massive-scale BGC analysis can be performed ad hoc given sufficient computational resource and expertise, we can convert the precalculated global analysis result into a publicly accessible “reference” GCF database (now available online as BiG-FAM database [122]), allowing the scientific community to benefit from the result in new ways. For example, by curating this reference database with structural and functional annotations derived from (known) BGCs, it can facilitate the functional characterization and dereplication of newly sequenced BGCs.

Acknowledgements

We thank Joris Louwen for adding 21 manually selected biosynthetic pfams to the library, Vittorio Tracanna for his constructive feedback on the study, and Jorge C. Navarro-Muñoz for his input on BiG-SCAPE and fungal BGCs.

Conflicts of Interest Statement

M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio.

Data Availability

Input BGCs, analysis results and python scripts used to generate all figures and tables in this study is available via the GigaScience repository, GigaDB [123]. An archived v1.0.0 release of the BiG-SLiCE software including the pHMM models used for this study can be downloaded from Zenodo [124].

Supplementary Material

All supporting information is available online at https://bit.ly/2NwuZ9s.

References

1. Demain AL. Importance of microbial natural products and the need to revitalize their discovery. J Ind Microbiol Biotechnol. 2014;41: 185–201.
2. Tanaka Y, Omura S. Agroactive compounds of microbial origin. Annu Rev Microbiol. 1993;47: 57–87.
3. Barker DJ, Stuckey DC. A review of soluble microbial products (SMP) in wastewater treatment systems. Water Res. 1999;33: 3063–3082.
4. Mukherjee AK, Das K. Microbial Surfactants and Their Potential Applications: An Overview. Biosurfactants. Springer, New York, NY; 2010. pp. 54–64.
5. WHO | No Time to Wait: Securing the future from drug-resistant infections. 2019 [cited 18 Feb 2020]. Available: http://www.who.int/antimicrobial-resistance/interagency-coordination-group/final-report/en/
6. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016;113: 5970–5975.
7. Larsen BB, Miller EC, Rhodes MK, Wiens JJ. Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life. Q Rev Biol. 2017;92: 229–265.
8. Li S, Hu X, Li L, Hu X, Wang J, Hu X, et al. 1-hydroxy-7-oxolavanducyanin and Δ 7″,8″ -6″-hydroxynaphthomevalin from Streptomyces sp. CPCC 203577. J Antibiot . 2020; 1–5.
9. Nguyen HT, Pokhrel AR, Nguyen CT, Pham VTT, Dhakal D, Lim HN, et al. Streptomyces sp. VN1, a producer of diverse metabolites including non-natural furan-type anticancer compound. Sci Rep. 2020;10: 1–14.
10. Sánchez-Hidalgo M, Martín J, Genilloud O. Identification and Heterologous Expression of the Biosynthetic Gene Cluster Encoding the Lasso Peptide Humidimycin, a Caspofungin Activity Potentiator. Antibiotics. 2020;9: 67.
11. Zhao X-L, Wang H, Xue Z-L, Li J-S, Qi H, Zhang H, et al. Two new glutarimide antibiotics from Streptomyces sp. HS-NF-780. J Antibiot . 2019;72: 241–245.
12. Han Y, Wang Y, Yang Y, Chen H. Shellmycin A–D, Novel Bioactive Tetrahydroanthra-γ-Pyrone Antibiotics from Marine Streptomyces sp. Shell-016. Mar Drugs. 2020;18: 58.
13. Yang L, Li X, Wu P, Xue J, Xu L, Li H, et al. Streptovertimycins A–H, new fasamycin-type antibiotics produced by a soil-derived Streptomyces morookaense strain. J Antibiot . 2020; 1–7.
14. Eckburg PB, Gill SR, Costello EK, Hsiao EY, Gopalakrishnan V, Matson V, et al. The Integrative Human Microbiome Project. Nature. 2019;569: 641–648.
15. Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JHM, et al. Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. Science. 2011;332: 1097–1100.
16. Amos GCA, Awakawa T, Tuttle RN, Letzel A-C, Kim MC, Kudo Y, et al. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. Proc Natl Acad Sci U S A. 2017;114: E11121–E11130.
17. Du C, van Wezel GP. Mining for Microbial Gems: Integrating Proteomics in the Postgenomic

Natural Product Discovery Pipeline. Proteomics. 2018;18: e1700332.

18. Rochfort S. Metabolomics reviewed: a new “omics” platform technology for systems biology and implications for natural products research. J Nat Prod. 2005;68: 1813–1820.
19. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87.
20. Christopher T. Walsh MAF. Natural Products Version 2.0: Connecting Genes to Molecules. J Am Chem Soc. 2010;132: 2469.
21. Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. 2009;6: 23–52.
22. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2019;16: 60–68.
23. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
24. Moghaddam JA, Crüsemann M, Alanjary M, Harms H, Dávila-Céspedes A, Blom J, et al. Analysis of the Genome and Metabolome of Marine Myxobacteria Reveals High Potential for Biosynthesis of Novel Specialized Metabolites. Sci Rep. 2018;8: 1–14.
25. Nielsen JC, Grijseels S, Prigent S, Ji B, Dainat J, Nielsen KF, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. Nature Microbiology. 2017;2: 1–9.
26. McClure RA, Goering AW, Ju K-S, Baccile JA, Schroeder FC, Metcalf WW, et al. Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using Metabolite/Gene Cluster Correlations. ACS Chem Biol. 2016;11: 3452–3460.
27. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. ACS Cent Sci. 2016;2: 99–108.
28. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
29. Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol. 2015;22: 460–471.
30. Parkinson EI, Tryon JH, Goering AW, Ju K-S, McClure RA, Kemball JD, et al. Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. ACS Chem Biol. 2018;13: 1029–1037.
31. Cao L, Shcherbin E, Mohimani H. A Metabolome- and Metagenome-Wide Association Network Reveals Microbial Natural Products and Microbial Biotransformation Products from the Human Microbiota. mSystems. 2019;4. doi:10.1128/mSystems.00387-19
32. Olm MR, Bhattacharya N, Crits-Christoph A, Firek BA, Baker R, Song YS, et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria. Science Advances. 2019;5: eaax5727.
33. Carrión VJ, Perez-Jaramillo J, Cordovez V, Tracanna V, de Hollander M, Ruiz-Buck D, et al. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science. 2019;366: 606–612.
34. The long view on sequencing. Nat Biotechnol. 2018;36: 287.
35. Blin K, Pascal Andreu V, de los Santos ELC, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2019;47: D625–D630.
36. Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res. 2020;48: D422–D430.
37. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017;45: W49–W54.
38. Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. Microb Genom. 2019;5. doi:10.1099/mgen.0.000260
39. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47: e110.

40. Papageorgiou L, Eleni P, Raftopoulou S, Mantaiou M, Megalooikonomou V, Vlachakis D. Genomic big data hitting the storage bottleneck. EMBnet J. 2018;24.
41. SQLite Home Page. [cited 27 Jan 2020]. Available: https://www.sqlite.org/index.html
42. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432.
43. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315: 972–976.
44. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31: 651–666.
45. Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. 1996 [cited 27 Jan 2020]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.2504
46. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. Available: http://arxiv.org/abs/1301.3781
47. Viehweger A, Krautwurst S, Parks DH, König B, Marz M. An encoding of genome content for machine learning. Cold Spring Harbor Laboratory. 2019. p. 524280. doi:10.1101/524280
48. Buchan DWA, Jones DT. Learning a functional grammar of protein domains using natural language word embedding techniques. Proteins. 2020;88: 616–624.
49. Caselles-Dupré H, Lesaint F, Royo-Letelier J. Word2vec applied to recommendation: hyperparameters matter. Proceedings of the 12th ACM Conference on Recommender Systems. New York, NY, USA: ACM; 2018. pp. 352–356.
50. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.
51. Alborzi SZ, Devignes M-D, Ritchie DW. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics. 2017;18: 107.
52. Katz L. Manipulation of Modular Polyketide Synthases. Chem Rev. 1997;97: 2557–2576.
53. Chen C, Natale DA, Finn RD, Huang H, Zhang J, Wu CH, et al. Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. PLoS One. 2011;6: e18910.
54. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.
55. Lorbeer B, Kosareva A, Deva B, Softić D, Ruppel P, Küpper A. Variations on the Clustering Algorithm BIRCH. Big Data Research. 2018;11: 44–53.
56. Rosenberg A, Hirschberg J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007. pp. 410–420.
57. Flask. In: Pallets [Internet]. [cited 27 Jan 2020]. Available: https://palletsprojects.com/p/flask/
58. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biol Evol. 2016;8: 1906–1916.
59. Mungan MD, Alanjary M, Blin K, Weber T, Medema MH, Ziemert N. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Res. 2020;48: W546–W552.
60. Krause J, Handayani I, Blin K, Kulik A, Mast Y. Disclosing the Potential of the SARP-Type Regulator PapR2 for the Activation of Antibiotic Gene Clusters in Streptomycetes. Front Microbiol. 2020;11. doi:10.3389/fmicb.2020.00225
61. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology. 2017;2: 1533–1542.
62. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Scientific Data. 2018;5: 1–8.
63. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified sequence catalogue of over 280,000 genomes obtained from the human gut microbiome. bioRxiv. 2019. p. 762682. doi:10.1101/762682
64. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37: 953–961.
65. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020;21: 1–16.

66. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2020;48: D454–D458.

67. Hervé V, Liu P, Dietrich C, Sillam-Dussès D, Stiblik P, Šobotník J, et al. Phylogenomic analysis of 589 metagenome-assembled genomes encompassing all major prokaryotic lineages from the gut of higher termites. PeerJ. 2020;8: e8614.

68. Singleton CM, Petriglieri F, Kristensen JM, Kirkegaard RH, Michaelsen TY, Andersen MH, et al. Connecting structure to function with the recovery of over 1000 high-quality activated sludge metagenome-assembled genomes encoding full-length rRNA genes using long-read sequencing. bioRxiv. 2020. p. 2020.05.12.088096. doi:10.1101/2020.05.12.088096

69. Anderson CL, Fernando SC. Insights into rumen microbial biosynthetic gene cluster diversity through genome-resolved metagenomics. bioRxiv. 2020. p. 2020.05.19.105130. doi:10.1101/2020.05.19.105130

70. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, Dauser S, Schuster N, Walzer C, et al. Large scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. bioRxiv. 2020. p. 2020.06.05.135962. doi:10.1101/2020.06.05.135962

71. Pamela Engelberts J, Robbins SJ, de Goeij JM, Aranda M, Bell SC, Webster NS. Characterization of a sponge microbiome using an integrative genome-centric approach. ISME J. 2020; 1–11.

72. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol. 2020;38: 701–707.

73. Liang R, Lau MCY, Saitta ET, Garvin ZK, Onstott TC. Genome-centric resolution of novel microbial lineages in an excavated Centrosaurus dinosaur fossil bone from the Late Cretaceous of North America. Environmental Microbiome. 2020;15: 4724.

74. Eze MO, Lütgert SA, Neubauer H, Balouri A, Kraft AA, Sieven A, et al. Metagenome Assembly and Metagenome-Assembled Genome Sequences from a Historical Oil Field Located in Wietze, Germany. Microbiol Resour Announc. 2020;9. doi:10.1128/MRA.00333-20

75. Newberry E, Bhandari R, Kemble J, Sikora E, Potnis N. Genome-resolved metagenomics to study co-occurrence patterns and intraspecific heterogeneity among plant pathogen metapopulations. Environ Microbiol. 2020;22: 2693–2708.

76. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0603-3

77. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019;176: 649–662.e20.

78. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568: 505–510.

79. Martínez-Romero E, Rodríguez-Medina N, Beltrán-Rojel M, Silva-Sánchez J, Barrios-Camacho H, Pérez-Rueda E, et al. Genome misclassification of Klebsiella variicola and Klebsiella quasipneumoniae isolated from plants, animals and humans. Salud Pública de México. 2017;60: 56–62.

80. Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol. 2018;68: 2386.

81. Mateo-Estrada V, Graña-Miraglia L, López-Leal G, Castillo-Ramírez S. Phylogenomics Reveals Clear Cases of Misclassification and Genus-Wide Phylogenetic Markers for Acinetobacter. Genome Biol Evol. 2019;11: 2531–2541.

82. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019 [cited 4 Mar 2020]. doi:10.1093/bioinformatics/btz848

83. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics. 2010;11: 538.

84. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9: 5114.

85. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119.

86. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5: e9490.
87. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132.
88. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47: W256–W259.
89. Yeh H-H, Ahuja M, Chiang Y-M, Oakley CE, Moore S, Yoon O, et al. Resistance Gene-Guided Genome Mining: Serial Promoter Exchanges in Aspergillus nidulans Reveal the Biosynthetic Pathway for Fellutamide B, a Proteasome Inhibitor. ACS Chem Biol. 2016;11: 2275–2284.
90. Lebar MD, Cary JW, Majumdar R, Carter-Wientjes CH, Mack BM, Wei Q, et al. Identification and functional analysis of the aspergillic acid gene cluster in Aspergillus flavus. Fungal Genet Biol. 2018;116: 14–23.
91. Cruz Morales P, Barona Gómez F, Ramos Aboites HE. GENETIC SYSTEM FOR PRODUCING A PROTEASES INHIBITOR OF A SMALL PEPTIDE ALDEHYDE TYPE. World Patent. 2016097957, 2016. Available: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016097957
92. Cruz-Morales P, Vijgenboom E, Iruegas-Bocardo F, Girard G, Yáñez-Guerra LA, Ramos-Aboites HE, et al. The genome sequence of Streptomyces lividans 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island. Genome Biol Evol. 2013;5: 1165–1175.
93. Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC Evol Biol. 2010;10: 26.
94. Begley TP, editor. Polyketide Biosynthesis: Fungi. Wiley Encyclopedia of Chemical Biology. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. p. 380.
95. Chen H, Du L. Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. Appl Microbiol Biotechnol. 2016;100: 541–557.
96. Fisch KM. Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. RSC Adv. 2013;3: 18228.
97. Shen B, Cheng Y-Q, Christenson SD, Jiang H, Ju J, Kwon H-J, et al. Polyketide Biosynthesis beyond the Type I, II, and III Polyketide Synthase Paradigms: A Progress Report: Biosynthesis, Biological Activity, and Genetic Engineering. In: Rimando AM, Baerson SR, editors. Polyketides. Washington, DC: American Chemical Society; 2007. pp. 154–166.
98. Liu W, Christenson SD, Standage S, Shen B. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science. 2002;297: 1170–1173.
99. Li X, Lei X, Zhang C, Jiang Z, Shi Y, Wang S, et al. Complete genome sequence of Streptomyces globisporus C-1027, the producer of an enediyne antibiotic lidamycin. J Biotechnol. 2016;222: 9–10.
100. Haft DH, Basu MK. Biological systems discovery in silico: radical S-adenosylmethionine protein families and their target peptides for posttranslational modification. J Bacteriol. 2011;193: 2745–2755.
101. Hudson GA, Burkhart BJ, DiCaprio AJ, Schwalen CJ, Kille B, Pogorelov TV, et al. Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New Cα, Cβ, and Cγ-Linked Thioether-Containing Peptides. J Am Chem Soc. 2019;141: 8228–8238.
102. Chen Y, Yang Y, Ji X, Zhao R, Li G, Gu Y, et al. The SCIFF-derived ranthipeptides participate in quorum sensing in solventogenic clostridia. Biotechnol J. 2020; e2000136.
103. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58: 4169–4182.
104. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai H-C, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol. 2017;13: 470–478.
105. Walker MC, Eslami SM, Hetrick KJ, Ackenhusen SE, Mitchell DA, van der Donk WA. Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. BMC Genomics. 2020;21: 387.
106. Kloosterman AM, Shelton KE, van Wezel GP, Medema MH, Mitchell DA. RRE-Finder: A Genome-Mining Tool for Class-Independent RiPP Discovery. Bioinformatics. bioRxiv; 2020. p. 11734.
107. Baltz RH. Gifted microbes for genome mining and natural product discovery. J Ind Microbiol Biotechnol. 2017;44: 573–588.
108. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural

products provides insights for future discovery trends. Proc Natl Acad Sci U S A. 2017;114: 5601–5606.

109. Park CJ, Smith JT, Andam CP. Horizontal Gene Transfer and Genome Evolution in the Phylum Actinobacteria. In: Villa TG, Viñas M, editors. Horizontal Gene Transfer. Cham: Springer International Publishing; 2019. pp. 155–174.

110. McDonald BR, Currie CR. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. MBio. 2017;8. doi:10.1128/mBio.00644-17

111. Tidjani A-R, Lorenzi J-N, Toussaint M, van Dijk E, Naquin D, Lespinet O, et al. Massive Gene Flux Drives Genome Diversity between Sympatric Streptomyces Conspecifics. MBio. 2019;10. doi:10.1128/mBio.01533-19

112. de Lima Procópio RE, da Silva IR, Martins MK, de Azevedo JL, de Araujo JM. Antibiotics produced by Streptomyces. Braz J Infect Dis. 2012;16: 466–471.

113. Rhodes MW, Kator H, McNabb A, Deshayes C, Reyrat J-M, Brown-Elliott BA, et al. Mycobacterium pseudoshottsii sp. nov., a slowly growing chromogenic species isolated from Chesapeake Bay striped bass (Morone saxatilis). Int J Syst Evol Microbiol. 2005;55: 1139–1147.

114. Kreutzer MF, Kage H, Gebhardt P, Wackler B, Saluz HP, Hoffmeister D, et al. Biosynthesis of a complex yersiniabactin-like natural product via the mic locus in phytopathogen Ralstonia solanacearum. Appl Environ Microbiol. 2011;77: 6117–6124.

115. Spraker JE, Sanchez LM, Lowe TM, Dorrestein PC, Keller NP. Ralstonia solanacearum lipopeptide induces chlamydospore development in fungi and facilitates bacterial entry into fungal tissues. ISME J. 2016;10: 2317–2330.

116. Prior P, Ailloud F, Dalsing BL, Remenant B, Sanchez B, Allen C. Genomic and proteomic evidence supporting the division of the plant pathogen Ralstonia solanacearum into three species. BMC Genomics. 2016;17: 1–11.

117. Remenant B, Coupat-Goutaland B, Guidot A, Cellier G, Wicker E, Allen C, et al. Genomes of three tomato pathogens within the Ralstonia solanacearum species complex reveal significant evolutionary divergence. BMC Genomics. 2010;11: 1–16.

118. Hayward AC. Characteristics of Pseudomonas solanacearum. J Appl Bacteriol. 1964;27: 265–277.

119. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560: 233–237.

120. Del Carratore F, Zych K, Cummings M, Takano E, Medema MH, Breitling R. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. Commun Biol. 2019;2: 83.

121. Almabruk KH, Dinh LK, Philmus B. Self-Resistance of Natural Product Producers: Past, Present, and Future Focusing on Self-Resistant Protein Variants. ACS Chem Biol. 2018;13: 1426–1437.

122. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. 2020 [cited 23 Oct 2020]. doi:10.1093/nar/gkaa812

123. Satria KA, van der Hooft J Justin J, de Dick R, Marnix MH. Supporting data for “BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million Biosynthetic Gene Clusters.” GigaScience Database; 2020. doi:10.5524/100826

124. Kautsar SA. medema-group/bigslice: Version 1.0.0. 2020. doi:10.5281/zenodo.3975432

Chapter 8

BiG-FAM: A New Online Database of Gene Cluster Families

With BiG-SLiCE, we were given the means to perform a global-scale homology analysis of BGCs and programmatically explore its information-rich results. However, facilitating this novel technology to reach the widest possible audience within the scientific community required an openly accessible and easy-to-use interface to present its results. In this chapter, I present BiG-FAM, an online database which, upon its initial release, holds information about 29,955 gene cluster families (GCFs) calculated from the previous BiG-SLiCE analysis of ~1.2 million BGCs. To facilitate easy access to its content, BiG-FAM provides an extensive search and query infrastructure along with straightforward visualization of its data. Finally, to allow direct analyses of newly identified BGCs, users may submit a completed antiSMASH job id and obtain a comprehensive result listing the closest distant relatives of each of the underlying BGCs.

Published as:

Kautsar, S. A., Blin, K., Shaw, S., Weber, T., & Medema, M. H. (2020).
BiG-FAM: the biosynthetic gene cluster families database.
Nucleic Acids Research, 49(D1), D490–D497*.*

8.1. Introduction

Microbial secondary metabolism produces a vast array of natural products (NPs) beneficial not only to the microbes themselves, but sometimes also to humans, for use as, e.g. antibiotics, chemotherapeutics, and crop protecting agents [1,2]. Enzyme-coding genes for these metabolic pathways, as well as genes encoding associated transporters and regulators, are often found physically co-located within a microbial genome, on loci referred to as biosynthetic gene clusters (BGCs). With the increasing availability of bacterial and fungal genome sequences, BGC identification tools like antiSMASH [3] and PRISM [4] have played a critical role in transforming NP discovery into a genome-based endeavor, as they allow the investigation of bioactive compounds a microorganism may produce even if the pathways are not expressed in the lab or when the genomes originate from uncultivated organisms.

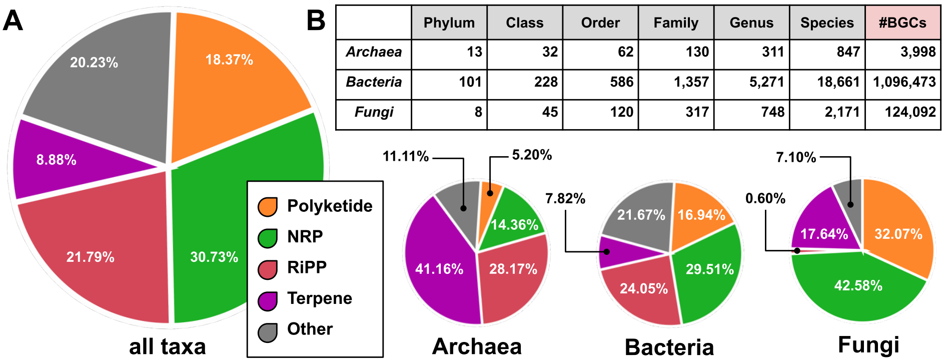
With the simultaneous sequencing of hundreds to thousands of microbial genomes becoming more common, the large quantities of BGC data resulting from this pose an opportunity as well as a challenge. Databases such as antiSMASH-DB (https://antismash-db.secondarymetabolites.org/) [5], IMG-ABC (https://img.jgi.doe.gov/cgi-bin/abc/main.cgi) [6] and MIBiG (https://mibig.secondarymetabolites.org/) [7] have a crucial role in the analysis of BGCs, as they allow comparing the sequences of newly sequenced BGCs against those of previously predicted and experimentally characterized ones [8]. However, while this sequence-based approach works well to identify closely related BGCs across different taxa [9,10], it does not facilitate global analysis of relationships between BGCs across taxa. This is exemplified by the ClusterBlast module in antiSMASH, which, for every detected BGC, outputs a visualized overview of an arbitrary number of top hits for its sequence similarity searches, while the actual number of homologous gene clusters may be smaller or (much) larger.

To identify groups of BGCs that are functionally closely related and encode the production of the same or very similar molecules, approaches have been developed to group BGCs into gene cluster families (GCFs) [10–12]. GCFs have been shown to be very useful in genome-based NP discovery efforts. By examining shared absence/presence patterns of GCFs and compound families (derived via molecular networking of MS/MS spectra [13–15]) across different microbial strains, one can connect BGCs to their expressed products [12,13,16–19]. Algorithms such as BiG-SCAPE [19] automate the GCF reconstruction process and provide detailed interactive sequence similarity networks that can be explored by the users. However, sequence similarity network approaches are not sufficiently scalable to perform global analysis of all BGCs publicly available data. Recently, we developed BiG-SLiCE [20], a tool that makes it possible to perform GCF reconstruction at very large scales in a computationally feasible way. While BiG-SCAPE uses a pairwise comparison strategy to build GCF networks of up to ~70,000 BGCs within ten days in a compute server, BiG-SLiCE applies BGC vectorization coupled with a near-linear clustering algorithm to process more than a million BGCs under the same computational runtime.

Here, we present BiG-FAM, an online database that leverages BiG-SLICE clustering of BGCs to enable GCF-based exploration and homology searching of >1.2 million BGCs harbored by >200,000 microbial genomes. Thus, it provides a complete picture of the “global” secondary metabolic diversity of microbial NPs. Using this web-based platform, scientists can explore the biosynthetic repertoires of specific taxa, investigate the taxonomic and architectural diversity of BGCs of known function, and obtain insights into the novelty of newly sequenced BGCs or their relationships to BGCs in publicly available genomes. For further analysis of the underlying BGCs, BiG-FAM provides cross-links to both the MIBiG and antiSMASH databases, which contain more detailed annotations on the BGCs, such as their predicted or characterized chemical products.

8.2. Database Features

8.2.1. GCF data on over 1.2 million BGCs



	Phylum	Class	Order	Family	Genus	Species	#BGCs
Archaea	13	32	62	130	311	847	3,998
Bacteria	101	228	586	1,357	5,271	18,661	1,096,473
Fungi	8	45	120	317	748	2,171	124,092

Figure 8.1. (A) Pie chart depicting the ratio of five generic BGC classes within the full dataset across three different microbial kingdoms (a total of 1,224,563 BGCs, excluding 16 plant BGCs from MIBiG and 492 BGCs with unassigned taxonomy). (B) Taxa covered by BGCs in BiG-FAM (total number of unique taxa represented by at least one BGC-containing genome per taxonomy level), with the total number of BGCs per kingdom provided in the far-right column of the table.

The BiG-FAM database contains 29,955 GCFs previously calculated using BiG-SLiCE version 1.0 (parameters `--complete-only --threshold 900`) from a collection of 1,225,071 BGCs [20]. These ~1.2 million BGCs were predicted by antiSMASH v5.1.1 from a set of 188,622 microbial genomes (181,521 bacterial, 5,939 fungal and 1,162 archaeal genomes from the NCBI RefSeq/GenBank database) and 20,584 MAGs (from several previously published studies [21–25]), and then complemented with 1,910 experimentally characterized BGCs from the MIBiG 2.0 database [7]. A complete list of all genomes along with their BGC counts and taxonomy according to the Genome Taxonomy Database (GTDB) [26] is provided in Supplementary Table S1. The included BGCs (and their corresponding GCFs) cover a wide range of biosynthetic classes (Figure 8.1A) and taxa (Figure 8.1B), thus providing extensive coverage of the secondary metabolic diversity of the “observable microbial universe”.

8.2.2. Seamless exploration of the database content

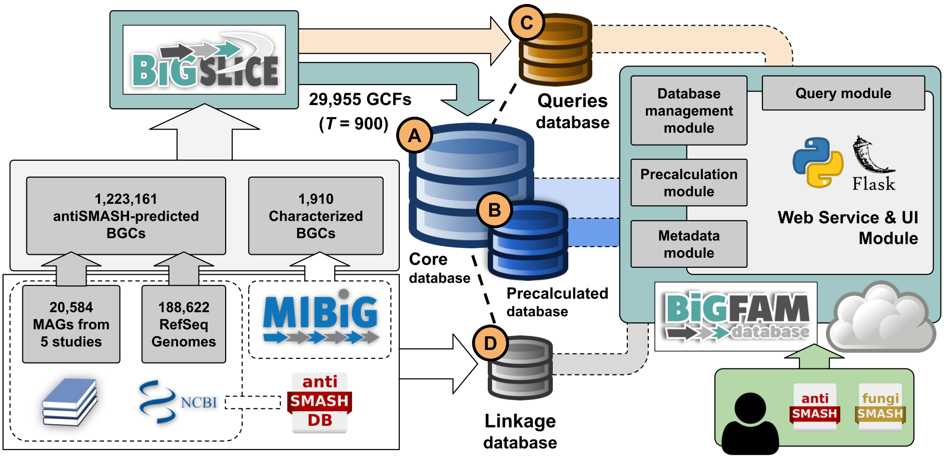


Figure 8.2. Workflow schema of BiG-FAM's architecture. Starting from a collection of ~1.2 million BGCs, BiG-SLiCE was used to perform a clustering analysis (with threshold parameter *T*=900), resulting in (A) 29,955 GCFs stored in an SQLite3 database file. This file is used as the "Core database" for BiG-FAM. To support BiG-FAM's extensive functionalities, three related database files were created, each managed by a specific module in the software package. (B) The "Precalculated database" summarizes complex SQL operations (i.e., calculation of taxonomy counts per GCF) to speed up page loads (detailed schema and procedures can be accessed from the "precalculation" module in BiG-FAM's source code). (C) The "Queries database" stores information related to user-submitted BGC queries, such as processed features from antiSMASH BGCs and the corresponding list of best-matched GCFs identified using BiG-SLiCE. (D) Finally, the "Linkage database" keeps tab on the cross-links to external databases (i.e. MIBiG and antiSMASH-DB), storing information such as the accession number of each linked BGC, which can be used to generate the correct URL addresses pointing to the correct entry within the specific database. These modules and databases were used to serve an online database written in Python using the Flask programming library.

Starting with the core SQLite3-based (https://sqlite.org/index.html) data storage produced by BiG-SLiCE version 1.0.0 (Figure 8.2A), we built a fully functional web server using the Python Flask library (https://palletsprojects.com/p/flask/). We implemented an extra layer of cache storage (Figure 8.2B) to prefetch complex SQL queries coming from various parts of the web server. This, in turn, provides a seamless browsing experience at "compute-heavy" web pages, such as the (database-wide and per-GCF) "Statistics" view. Furthermore, this setup allowed a fairly lightweight (each request is returned within 0.5-5 seconds on average) implementation of the "Search and Filter" function on both BGCs and GCFs, giving users the ability to look for BGCs or GCFs annotated with specific taxa, source dataset types, biosynthetic classes or protein domains.

8.2.3. Querying user-supplied BGCs for rapid GCF placement

One major advantage of using BiG-SLiCE is that the shared BGC features of each GCF are summarized in the same Euclidean-based feature matrix as used for the underlying BGCs, forming what are known as the GCF models (or GCF centroids). This in turn enabled a linear BGC-to-GCF matching, which allows placing dozens of newly sequenced BGCs from a typical microbial genome onto the global map of precalculated biosynthetic diversity within seconds of compute time. To enable easy access to this powerful feature, BiG-FAM incorporates a web-based "Query" submission system, for which the supporting data infrastructure is stored in a separate database (Figure 8.2C), where users can directly take their antiSMASH-predicted BGCs (using job IDs from the antiSMASH web server) and submit them for a GCF analysis. The produced BGC-to-GCF hits will reveal the close (i.e., near-duplicate) and distant relatives of the queried BGCs, which provides insights into their novelty and their relationships to other BGCs and helps studying their distribution and evolution across taxa.

8.2.4. Direct links to BGC and genome databases

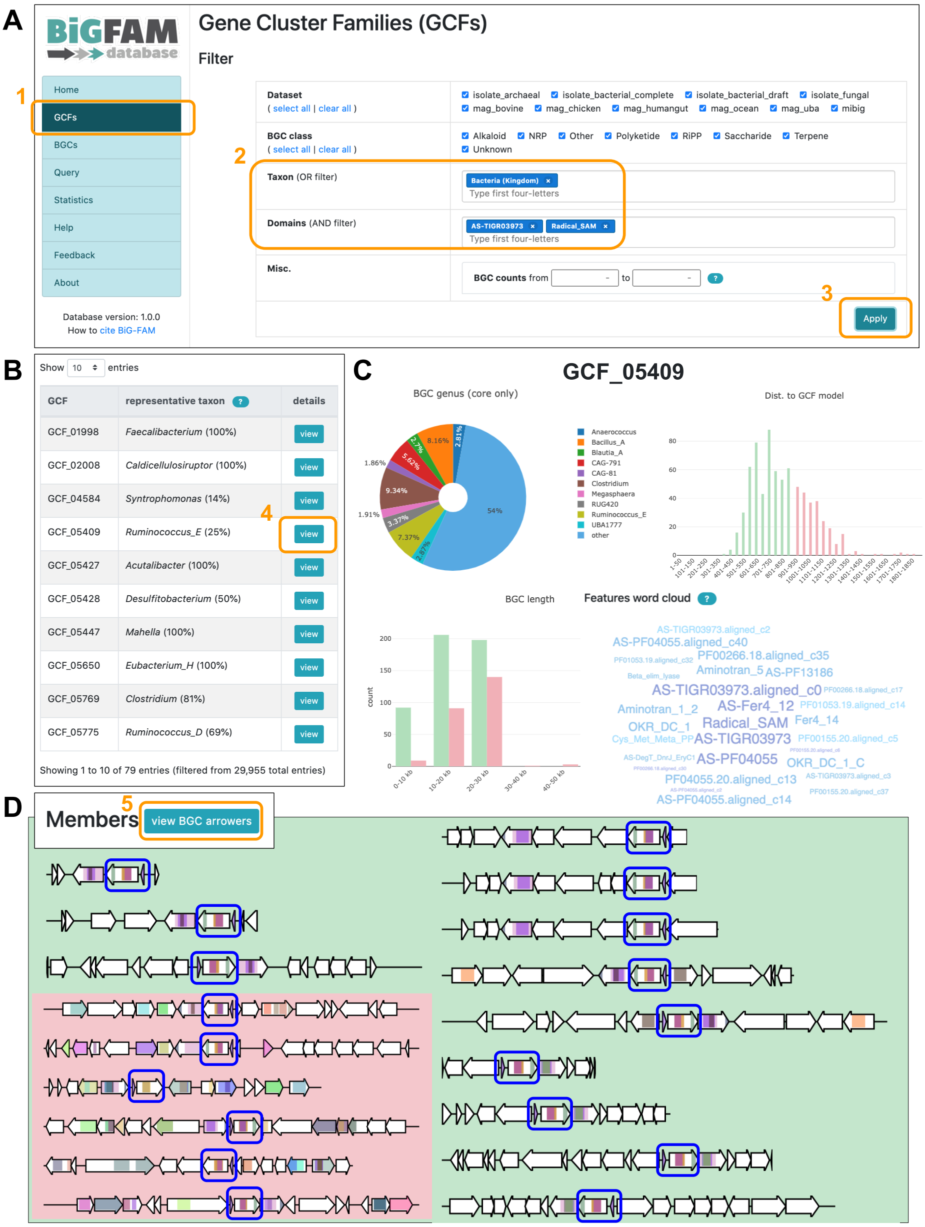
While BiG-FAM stores and displays a lot of useful BGC-related information (such as protein sequences and biosynthetic domain hits), the database was never intended to be a BGC database, and therefore does not include information not directly related to the GCFs, such as nucleotide sequences of the BGCs. To support users who are looking for these data, BiG-FAM stores links to 1,910 known BGCs from MIBiG and 43,117 NCBI-derived BGCs from the antiSMASH database as metadata (Figure 8.2D, Supplementary Table S2). These cross-links can be used to acquire, for example, further information about the characterized (i.e., user-curated information from the MIBiG database) and predicted (for BGCs in antiSMASH database) core structure of the BGC products. Finally, links to the original genome sources (NCBI nucleotide database for isolate datasets, publication's URL for MAG datasets) are also accessible from each BGC's summary page (or as a merged URL list available for download from the GCF page).

8.3. Example Use Cases

There are several ways in which BiG-FAM can be used to answer scientific questions related to microbial secondary metabolism. The "Search and Filter" function can be used to track the distribution of specific groups of BGCs (i.e., based on their generic classes, or the queried combination of their biosynthetic domains) across different taxa. Alternatively, the "Query" page can be used to rapidly match user-supplied BGCs against the set of precalculated GCFs, providing useful information for the characterization and dereplication of those BGCs. Here, we describe two real-world use cases to demonstrate how such analyses could be done by the database users.

8.3.1. Example use case 1: exploring ranthipeptide BGC diversity

Ranthipeptides (previously known as "SCIFF peptides") are ribosomally synthesized and post-translationally modified peptides (RiPPs) prevalent in the taxonomic class *Clostridia* [27], although GC -content analysis indicated that their biosynthetic genes might be horizontally transferred to other taxa as well [28]. Recent analysis shows that these peptides played an important role in regulation at the population level, i.e., via a quorum sensing mechanism [29]. During our previous effort in charting the global diversity of 1.2 million BGCs [20], we captured a large group (6,800) of putative ranthipeptide BGCs with diverse patterns of gene neighborhoods flanking the precursor peptides. To explore this diversity, we can use BiG-FAM's "GCF search" function and use the two signature domains of this BGC class (AS-TIGR03973 and Radical_SAM) as query baits (Figure 8.3A). The search result shows 79 GCFs, each representing a distinct pattern of underlying BGCs distribution across taxonomy (Figure 8.3B). By clicking on the link to each GCF's detail page, information is provided about the taxonomic origins, nucleotide length, calculated radius and biosynthetic features shared by BGCs within the GCF (Figure 8.3C). Additionally, an overview of all BGCs and links to obtain their sequences can be downloaded in TSV format. Furthermore, a comparative multi-gene visualization of the BGCs provides a collated view on the diversity of gene neighborhoods flanking the ranthipeptide precursor genes (Figure 8.3D).



Database version: 1.0.0
How to cite BiG-FAM

Gene Cluster Families (GCFs)

Filter

Dataset (select all | clear all): ☑ isolate_archaeal ☑ isolate_bacterial_complete ☑ isolate_bacterial_draft ☑ isolate_fungal ☑ mag_bovine ☑ mag_chicken ☑ mag_humangut ☑ mag_ocean ☑ mag_uba ☑ mibig

BGC class (select all | clear all): ☑ Alkaloid ☑ NRP ☑ Other ☑ Polyketide ☑ RiPP ☑ Saccharide ☑ Terpene ☑ Unknown

Taxon (OR filter): Bacteria (Kingdom) × Type first four-letters

Domains (AND filter): AS-TIGR03973 × Radical_SAM × Type first four-letters

Misc.: BGC counts from - to -

Apply

Show 10 entries

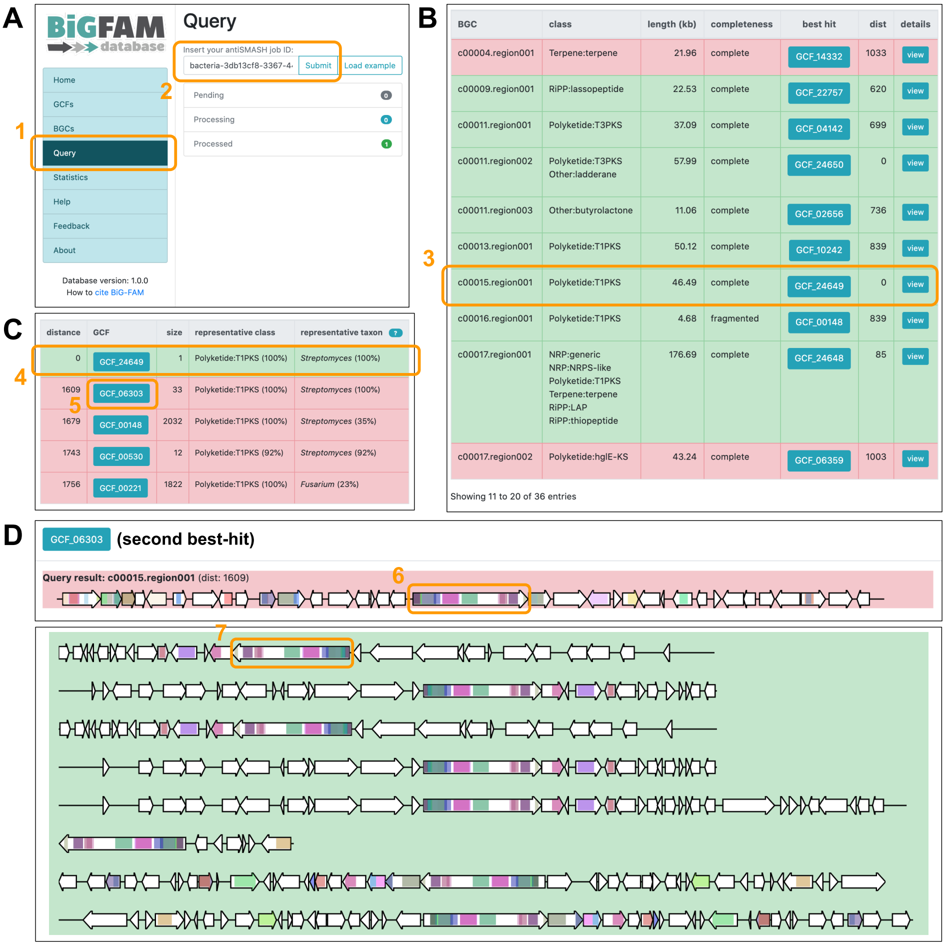
GCF	representative taxon	details
GCF_01998	*Faecalibacterium* (100%)	view
GCF_02008	*Caldicellulosiruptor* (100%)	view
GCF_04584	*Syntrophomonas* (14%)	view
GCF_05409	*Ruminococcus_E* (25%)	view
GCF_05427	*Acutalibacter* (100%)	view
GCF_05428	*Desulfitobacterium* (50%)	view
GCF_05447	*Mahella* (100%)	view
GCF_05650	*Eubacterium_H* (100%)	view
GCF_05769	*Clostridium* (81%)	view
GCF_05775	*Ruminococcus_D* (69%)	view

Showing 1 to 10 of 79 entries (filtered from 29,955 total entries)

Figure 8.3. (A) By clicking on the ‘GCF’ page link (box 1) from the main menu, users will be provided with an interface to search GCFs based on multiple criteria; in this case we search for “bacterial GCFs harboring AS-TIGR03973 and Radical_SAM biosynthetic domains in at least ~80% of their BGCs” (box 2). (B) After applying the filter function (box 3), BiG-FAM returned a list of 79 GCFs satisfying the criteria. (C) Clicking on the “view” button of a GCF (box 4) will take users to a detail page that shows several statistics related to the GCF’s taxonomic distribution, length of its BGCs, and features (domains) distribution. (D) In the GCF detail page, users may also choose to view an “arrower” visualization of the BGCs (box 5), which in this case shows the occurrence of neighboring biosynthetic genes (depicted in colored arrows) flanking the queried cysteine-rich precursor and SAM gene pairs (blue boxes).

8.3.3. Example use case 2: GCF analysis on a newly sequenced Streptomyces strain

Recently, a draft genome was published [30] for *Streptomyces tunisialbus*, a new streptomycete isolated from the rhizospheric soil of lavender plants (*Lavandula officinalis*) in Tunisia [31]. To showcase how BiG-FAM can be used to assess biosynthetic novelty and capture distant relationships of newly sequenced BGCs, we downloaded the assembled genome from ENA (accession: OKRJ01) and uploaded it to the antiSMASH web server (http://antismash.secondarymetabolites.org/), returning a unique job id ("bacteria/fungi-xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx") which (after the run is done) can then directly be used to perform GCF analysis in BiG-FAM (https://bigfam.bioinformatics.nl/query) (Figure 8.4A). The entire analysis for the 36 antiSMASH-predicted BGCs was completed in less than a minute, resulting in a summary table of the best BGC-to-GCF hit pairs (Figure 8.4B). One interesting BGC in this genome is the complete, 46.5 kb long Type-I PKS protocluster from "Region 15.1", which shows an overall low hit rate to gene clusters from public databases in both its ClusterBlast and KnownClusterBlast results (Supplementary Figure S1). A quick look at the GCF analysis result for the BGC shows a significant hit only to one singleton GCF (Figure 8.4C), which after a follow-up inspection turned out to originate from the NCBI-submitted entry of the same genome (accession: GCA_900290435.1). This suggests that the PKS BGC in question represents a novel type of BGC, as it is not closely related to any GCFs with members from other genomes. Relationships to more distantly related GCFs and BGCs can be analyzed by "tracking" of biosynthetic domains of the query BGC across hundreds to thousands of distant BGCs, showing the domain architectural similarity shared between the genes (Figure 8.4D).



BGC	class	length (kb)	completeness	best hit	dist	details
c00004.region001	Terpene:terpene	21.96	complete	GCF_14332	1033	view
c00009.region001	RiPP:lassopeptide	22.53	complete	GCF_22757	620	view
c00011.region001	Polyketide:T3PKS	37.09	complete	GCF_04142	699	view
c00011.region002	Polyketide:T3PKS Other:ladderane	57.99	complete	GCF_24650	0	view
c00011.region003	Other:butyrolactone	11.06	complete	GCF_02656	736	view
c00013.region001	Polyketide:T1PKS	50.12	complete	GCF_10242	839	view
c00015.region001	Polyketide:T1PKS	46.49	complete	GCF_24649	0	view
c00016.region001	Polyketide:T1PKS	4.68	fragmented	GCF_00148	839	view
c00017.region001	NRP:generic NRP:NRPS-like Polyketide:T1PKS Terpene:terpene RiPP:LAP RiPP:thiopeptide	176.69	complete	GCF_24648	85	view
c00017.region002	Polyketide:hglE-KS	43.24	complete	GCF_06359	1003	view

Showing 11 to 20 of 36 entries

distance	GCF	size	representative class	representative taxon ?
0	GCF_24649	1	Polyketide:T1PKS (100%)	*Streptomyces* (100%)
1609	GCF_06303	33	Polyketide:T1PKS (100%)	*Streptomyces* (100%)
1679	GCF_00148	2032	Polyketide:T1PKS (100%)	*Streptomyces* (35%)
1743	GCF_00530	12	Polyketide:T1PKS (92%)	*Streptomyces* (92%)
1756	GCF_00221	1822	Polyketide:T1PKS (100%)	*Fusarium* (23%)

GCF_06303 (second best-hit)

Query result: c00015.region001 (dist: 1609)

Figure 8.4. (A) When users click on the "Query" section of the main menu (box 1), they will be presented with a form to input the job ID of a finished antiSMASH run. After pressing "Submit", BiG-FAM will immediately execute (or put into queue) the downloading, preprocessing and GCF matching of all BGCs (i.e., regions) included in the submitted run. (B) A list will then be shown with the summary of all best BGC-to-GCF pairings with distance lower than 900 (original threshold value) highlighted in green, depicting a good match to at least one GCF in the database. A particular query BGC, "Region 15.1" was selected for a detailed look (box 3) as mentioned in the main text. (C) A list of five best-matching GCFs and their model distances to the query BGC, showing an exact match (d=0) to a singleton GCF from *Streptomyces* (GCF_24649, box 4) which turned out to be the same BGC from the same genome. Looking at the visualization of the second closest GCF on the list (GCF_06303 with d=1609, box 5), we can see (D) co-occurrence of protein domains across the distantly related BGCs, where some similar but non-identical PKS genes (longest multi-domain gene in each GCF) seems to act as an "anchor" that defines the GCF. While this group of anchor genes have a similar domain architecture to the PKS gene of the queried BGC (box 6), a quick BLASTp analysis against one example gene (box 7) shows only 52.63% amino acid identity (Supplementary Text 1). Along with the differences in non-PKS genes between the query BGC and the gene clusters in the GCF, this suggests that, while the BGC is (distantly) related to this GCF, it does not actually belong to it and constitutes a novel gene cluster architecture.

8.4. Discussion

Being the first resource to offer unprecedented access to the “global” biosynthetic space of microbial BGC families, we expect BiG-FAM to become a relevant resource for NP discovery. With its feature-rich web interface, BiG-FAM facilitates user-friendly exploration and querying of its GCFs and BGCs. In the future, “Wikipedia-style” manually curated or semi-automatically generated annotations of precalculated GCFs (e.g. based on the presence of known BGCs or enrichment of signature proteins) may make the database even more useful for end-users. Moreover, with the recent emergence of metadata-rich microbial NP structure databases like the NPAtlas (https://www.npatlas.org/) [32] and databases of mass-spectrometric data like GNPS [33], it may become feasible to perform global meta-analyses to link (taxonomically conserved) BGCs to compounds based on their species/genus level presence/absence patterns observed across these databases [34].

To improve BiG-FAM in subsequent releases, additional useful features for users are planned, such as a REST-based API to support programmatic access to the data and more detailed downloadable summaries (e.g., in a tab-separated text file) that can be used for downstream analyses of the GCFs. Furthermore, there are opportunities to further extend the coverage of precalculated BGCs and GCFs in BiG-FAM by incorporating additional data sources. For example, IMG-ABC currently holds >400,000 BGCs from their >60,000 bacterial genomes, with some degree of overlap against NCBI and MIBiG. There are also other microbial genome databases like MycoCosm [35] that contain genomes not submitted to GenBank or the ENA. Finally, future efforts should also incorporate more data from shotgun metagenomic studies to cover a greater extent of the unculturable microbial biosphere.

Funding

Graduate School for Experimental Plant Sciences (EPS); Novo Nordisk Foundation [NNF10CC1016517 to T.W., K.B., S.S., NNF16OC0021746 to T.W.]; Danish National Research Foundation [DNRF137 to T.W.]. Funding for open access charge: Wageningen University.

Conflicts of Interest Statement

M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio.

Data Availability

The BiG-FAM database is publicly available online and can be accessed without login requirements at https://bigfam.bioinformatics.nl, while the Python script and SQLite schema used to construct the database is available as open source at https://github.com/medema-group/bigfamdb. All data in BiG-FAM are freely available under the Creative Commons CC-BY license.

Supplementary Material

All supporting information is available online at https://bit.ly/3blpzb4.

References

1. Demain AL. Importance of microbial natural products and the need to revitalize their discovery. J Ind Microbiol Biotechnol. 2014;41: 185–201.
2. Vicente MF, Basilio A, Cabello A, Peláez F. Microbial natural products as a source of antifungals. Clin Microbiol Infect. 2003;9: 15–32.
3. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47: W81–W87.
4. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017;45: W49–W54.
5. Blin K, Pascal Andreu V, de Los Santos ELC, Del Carratore F, Lee SY, Medema MH, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2019;47: D625–D630.
6. Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res. 2020;48: D422–D430.
7. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res. 2020;48: D454–D458.
8. Medema MH, Takano E, Breitling R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol. 2013;30: 1218–1223.
9. Doroghazi JR, Metcalf WW. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. BMC Genomics. 2013;14: 611.
10. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc Natl Acad Sci U S A. 2014;111: E1130–9.
11. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
12. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
13. Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, Zhao X, et al. MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci U S A. 2013;110: E2611–20.
14. Winnikoff JR, Glukhov E, Watrous J, Dorrestein PC, Gerwick WH. Quantitative molecular networking to profile marine cyanobacterial metabolomes. J Antibiot . 2014;67: 105–112.
15. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular networking as a dereplication strategy. J Nat Prod. 2013;76: 1686–1699.
16. Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol. 2015;22: 460–471.
17. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. ACS Cent Sci. 2016;2: 99–108.
18. Eldjárn GH, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. Cold Spring Harbor Laboratory. 2020. p. 2020.06.12.148205. doi:10.1101/2020.06.12.148205
19. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020;16: 60–68.
20. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. Gigascience. 2021;10.

doi:10.1093/gigascience/giaa154

21. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2: 1533–1542.
22. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5: 170203.
23. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2021;39: 105–114.
24. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37: 953–961.
25. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020;21: 34.
26. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38: 1079–1086.
27. Haft DH, Basu MK. Biological systems discovery in silico: radical S-adenosylmethionine protein families and their target peptides for posttranslational modification. J Bacteriol. 2011;193: 2745–2755.
28. Hudson GA, Burkhart BJ, DiCaprio AJ, Schwalen CJ, Kille B, Pogorelov TV, et al. Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New Cα, Cβ, and Cγ-Linked Thioether-Containing Peptides. J Am Chem Soc. 2019;141: 8228–8238.
29. Chen Y, Yang Y, Ji X, Zhao R, Li G, Gu Y, et al. The SCIFF-Derived Ranthipeptides Participate in Quorum Sensing in Solventogenic Clostridia. Biotechnol J. 2020;15: e2000136.
30. Ayed A, Wibberg D, Zendah El Euch I, Frese M, Limam F, Sewald N. Draft genome sequence of Streptomyces tunisialbus DSM 105760T. Arch Microbiol. 2020;202: 2013–2017.
31. Ayed A, Slama N, Mankai H, Bachkouel S, ElKahoui S, Tabbene O, et al. Streptomyces tunisialbus sp. nov., a novel Streptomyces species with antimicrobial activity. Antonie Van Leeuwenhoek. 2018;111: 1571–1581.
32. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019;5: 1824–1833.
33. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34: 828–837.
34. van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev. 2020;49: 3297–3314.
35. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42: D699–704.

Chapter 9

Discussion

Here I discuss, in a broader sense, the significance and scientific contribution of the work presented in this thesis. This includes a short synthesis explaining how the various pieces of work complement each other in answering the original question of the thesis: "can we use genomics to map (the biochemistry of) natural product diversity?". Finally, I reflect upon the direction of the field as we move forward in this exciting era of data-generating technological breakthroughs.

9.1. General comments

Over the last decades, data science has transformed and has been integrated into many fields: from business and management [1] to astronomy [2,3]. In biology, the technology-fueled explosion of available whole-genome sequence data marked a new era of bioinformatics and systems biology. As enzyme-coding genes can be thought of as ingredients of a recipe for a biosynthetic pathway, one can imagine using the genome as a reliable source for predicting an organism's metabolic repertoire. With a mountain of genomic data being made available for many antibiotic-producing organisms, data-driven approaches can potentially make a significant impact on natural product (NP)-based drug discovery. This thesis offers various contributions in providing key technologies to generate, store, and process massive amounts of biosynthetic gene cluster (BGC) data. By leveraging more than 200,000 publicly available genomes, it shows the most global map of (microbial) NP diversity thus far.

In ***"Bigger data, better data(bases)"*** I touched upon the topic of generation, storage, and provision of BGC-related data. Using plantiSMASH (**chapter 2**), we identified more than two thousand BGC-like loci from 47 high-quality plant genomes and went on to characterize several of them (**chapter 3**). As more computationally predicted and experimentally discovered BGCs have become publicly available, scalable database platforms (**chapter 4-5**) will be crucial to deal with the incoming data deluge. Then, in ***"Mapping NP diversity through genomics,"*** I addressed the central question raised in the introduction to this thesis: *"how can we leverage (massive-scale) genomic data to map the global NP diversity?"*. Using BiG-SCAPE, we performed a high-resolution clustering of more than seventy thousand BGCs, allowing us to map the biosynthetic diversity within the gene cluster family (GCF) associated with the production of detoxin-rimosamide compound family (**chapter 6**). Later in **chapter 7**, we introduced BiG-SLiCE, which implements a less sensitive but orders of magnitude faster clustering algorithm that can process 1.2 million BGCs within reasonable time and resource constraints. To serve the resulting map of microbial NP diversity produced by the analysis, we built a publicly accessible GCF database, BiG-FAM, which facilitates interactive exploration and user-submitted querying of newly sequenced BGCs (**chapter 8**).

9.2. What does the global analysis tell us?

One main question that this thesis tries to answer with the global-scale analysis was whether or not we have exhausted the NP repertoire of highly-cultivated (micro)organisms. The answer to this question has a far-reaching impact, as it will determine the fate of billions of euros of funding [4] dedicated to tackling antimicrobial resistance. There have been several attempts to answer this question in the recent past. Analyzing the historical record of screened *Streptomyces* isolates and the number of discovered antibiotics around the golden era, Baltz [5] estimated that in order to find a novel broad-spectrum antibiotic, one needs to screen at least ten million (10^6) isolates, which sounds like an intractable effort even for the largest pharmaceutical companies. Lewis [6] further supported this notion and argued that no more broad-spectrum antibiotics can be mined from highly-exploited taxa like Actinobacteria. While he was aware that many actinobacterial species contain a distinctly high number of silent BGCs, he predicted that most antibiotics remaining to be found will be narrow-spectrum and unlikely to work against Gram-negative bacteria.

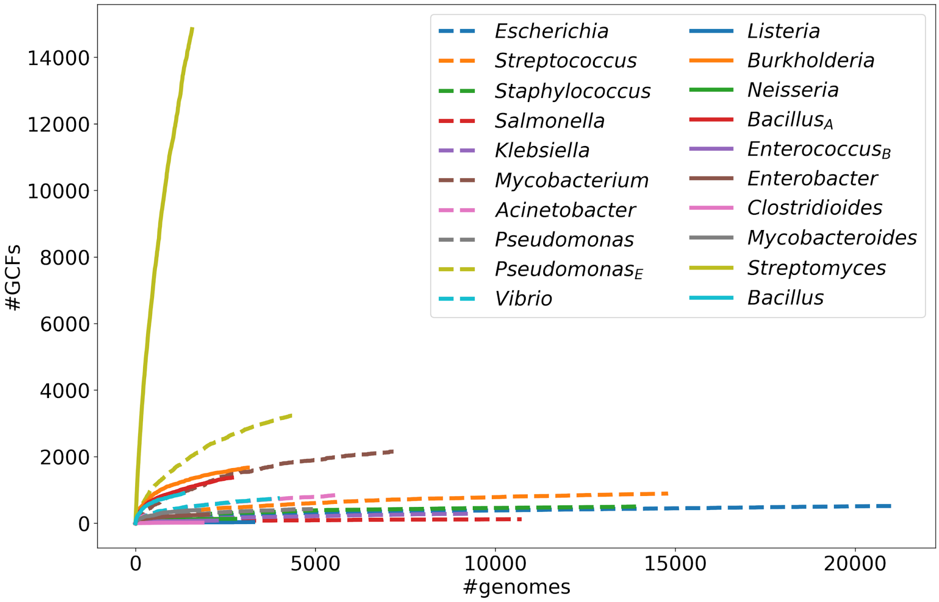


Figure 9.1. Rarefaction curves of twenty biggest genera (i.e., having the most number of available genomes) from the global analysis of 1.2 million BGCs. Here, BiG-SLiCE's clustering result from *T*=300 (producing 121,299 GCFs) was taken and processed using a custom Python script (available via Zenodo [7]).

From the global analysis of 1.2 million BGCs in **chapter 7**, we see a substantial proportion of underexplored biosynthetic diversity, including those from classic NP producers like the Actinobacteria. Taking a rarefaction analysis of 786,465 BGCs harbored by the ten largest genera in the dataset (comprising a total of 118,769 genomes), the potential NP diversity of *Streptomyces* is immense and shows no clear sign of a plateau (Figure 9.1), which corroborates the previous model-based estimate of ~100,000 antimicrobial compounds predicted to be harbored by the genus [3]. While it is hard to tell whether an unknown compound or BGC will show broad-spectrum antibiotic activity (I will briefly discuss this topic in the next subchapter), the sheer breadth of diversity suggests that one should still expect that some, if not many, new promising drug-like compounds can be unearthed from these taxa. With the increasing success of co-culture and genetic engineering techniques in expressing silent BGCs [8–10], this huge trove of uncharacterized BGCs holds promise for finding many new antibiotics classes to fight AMR pathogens. For this, I expect BiG-SLiCE will play a significant role in sifting through the large amounts of BGC data and selecting the most potent strains and BGCs for downstream expression study. In order to do that, however, some challenges and limitations will first need to be addressed.

With BiG-SLiCE, one needs to be aware of its current limitation regarding the algorithm's low sensitivity towards several compound classes (RiPPs, bacteriocin, and highly-modular PKs/NRPs). While this likely would not change the previous conclusion about *Streptomyces'* unrivaled NP diversity, the algorithm will tend to underestimate those poorly-covered classes' chemical diversity when trying to compare the NP potential of different genera, species, or strains. This disparity may lead people to overlook taxa rich in those compound classes, such as the many RiPP-producing commensal bacteria of the human gut [11,12]. An immediate solution would be to manually increase the number of captured sub-Pfam features for the core genes of these classes, which will increase the perceived weight of sequence-level (as opposed to domain-level) diversity for these specific classes within BiG-SLiCE's clustering algorithm. At the same time, some normalization strategies can be implemented in order to perform a more balanced comparison of BGC features. Cosine similarity, which measures the angle between two vectors as opposed to their absolute distance, could theoretically provide that balance. By applying a least-squares normalization [13] to the original feature vectors, one should be able to have an equivalent of a cosine similarity measure using the Euclidean-based clustering algorithm of BiG-SLiCE.

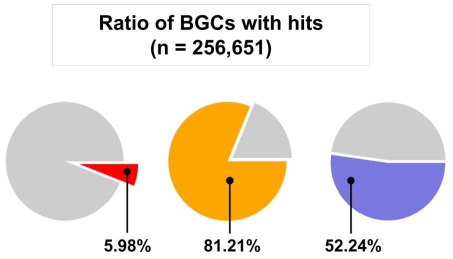
Furthermore, the 1.2 million BGCs processed by BiG-SLiCE in **chapter 7** are still nowhere close to representing the true global diversity of microbial NPs. First, a considerable majority of environmental microbes (some potentially culturable) are not represented by the 14,405 anecdotal metagenome-assembled genomes (MAGs) used in the study. For example, the Joint Genome Institute recently published a catalog of 52,515 MAGs representing 18,028 putative species (~12,000 novel) that inhabit various environments on earth [11]. Second, due to its strict preference towards experimentally proven BGCs, the antiSMASH-based prediction only covers well-known NP classes and might potentially miss other "non-canonical" BGCs. To get truly comprehensive coverage of all microbial NPs, we may need to manually extend antiSMASH's scope or use other alternative BGC prediction methods (I will discuss this more in **subchapter 9.6**). All in all, the aforementioned algorithm modification and coverage extension will likely increase, not decrease, our estimate for the total NP diversity to be mined from cultivable microorganisms on earth. While BiG-SLiCE would theoretically be able to handle input data three to four times larger than the current volume, a follow-up prioritization strategy would need to be put in place to select the most promising GCFs and BGCs for downstream analyses.

9.3. GCF prioritization for drug discovery

An overarching goal of this thesis, and of NP discovery in general, is to find novel compounds that can be used or adapted as drugs to combat untreatable diseases. As our global BGC analysis managed to produce a navigable map in the form of GCFs and their feature vectors, the next challenge is to find ways to prioritize BGCs most promising to produce novel drug-like compounds. Using reference databases like MIBiG, GCFs harboring known BGCs can be studied to find naturally-derived antibiotic analogs that can potentially be active against clinical AMR bacteria. The search for detoxin analogs using BiG-SCAPE in

chapter 6 may serve as a pilot model for this kind of approach. Conversely, large-scale comparison analyses can be performed to highlight unique GCFs (and BGCs) remotely related to any known BGCs. Crusemann et al. [14] recently used this approach to show the uniqueness of two BGCs, one of which came from a cultivated *Chromobacterium vaccinii* strain, that encoded the production of a potent Gq protein inhibitor compound, FR900359.

While the predicted novelty of BGC products may highlight areas with untapped natural product diversity in general, specific prioritization of compounds with actual antimicrobial or cytotoxic activity is mandatory for drug discovery. Previously, screening of BGCs and producer strains that harbor antibiotic resistance genes have been successfully used to unearth a novel proteasome inhibitor from *Salinispora tropica* [15], a new glycopeptide antibiotic from *Streptomyces sp.* WAC1420 [16] and herbicidal compounds from *Aspergillus* [17], among many others [18]. Colloquially termed “target-directed genome mining,” this approach relies upon the fact that many organisms coevolved self-resistance mechanisms to prevent getting self-harmed by their antimicrobial products [18]. Although state-of-the-art computational tools such as ARTS [19,20] and SYN-View [21] can help in identifying self-resistance gene families within a limited set (i.e., dozens to hundreds) of phylogenetically-related strains, a much greater potential lies in the global survey of those genes across all available microbial genomes. In addition to BGCs harboring known AMR gene families such as those documented in the Resfam database [22], there is a vast number of potential self-resistance genes co-located with secondary metabolic BGCs from Actinobacteria (Figure 9.2), representing a potentially important source for antibiotic discovery.



Most enriched pHMMs

Resfam	Resfam (extended)	ARTS
APH3	ABC_tran	TIGR00009
ClassA	Methyltransf_18	TIGR00055
ABC_efflux	MFS_1	TIGR00120
Chlor_Efflux_Pump	MarR_2	TIGR00558
SubclassB1	Lactamase_B	TIGR00043

Figure 9.2. Distribution of gene families with a putative role in self-resistance that are co-located with BGCs from Actinobacteria: (red) known, manually-curated AMR subfamilies from Resfam, (orange) families from Pfam included within Resfam’s “extended” dataset which cover much broader putative AMR genes with the tradeoff of giving a much larger number false-positives, and (purple) families from TIGRFAM included within ARTS’s putative resistant factors, which may be put into a similar category with the Resfam’s extended dataset. In many cases, these putative self-resistance protein families (except for the Resfam’s core set) are populated by a greater number of housekeeping and non-resistance related genes, e.g., the ABC transporters (ABC_tran), a very widespread gene family whose main task is to export the produced NP out of the bacterial cell.

However, as many resistant genes often arose from a duplication event of a corresponding housekeeping gene [23] or protein families related to the export and biosynthesis of BGC products, there is an inherent challenge in identifying them from the hundreds to thousands other genes commonly found within a single bacterial genome. In fact, internally located self-resistance genes might not be as common as one might think for most antibiotic-producing BGCs [24]. A back-of-the-envelope calculation we did for MIBiG BGCs suggests that only 5-10% of known antibiotic BGCs harbored an internal self-resistance gene, such as the BGCs for kanamycin (BGC0000704) [24] and nisin (BGC0000535) [25,26] production. There is a fair chance that a corresponding self-resistance gene (or operon) is located elsewhere in the genome, as is the case for the thiostrepton BGC [25], or that the antiSMASH detection algorithm simply missed some genes that are located just outside its predicted cluster border. Finally, in some cases there is no need for a producer to evolve self-resistance at all, i.e., for many Gram-negative bacteria that produce antibiotics to specifically target Gram-positives [26].

To address this challenge, one could perform a large-scale multi-genome analysis to identify proteins that are associated (e.g., co-evolved) with a specific GCF. Additional evidence such as being an expanded orthologue [27] or being a horizontally transferred gene [19] may further support their assignment as a self-resistance factor. Typically, such an analysis would start with the construction of a sequence similarity network (SSN) [28] of all proteins in the genomes, which requires pairwise comparisons and thus may cause a bottleneck when applied on the scale of tens to hundreds of thousand genomes. In this case, greedy approaches offering significantly less runtime at the expense of accuracy such as CD-HIT [29], UCLUST [30], or Linclust [31] may be used as a promising alternative. An interesting idea would be to use the combined sub-Pfam signatures of the putative protein families (mentioned in Figure 9.2) rather than performing an actual sequence alignment, which may offer further speed up as we have shown in **chapter 7**.

On top of the aforementioned target-directed approach, one should also look for other identifiable criteria to support a BGC's (and the related GCF) potential as a novel drug source. For example, Banfield et al. [32] built machine learning models to identify specific families of transporter genes that can signify the siderophore activity of the product of a BGC (transporter genes were also previously associated with several self-resistance mechanisms in many bacteria [18]). Using a support vector machine (SVM) model trained on 1,281 known BGCs with manually curated activities, the fourth iteration of PRISM (another long-running BGC prediction tool) introduced a new feature that enables the direct prediction of BGC product activity based on either protein domain content or predicted chemical structures [33]. The tool's publication also mentions the calculation of Lipinski's "rule of five" [34] on the predicted BGC products as a way to measure the compound's viability as an orally-available drug. Moore et al. [35] revealed the synergistic antifungal activity of two phylogenetically-conserved BGCs (one of them produces the known immunosuppressant compound rapamycin) in *Streptomyces rapamycinicus*, which opens up a bigger question of finding such a "supercluster" in other genomes and taxa [36]. Finally, highly-amenable universal regulatory mechanisms like the SARP (*Streptomyces antibiotic* *regulatory protein*) gene family may also be used as a screening target for large-scale BGC expression studies [37].

9.4. Understanding the ecological context of BGCs

Aside from their narrower scope as potential drugs, NPs from microbes also serve a more fundamental role in microbial interactions. Plant-associated bacteria has been known to produce a vast array of NPs that can deter fungal and insecticidal pathogens [38,39], stimulate a plant's immune response [40], or promote its growth [41] in exchange for nutrients released by the plant into the soil. Similarly, NPs produced by commensal microbes have been found to be essential for their symbiosis with marine invertebrates [42,43], within the human gut [44,45], and most recently, with insects [46]. By studying the abundance and expression of BGCs across metagenomes, one could better understand microbiome function, which can further be translated to improving agriculture and human health.

Recently, BiG-MAP [47] was developed to study the abundance and expression of BGCs in metagenomic and metatranscriptomic data. To demonstrate the software's functionalities, the authors mapped and analyzed 96 publicly available shotgun metagenomic reads of disease-related and healthy human oral microbiome to a reference dataset of 1,544 representative BGCs (redundancy-filtered from an initial count of 3,352 BGCs using a combination of MASH [48] and BiG-SCAPE) captured from the Human Microbiome Oral Database [49]. Although the analysis did not manage to capture a strong enough signal to confidently assign a BGC onto the microbiome's phenotype, it demonstrated the tool's applicability for use in other, similar analyses. With the continuous drop in sequencing cost, improvements in binning/assembly algorithms, and the emergence of amplicon-free long-read sequencing technologies, a growing number of large-scale shotgun metagenomic studies has been published, covering a wide range of environments: the human gut [50–52], ocean water [53], livestock [54,55], or a meta-collection [11,56]. This combination of tools and available data could provide unprecedented access to study microbial ecology beyond the limitation of marker-based metagenomics.

However, as opposed to the practical use in drug discovery for which the analysis of well-known (even if somewhat remotely related) BGC classes is considered sufficient, full coverage of all potentially relevant biosynthetic pathways is required for ecological studies, as they will give a truly comprehensive overview of the microbiome dynamics. As mentioned before, antiSMASH relies on manually curated signature genes known to be involved in well-characterized biosynthetic pathways, typically from highly-studied taxa like Actinobacteria. This means that, although manual updates can be rolled out as soon as new information becomes available (e.g., with the recent addition of detection rules for ranthipeptides [57], pyrrolidines [58], and lanthidines [59]), the algorithm does not comprehensively cover the biochemical vastness of microbial NPs. Furthermore, small compounds from the specialized primary metabolism of anaerobic bacteria, such as secondary bile acids, were found to be crucial for gut microbiome function and have been directly linked to human health and disease (which have been the focus of the gutSMASH tool used in the aforementioned BiG-MAP study). To move forward, we may need to extend the manual curation approach to cover those overlooked BGC classes. Alternatively, semi-supervised machine learning algorithms may be leveraged [60–63] to capture novel BGC classes with potential ecological relevance. By applying a downstream BiG-SLiCE analysis of BGCs predicted from these tools, we may select for phylogenetically-enriched BGCs and work around the large false-positive discovery rates of these tools. The viability of this concept has been successfully demonstrated, although on a smaller scale, by Barona-Gómez et al. with their EvoMining tool [27], and recently by Medema and Wezel et al. [62] for discovery of novel RiPPs (ribosomally-translated, post-translationally modified peptidic NPs).

9.5. Towards the full utilization of NP databases

There was a major collective effort behind the >70% data volume increase of the MIBiG database in **chapter 4**. The bulk of these data came not only from the accumulated individual submissions by the community, but also from multiple dedicated in-house "annotathons" to manually collect and record experimentally characterized BGCs from the literature. Despite these combined efforts, however, there is a lot of scattered information yet to be included in the database. This particularly holds true for many recently reported BGC characterization experiments, because MIBiG normally relies on the authors' willingness to go an extra mile to manually submit their data into the database. In the ideal scenario, submission of characterized BGC data along with the product's chemical structure information should be enforced as a mandatory requirement for journal publication. For the foreseeable future, it would be worthwhile to consider implementing an automation strategy to track and pre-fill (i.e., via web crawlers [64] in combination with text mining algorithms [65]) the majority of the required information items to be completed by database curators, which will significantly cut the required efforts to input new data into MIBiG.

In **chapter 8**, I introduced BiG-FAM, which was the first and only database to comprehensively cover BGC families harbored by >200,000 publicly available microbial genomes. While the database may have served its purpose to allow data exploration and BGC-based queries of related GCFs, there is still much room for improvement for the database's future updates. The incorporation of a less biased clustering algorithm and an increase in coverage of overlooked NP classes (as described before in **subchapter 9.2**) would be crucial to increase the overall data quality and usefulness of the database. With the size of genome and BGC databases expected to continue growing over time, input data synchronization should also be performed on a regular basis to keep the database relevant for users. As we continue to roll data updates, however, we should ensure that all modifications are properly documented and every release snapshot is provided as a downloadable archive. On the database features side, automated annotation of GCFs using the cumulative information of their known and computationally predicted members would go a long way to providing a richer user experience. Furthermore, we may also implement a user-editable information page (i.e., Wikipedia-style), allowing manual curation of precomputed GCFs. Most importantly, we should make the database programmatically accessible, i.e., by providing an application programming interface (API) to enable true interoperability between BiG-FAM and other related databases.

In order to obtain a complete picture of a biological system, one often needs to combine data from multiple types of omics experiments. For example, transcriptomic analysis has been used alongside genome or exome sequencing to pinpoint causative genes in undiagnosed Mendelian diseases [66,67]. On the NP discovery side, the integration of metabolomics and genomics would allow linking between BGCs and their chemical products, as demonstrated in **chapter 6** and several preceding studies [68–71]. Furthermore, the integration of transcriptomic data has been used to delineate the boundary of many BGCs in *Aspergillus niger [72]*. Ideally, NP-related databases as mentioned in **chapter 5** should be connected to allow such links to be added automatically as more data is stored publicly. Practically, however, it is very challenging to implement such an integrated but decentralized system, of which the problems mainly come down to incompatible data standards (or the lack thereof) and conflicting data distribution policies. The "Minimum Information about any Sequence" (MIxS) framework that the Genomic Standards Consortium implemented back in 2011 was a step in the right direction. In 2014, the FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles were coined [73] to solve this type of problem and has been gaining some traction in the research community. Furthermore, the RDF (Resource Data Framework) specification [74] can be implemented to allow a flexible yet interoperable transfer of metadata across database endpoints. Finally, a more pragmatic route can instead be taken by building a centralized hub that collects, transforms, and links data from multiple sources. The recently published Paired Omics data platform [75] is a perfect example of this approach, although it now still largely relies on manual submissions from the community.

9.6. The future of plant BGC analysis

Previously thought to be limited to bacteria and fungi, increasing evidence suggests that non-random gene organization also happens in eukaryotes [76–78], which supports the arguments for the evolution of plant BGCs [79]. Indeed, while the expression of plant BGCs is not as tightly controlled as bacterial operons [80], recent studies in *Arabidopsis thaliana* by Osbourn et al. suggest that they are still controlled by some sort of regulatory mechanism, most likely at the chromatin level [81,82]. Diving deeper into this hypothesis, the authors used chromosome conformation capture experiments to show that the clustered biosynthetic genes are embedded in a dynamic hotspot of the chromosome with conformational changes between the active (expressed) and inactive states [83]. The interest of the community in studying the evolution of plant BGCs [84–86] and mining them from newly sequenced plant genomes [87,88] may increase. Computational tools like plantiSMASH (**chapter 2**) can play an important role in embracing the new paradigm.

Three years after plantiSMASH was published along with the precomputed BGC prediction of 48 high-quality plant genomes in 2017, at least 278 new plant genomes with contig N50 values > 1MB have been submitted to the NCBI database. This sudden increase of high-quality genome assemblies is particularly driven by the increasing use of long-read sequencing technology from PacBio and Oxford Nanopore Technology [89]. Over this time, plantiSMASH's webserver has processed more than 3,200 job submissions, which reflects the aforementioned increasing interests of the scientific community to learn about plant BGCs. However, by looking at the webserver's log, it becomes clear that a large bulk of these job submissions were spent on redundantly processing publicly available plant genomes from NCBI. While this is rarely an issue for antiSMASH, the much longer runtime (~15-60 minutes) plantiSMASH requires per job means that a lot of resources are wasted. In the future, the development of a precomputed database (similar to the antiSMASH-DB [59] for bacterial and fungal genomes) will allow automated redirection, saving significant time and resources in the long run. Such a database will also allow precomputing publicly available transcriptomics data on a large scale, giving its users a complete insight into the expression of the computationally-predicted BGCs.

On the algorithmic side, plantiSMASH might also benefit from covering a larger set of NP classes. Plant RiPPs , which include the head-to-tail cyclized cyclotides and orbitides [90], constitute one of these promising but currently missed classes. Similar to their bacterial counterparts, plant RiPPs are composed of a highly-conserved leader peptide sequence followed by a hypervariable core sequence [91], the combination of which may be computationally mined using unsupervised machine learning approaches like NeuRiPP [92] and decRiPPter [62] or motif-based tools like RODEO [93], BAGEL, and antiSMASH [94]. In 2019, Kersten and Weng [95] proposed the biosynthetic basis of lyciumins, a seemingly widespread plant RiPP that were initially isolated from the roots of Chinese wolfberry (*Lycium barbarum*). Interestingly, precursor genes of this new plant RiPP class harbor a specific protein domain (commonly known as BURP [96]) alongside some distinctly repetitive precursor motifs, suggesting the possibility for a pattern-based detection rule similar to the likes of lanthipeptides, thiopeptides, and other bacterial RiPPs in antiSMASH.

9.7. Closing remarks

Nearly two decades after the completion of the Human Genome Project, we are still struggling to fulfill the great promise of gaining a complete understanding of an organism just by looking at its genome sequence. Nevertheless, we are now experiencing what is probably one of the most exciting times in the history of biology. Not only do rapid technological advances continue to take place to fuel an exponential increase of available omics data, tools and algorithms in bioinformatics follow suit to tackle the challenges brought by this data deluge. Indeed, although the fields of computer science and software engineering have always been marked by rapid innovation, computational biology has never been far behind since its "birth" in the 1950s.

Following in the great footsteps of many tools, databases, and algorithms that were developed to answer the challenges of their time in the past, this thesis has successfully answered the present challenge by extending upon that accumulated knowledge, methods and data. It is very likely that in the not-so-distant future, many new challenges will surface, requiring us to come up with newer, more sophisticated approaches. In that case, the work presented here would surely lend its shoulder to those next generation of bioinformatics tools and databases.

References

1. Frizzo-Barker J, Chow-White PA, Mozafari M, Ha D. An empirical study of the rise of big data in business scholarship. Int J Inf Manage. 2016;36: 403–413.
2. Hailey CJ, Mori K, Bauer FE, Berkowitz ME, Hong J, Hord BJ. A density cusp of quiescent X-ray binaries in the central parsec of the Galaxy. Nature. 2018;556: 70–73.
3. Zhang Y, Zhao Y. Astronomy in the big data era. Data Sci J. 2015;14: 11.
4. Kelly R, Zoubiane G, Walsh D, Ward R, Goossens H. Public funding for research on antibacterial resistance in the JPIAMR countries, the European Commission, and related European Union agencies: a systematic observational analysis. Lancet Infect Dis. 2016;16: 431–440.
5. Baltz RH. Antimicrobials from actinomycetes: back to the future. Microbe-American Society For Microbiology. 2007;2: 125.
6. Lewis K. The Science of Antibiotic Discovery. Cell. 2020;181: 29–45.
7. Kautsar SA. Phyton scripts to generate Figure 9.1: Rarefaction curves of twenty biggest genera (i.e., having the most number of available genomes) from the global analysis of 1.2 million BGCs. 2021. doi:10.5281/zenodo.4474950
8. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nat Rev Microbiol. 2015;13: 509–523.
9. Nah H-J, Pyeon H-R, Kang S-H, Choi S-S, Kim E-S. Cloning and Heterologous Expression of a Large-sized Natural Product Biosynthetic Gene Cluster in Streptomyces Species. Front Microbiol. 2017;8: 394.
10. Xu W, Klumbys E, Ang EL, Zhao H. Emerging molecular biology tools and strategies for engineering natural product biosynthesis. Metab Eng Commun. 2020;10: e00108.
11. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0718-6
12. Li Y, Rebuffat S. The manifold roles of microbial ribosomal peptide-based natural products in physiology and ecology. J Biol Chem. 2020;295: 34–54.
13. Thompson WJ. Algorithms for normalizing by least squares. Computers in Physics. 1992;6: 386–388.
14. Hermes C, Richarz R, Wirtz DA, Patt J, Hanke W, Kehraus S, et al. Thioesterase-mediated side chain transesterification generates potent Gq signaling inhibitor FR900359. Nat Commun. 2021;12: 144.
15. Kale AJ, McGlinchey RP, Lechner A, Moore BS. Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. ACS Chem Biol. 2011;6: 1257–1264.
16. Thaker MN, Wang W, Spanogiannopoulos P, Waglechner N, King AM, Medina R, et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat Biotechnol. 2013;31: 922–927.
17. Yan Y, Liu Q, Zang X, Yuan S, Bat-Erdene U, Nguyen C, et al. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. Nature. 2018;559: 415–418.
18. O'Neill EC, Schorn M, Larson CB, Millán-Aguiñaga N. Targeted antibiotic discovery through biosynthesis-associated resistance determinants: target directed genome mining. Crit Rev Microbiol. 2019;45: 255–277.
19. Alanjary M, Kronmiller B, Adamek M, Blin K, Weber T, Huson D, et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. Nucleic Acids Res. 2017;45: W42–W48.
20. Mungan MD, Alanjary M, Blin K, Weber T, Medema MH, Ziemert N. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Res. 2020;48: W546–W552.
21. Stahlecker J, Mingyar E, Ziemert N, Mungan MD. SYN-View: A Phylogeny-Based Synteny Exploration Tool for the Identification of Gene Clusters Linked to Antibiotic Resistance. Molecules. 2020;26. doi:10.3390/molecules26010144
22. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance

determinants reveals microbial resistomes cluster by ecology. ISME J. 2015;9: 207–216.
23. Tang X, Li J, Millán-Aguiñaga N, Zhang JJ, O'Neill EC, Ugalde JA, et al. Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. ACS Chem Biol. 2015;10: 2841–2849.
24. Tran PN, Yen M-R, Chiang C-Y, Lin H-C, Chen P-Y. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. Appl Microbiol Biotechnol. 2019;103: 3277–3287.
25. Smith TM, Jiang YF, Shipley P, Floss HG. The thiostrepton-resistance-encoding gene in Streptomyces laurentii is located within a cluster of ribosomal protein operons. Gene. 1995;164: 137–142.
26. Masschelein J, Jenner M, Challis GL. Antibiotics from Gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights. Nat Prod Rep. 2017;34: 712–783.
27. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biol Evol. 2016;8: 1906–1916.
28. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One. 2009;4: e4345.
29. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152.
30. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26: 2460–2461.
31. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9: 2542.
32. Crits-Christoph A, Bhattacharya N, Olm MR, Song YS, Banfield JF. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. Genome Res. 2020. doi:10.1101/gr.268169.120
33. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 2020;11: 6058.
34. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46: 3–26.
35. Mrak P, Krastel P, Pivk Lukančič P, Tao J, Pistorius D, Moore CM. Discovery of the actinoplanic acid pathway in Streptomyces rapamycinicus reveals a genetically conserved synergism with rapamycin. J Biol Chem. 2018;293: 19982–19995.
36. Alanjary M, Medema MH. Mining bacterial genomes to reveal secret synergy. J Biol Chem. 2018;293: 19996–19997.
37. Krause J, Handayani I, Blin K, Kulik A, Mast Y. Disclosing the Potential of the SARP-Type Regulator PapR2 for the Activation of Antibiotic Gene Clusters in Streptomycetes. Front Microbiol. 2020;11: 225.
38. Duke SO. Interaction of Chemical Pesticides and Their Formulation Ingredients with Microbes Associated with Plants and Plant Pests. J Agric Food Chem. 2018;66: 7553–7561.
39. Carrión VJ, Perez-Jaramillo J, Cordovez V, Tracanna V, de Hollander M, Ruiz-Buck D, et al. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science. 2019;366: 606–612.
40. Ryu C-M, Farag MA, Hu C-H, Reddy MS, Kloepper JW, Paré PW. Bacterial volatiles induce systemic resistance in Arabidopsis. Plant Physiol. 2004;134: 1017–1026.
41. Ryu C-M, Farag MA, Hu C-H, Reddy MS, Wei H-X, Paré PW, et al. Bacterial volatiles promote growth in Arabidopsis. Proc Natl Acad Sci U S A. 2003;100: 4927–4932.
42. Taylor MW, Radax R, Steger D, Wagner M. Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. Microbiol Mol Biol Rev. 2007;71: 295–347.
43. McCauley EP, Piña IC, Thompson AD, Bashir K, Weinberg M, Kurz SL, et al. Highlights of marine natural products having parallel scaffolds found from marine-derived bacteria, sponges, and tunicates. J Antibiot . 2020;73: 504–525.
44. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell. 2014;158: 1402–1414.
45. Milshteyn A, Colosimo DA, Brady SF. Accessing Bioactive Natural Products from the Human Microbiome. Cell Host Microbe. 2018;23: 725–736.

46. Beemelmanns C, Guo H, Rischer M, Poulsen M. Natural products from microbes associated with insects. Beilstein J Org Chem. 2016;12: 314–327.
47. Andreu VP, Augustijn HE, van den Berg K, van der Hooft JJJ, Fischbach MA, Medema MH. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. Cold Spring Harbor Laboratory. 2020. p. 2020.12.14.422671. doi:10.1101/2020.12.14.422671
48. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17: 132.
49. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. mSystems. 2018;3. doi:10.1128/mSystems.00187-18
50. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. 2020. doi:10.1038/s41587-020-0603-3
51. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568: 499–504.
52. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, et al. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science. 2019;366. doi:10.1126/science.aax9176
53. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data. 2018;5: 170203.
54. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol. 2019;37: 953–961.
55. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020;21: 34.
56. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2: 1533–1542.
57. Hudson GA, Burkhart BJ, DiCaprio AJ, Schwalen CJ, Kille B, Pogorelov TV, et al. Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New Cα, Cβ, and Cγ-Linked Thioether-Containing Peptides. J Am Chem Soc. 2019;141: 8228–8238.
58. Zheng X, Cheng Q, Yao F, Wang X, Kong L, Cao B, et al. Biosynthesis of the pyrrolidine protein synthesis inhibitor anisomycin involves novel gene ensemble and cryptic biosynthetic steps. Proc Natl Acad Sci U S A. 2017;114: 4135–4140.
59. Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Res. 2021;49: D639–D643.
60. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158: 412–421.
61. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47: e110.
62. Kloosterman AM, Cimermancic P, Elsayed SS, Du C, Hadjithomas M, Donia MS, et al. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. PLoS Biol. 2020;18: e3001026.
63. Kloosterman AM, Shelton KE, van Wezel GP, Medema MH, Mitchell DA. RRE-Finder: a Genome-Mining Tool for Class-Independent RiPP Discovery. mSystems. 2020;5. doi:10.1128/mSystems.00267-20
64. Hernández I, Rivero CR, Ruiz D. Deep Web crawling: a survey. World Wide Web J Biol. 2019;22: 1577–1610.
65. Krallinger M, Rabal O, Lourenço A, Oyarzabal J, Valencia A. Information Retrieval and Text Mining Technologies for Chemistry. Chem Rev. 2017;117: 7673–7761.
66. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9. doi:10.1126/scitranslmed.aal5209
67. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat Commun. 2017;8: 15824.

68. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10: 963–968.
69. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, et al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Monomer. ACS Cent Sci. 2016;2: 99–108.
70. Duncan KR, Crüsemann M, Lechner A, Sarkar A, Li J, Ziemert N, et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol. 2015;22: 460–471.
71. Parkinson EI, Tryon JH, Goering AW, Ju K-S, McClure RA, Kemball JD, et al. Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. ACS Chem Biol. 2018;13: 1029–1037.
72. Kwon MJ, Steiniger C, Cairns TC, Wisecaver JH, Lind A, Pohl C, et al. Beyond the biosynthetic gene cluster paradigm: Genome-wide co-expression networks connect clustered and unclustered transcription factors to secondary metabolic pathways. Cold Spring Harbor Laboratory. 2020. p. 2020.04.15.040477. doi:10.1101/2020.04.15.040477
73. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.
74. Miller E. An introduction to the resource description framework. Bull Am Soc Inf Sci. 2005;25: 15–19.
75. Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, et al. A community resource for paired genomic and metabolomic data mining. Nat Chem Biology. 2021. doi:10.1038/s41589-020-00724-z
76. Pál C, Hurst LD. Evidence for co-evolution of gene order and recombination rate. Nat Genet. 2003;33: 392–395.
77. Hurst LD, Pál C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 2004;5: 299–310.
78. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008;91: 243–248.
79. Boycheva S, Daviet L, Wolfender J-L, Fitzpatrick TB. The rise of operon-like gene clusters in plants. Trends Plant Sci. 2014;19: 447–459.
80. Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, Rokas A. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. Plant Cell. 2017;29: 944–959.
81. Yu N, Nützmann H-W, MacDonald JT, Moore B, Field B, Berriri S, et al. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. Nucleic Acids Res. 2016;44: 2255–2265.
82. Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. Formation of plant metabolic gene clusters within dynamic chromosomal regions. Proc Natl Acad Sci U S A. 2011;108: 16116–16121.
83. Nützmann H-W, Doerr D, Ramírez-Colmenero A, Sotelo-Fonseca JE, Wegel E, Di Stefano M, et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. Proc Natl Acad Sci U S A. 2020;117: 13800–13809.
84. Mao L, Kawaide H, Higuchi T, Chen M, Miyamoto K, Hirata Y, et al. Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. Proc Natl Acad Sci U S A. 2020;117: 12472–12480.
85. Liu Z, Suarez Duran HG, Harnvanichvech Y, Stephenson MJ, Schranz ME, Nelson D, et al. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. New Phytol. 2020;227: 1109–1123.
86. Liu Z, Cheema J, Vigouroux M, Hill L, Reed J, Paajanen P, et al. Formation and diversification of a paradigm biosynthetic gene cluster in plants. Nat Commun. 2020;11: 5354.
87. Kong D, Li S, Smolke CD. Discovery of a previously unknown biosynthetic capacity of naringenin chalcone synthase by heterologous expression of a tomato gene cluster in yeast. Sci Adv. 2020;6. doi:10.1126/sciadv.abd1143
88. Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, et al. The opium poppy genome and morphinan production. Science. 2018;362: 343–347.
89. Li C, Lin F, An D, Wang W, Huang R. Genome Sequencing and Assembly by Long Reads in Plants. Genes . 2017;9. doi:10.3390/genes9010006
90. Luo S, Dong S-H. Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic

RiPP Natural Products. Molecules. 2019;24. doi:10.3390/molecules24081541

91. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30: 108–160.
92. de Los Santos ELC. NeuRiPP: Neural network identification of RiPP precursor peptides. Sci Rep. 2019;9: 13406.
93. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai H-C, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol. 2017;13: 470–478.
94. Russell AH, Truman AW. Genome mining strategies for ribosomally synthesised and post-translationally modified peptides. Comput Struct Biotechnol J. 2020;18: 1838–1851.
95. Kersten RD, Weng J-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. Proc Natl Acad Sci U S A. 2018;115: E10961–E10969.
96. Li Y, Chen X, Chen Z, Cai R, Zhang H, Xiang Y. Identification and Expression Analysis of BURP Domain-Containing Genes in Medicago truncatula. Front Plant Sci. 2016;7: 485.

Summary

Natural products (NPs) from plants and microbes have become an integral part of human civilization, with a distinct contribution in healthcare as antibiotics or other medicinal drugs. With a threatening antibiotic resistance crisis, there is great pressure to mine promising new drugs from nature, which has been postulated to harbor a much greater NP diversity than what has been characterized thus far. One promising way to explore this hidden diversity is through genomics. All living organisms encode metabolic proteins as biosynthetic genes in their genome, and in microbes, genes for each metabolic pathway are often co-located in regions known as biosynthetic gene clusters (BGCs). With the increasing availability of DNA sequencing technologies, bioinformatic detection and analysis of these BGCs have started to play a more and more crucial role in the genomics-based discovery of new NPs. However, new tools and databases are needed to keep up with the speed at which new genomic data is generated.

In **chapter 1**, I introduce the field of NP genomics and the main problem of the thesis and describe several state-of-the-art tools on which this thesis intends to improve. In **chapter 2**, I introduce plantiSMASH (plant Secondary Metabolite Analysis SHell), which can be used to predict and analyze BGC-like regions in assembled plant genomes. Using the tool, we identify more than two thousand candidate BGCs from 47 high-quality plant genomes and provide a way to prioritize them using transcriptomic data. We further demonstrate the usefulness and flexibility of the tool in **chapter 3**, customizing the algorithm to investigate putative sesterterpene synthetases in *Brassicaceae*, encoded by BGC-like gene pairs that each harbor the required prenyltransferase and terpene cyclase domain. We expressed seven of these BGCs in tobacco, revealing fungal-like sesterterpenes with tri-, tetra-, and pentacyclic scaffolds.

In **chapter 4**, I present a major update on the MIBiG (Minimum Information about a Biosynthetic Gene cluster) database for experimentally characterized BGCs. Thanks to community contribution and several dedicated annotation workshops, the update encompasses a >70% increase of manually curated entries, putting the database at a total number of 2,201 BGCs. Moreover, the overall annotation quality has been improved and the infrastructure has been redesigned to allow for a more dynamic and feature-rich user experience. In **chapter 5 I** provide a comprehensive review of publicly available and commercial databases relevant for NP discovery and highlight some of the challenges and opportunities moving forward in the multi-omics era.

In **chapter 6**, I describe a large joint effort in the development and subsequent validation of BiG-SCAPE - CORASON, a first dedicated pipeline to perform similarity networking analysis of ~70,000 BGCs, eventually grouping them into gene cluster families (GCFs) and providing a feature-rich user interface to explore the results. While BiG-SCAPE marked a significant improvement over preceding approaches, it faces a bottleneck when analyzing large sets of BGCs. To tackle

this issue, in **chapter 7** I developed BiG-SLiCE, which implements an ultra-scalable clustering algorithm to perform a clustering analysis of 1.2 million BGCs predicted from ~209,000 genomes within less than ten days runtime, which allowed charting a first (near) global map of microbial NP diversity. Finally, to provide this new resource to the wider scientific community, in **chapter 8** I introduce the BiG-FAM (Biosynthetic Genes FAMilies) database, which offers not only a user-friendly way to explore the precomputed GCFs and BGCs data but also to query the closest matches of user-submitted BGCs.

Chapter 9 concludes this thesis with a brief discussion on the potential impact of my work and an overview of topics in which my work may catalyze further scientific advances in combination with other emerging methods and technologies.

Samenvatting

Moleculen gemaakt door planten en microben leveren een grote bijdrage aan de maatschappij, bijvoorbeeld als antibiotica in de gezondheidszorg. Vanwege de dreigende antibioticaresistentiecrisis is het hard nodig dat er nieuwe medicijnen ontwikkeld worden uit natuurlijke bronnen, die een veel grotere diversiteit van moleculen kunnen aanmaken dan wat tot nu toe is gekarakteriseerd. Een veelbelovende manier om deze verborgen diversiteit te verkennen, is door middel van genomics. Alle levende organismen coderen eiwitten in hun genoom die enzymatische reacties katalyseren om moleculen in elkaar te zetten, en in microben bevinden genen voor elke metabole route zich vaak naast elkaar in regio's die bekend staan als biosynthetische genclusters. Met de toenemende beschikbaarheid van DNA-sequencingtechnologieën, zijn bioinformatische detectie en analyse van deze genclusters een steeds belangrijkere rol gaan spelen in de ontdekking van nieuwe natuurlijke moleculen op basis van genoominformatie. Er zijn echter nieuwe tools en databases nodig om gelijke tred te houden met de snelheid waarmee nieuwe genomische data worden gegenereerd.

In **hoofdstuk 1** introduceer ik het wetenschappelijke veld dat genomische analyse gebruikt om natuurlijke moleculen en hun biosynthese te bestuderen. Ik introduceer voorts het hoofdonderwerp van het proefschrift en ik beschrijf verschillende recent ontwikkelde tools waarop dit proefschrift voortbouwt. In **hoofdstuk 2** introduceer ik plantiSMASH (Plant Secondary Metabolite Analysis SHell), dat kan worden gebruikt om gencluster-achtige regio's in geassembleerde plantengenomen te voorspellen en analyseren. Met behulp van de tool identificeren we meer dan tweeduizend kandidaat-genclusters uit 47 plantengenomen van hoge kwaliteit, en bieden we een manier om deze te prioriteren met behulp van genexpressie-data. We demonstreren het nut en de flexibiliteit van de tool verder in **hoofdstuk 3**, waarbij we het algoritme aanpassen om vermeende sesterterpeen-producerende enzymen in Brassicaceae te onderzoeken, die gecodeerd worden door genomisch geclusterde sets genen en domeinen bevatten die geassocieerd zijn met prenyltransferase- en terpeencyclase-enzymactiviteiten. We brachten zeven van deze genclusters tot expressie in tabak, waarbij schimmelachtige sesterterpenen werden gedetecteerd met tri-, tetra- en pentacyclische scaffolds.

In **hoofdstuk 4** presenteer ik een belangrijke update van de MIBiG-database (Minimum Information about a Biosynthetic Gene cluster) voor experimenteel gekarakteriseerde genclusters. Dankzij bijdragen van de wetenschappelijke gemeenschap en het organiseren van meerdere annotatieworkshops, omvat de update een toename van >70% van handmatig samengestelde inzendingen, waardoor de database op een totaal aantal van 2,201 genclusters komt. Bovendien is de algehele annotatiekwaliteit verbeterd en is de infrastructuur opnieuw ontworpen om een meer dynamische gebruikerservaring mogelijk te maken met meer functies dan voorheen. In **hoofdstuk 5** geef ik een uitgebreid overzicht van openbaar beschikbare en commerciële databases die relevant zijn voor het ontdekken van natuurlijke moleculen en belicht ik enkele van de uitdagingen en kansen die het er liggen voor het veld in een tijdperk waarin steeds vaker meerdere soorten omics-data tegelijkertijd verzameld worden.

In **hoofdstuk 6** beschrijf ik een grote gezamenlijke inspanning bij de ontwikkeling en daaropvolgende validatie van BiG-SCAPE-CORASON, een set nieuwe softwareapplicaties die het mogelijk maakte om netwerkanalyse uit te voeren om de relaties tussen ~70.000 genclusters in kaart te brengen, hen te groeperen in ‘families’ van sterk op elkaar lijkende genclusters en een rijke gebruikersinterface te bieden om de resultaten te verkennen. Hoewel BiG-SCAPE een aanzienlijke verbetering bood ten opzichte van eerder gepubliceerde methoden, zijn er nog wel beperkingen voor het analyseren van zeer grote sets genclusters. Om dit probleem aan te pakken, heb ik in **hoofdstuk 7** BiG-SLiCE ontwikkeld, dat een uiterst schaalbaar clusteralgoritme implementeert om een clusteringanalyse uit te voeren van 1.2 miljoen genclusters uit ~209.000 genomen binnen minder dan tien dagen. Hierdoor wist ik voor het eerst de microbiële diversiteit van biosynthetische genclusters voor de aanmaak van natuurlijke moleculen (bijna) geheel in kaart te brengen. Ten slotte introduceer ik in **hoofdstuk 8** de BiG-FAM-database (Biosynthetic Genes FAMilies) om deze nieuwe bron toegankelijk te maken voor de bredere wetenschappelijke gemeenschap. BiG-FAM biedt niet alleen een gebruiksvriendelijke manier om gegevens over vooraf berekende genclusters en genclusterfamilies te verkennen, maar gebruiker ook vertellen op welke andere genclusters in de database een gencluster waarin hij of zij in geïnteresseerd is het meest op lijkt.

Hoofdstuk 9 sluit dit proefschrift af met een korte bespreking van de mogelijke impact van mijn werk en een overzicht van onderwerpen waarin mijn werk een katalysator kan zijn voor verdere wetenschappelijke vooruitgang, in combinatie met andere nieuwe methoden en technologieën.

Ringkasan

Senyawa alami (lazim disebut *natural products*, atau NPs) yang dihasilkan oleh tumbuhan dan mikroba memiliki peran yang sangat penting di dalam sejarah peradaban manusia. Bagi dunia kesehatan, senyawa-senyawa tersebut berperan penting menjadi bahan langsung dan inspirasi untuk pengembangan obat-obatan khususnya antibiotika. Saat ini, naiknya jumlah patogen yang resisten terhadap mayoritas antibiotik yang beredar di pasaran telah menjadi ancaman nyata bagi masa depan peradaban dunia. Untuk mengatasi krisis ini, ilmuwan dan industri farmasi berlomba untuk kembali mencari kandidat obat dari alam, khususnya mikroba, yang diprediksi masih menyimpan potensi yang sangat besar terutama dengan didukung teknologi *genome sequencing* dan ilmu bioinformatika.

Setiap makhluk hidup memiliki "cetak biru" berupa gen-gen penyandi enzim yang terlibat dalam rangkaian metabolisme primer maupun sekunder untuk menghasilkan ribuan macam senyawa alami. Gen-gen tersebut tersimpan sebagai kombinasi nukleotida (DNA), di mana pada umumnya, gen-gen yang terlibat dalam rangkaian reaksi yang sama dapat ditemukan dalam lokasi yang berdekatan (lebih lanjut disebut *biosynthetic gene clusters* atau BGCs) di dalam genom. Dengan bertambah pesatnya jumlah genom mikroba maupun tanaman yang tersimpan di basis data publik seperti NCBI, analisa dan prediksi BGC kini menjadi ujung tombak bagi pencarian senyawa obat baru berbasis bioinformatika. Walaupun pengembangan kakas bioinformatika untuk identifikasi senyawa baru telah berjalan sejak awal tahun 2000, masih terdapat celah dan kesempatan untuk pengembangan kakas maupun basis data baru terkait analisis dan prediksi BGC, terutama dalam skala besar.

Pada **bab 1**, saya mengakrabkan pembaca kepada topik pencarian senyawa baru berbasis genomik, dilanjutkan dengan memperkenalkan masalah utama yang dicoba diatasi oleh disertasi ini. Selanjutnya, saya memperkenalkan beberapa kakas bioinformatika dan basis data mutakhir, dan bagaimana disertasi ini mencoba menyempurnakan maupun melengkapi kekurangan dari kakas dan basis data tersebut. Pada **bab 2**, saya memperkenalkan *plantiSMASH*, yaitu kakas yang dapat digunakan untuk identifikasi dan prediksi BGC pada genom tumbuhan. Dengan menggunakan *plantiSMASH*, saya dan kolaborator mengidentifikasi lebih dari dua ribu kandidat BGC dari 47 genom tumbuhan serta mengemukakan beberapa ide untuk memprioritaskan BGC yang paling mungkin menyandikan metabolisme sekunder penghasil senyawa alami. Lebih jauh, kami mendemonstrasikan kegunaan *plantiSMASH* di **bab 3** dalam investigasi gen-gen penyandi *sesterterpenoid* di famili *Brassicaceae*. Kami menemukan hampir dua lusin pasangan gen penyandi *prenyltransferase* dan *terpene cyclase*, mengekspresikan tujuh diantaranya ke dalam vektor tembakau dan berhasil mengidentifikasi sekumpulan *sesterterpenes* dengan struktur *tri*-, *tetra*-, dan *pentacyclic*.

Pada **bab 4**, saya menjabarkan pembaharuan menyeluruh yang kami lakukan untuk basis data MIBiG, yang merupakan basis data referensi untuk BGC dan senyawa yang telah tervalidasi lewat eksperimen laboratorium. Berkat partisipasi yang tinggi oleh komunitas ilmuwan di berbagai belahan dunia, ditambah dengan beberapa lokakarya yang kami adakan untuk anotasi data MIBiG, terdapat peningkatan lebih dari 70% data dibandingkan versi sebelumnya (menjadi 2,201 BGC beranotasi). Selain itu, kualitas anotasi secara keseluruhan juga telah kami tingkatkan dan lengkapi dengan pembaharuan infrastruktur *web service* untuk mendukung pengembangan basis data MIBiG di masa depan. Pada **bab 5**, saya menuliskan rangkuman yang komprehensif terkait basis data publik maupun komersial yang relevan bagi usaha pencarian senyawa baru. Bersama kolaborator, kami menjabarkan beberapa tantangan dan kesempatan untuk pengembangan dan anotasi basis data terkait senyawa alami pada era *big data* dan *multi-omics*.

Pada **bab 6**, saya menjabarkan sebuah proyek kolaborasi besar dalam pengembangan dan validasi kakas BiG-SCAPE/CORASON, yang merupakan kakas pertama untuk melakukan analisis kesamaan dan kekerabatan (*clustering*) dari sekitar 70 ribu BGC. Kakas ini mengelompokkan BGC ke dalam *gene cluster families* (GCF), di mana masing-masing GCF merepresentasikan satu jenis reaksi unik yang dapat menghasilkan sebuah struktur senyawa tertentu. Pada konsepnya, analisis ini memungkinkan kita memetakan keberagaman senyawa alami yang dikode oleh lebih dari 1.2 juta BGC yang terdapat dalam sekitar 209 ribu genom mikroba yang tersimpan di basis data NCBI. Sayangnya, algoritma berbasis *networking* yang digunakan oleh BiG-SCAPE memiliki *bottleneck* yang menghalanginya untuk dapat melakukan analisis berskala global seperti itu. Untuk mengatasi keterbatasan ini, pada **bab 7**, saya mengembangkan BiG-SLiCE, yang memanfaatkan algoritma *clustering* alternatif berskala hampir-linier yang memungkinkan analisis keragaman 1.2 juta BGC diselesaikan dalam waktu kurang dari sepuluh hari. Dari analisis inilah, untuk pertama kalinya, kita bisa melihat bagaimana beragamnya senyawa alami yang dapat dihasilkan oleh mikroba-mikroba “kaya” seperti *Actinobacteria*. Selain itu, untuk memfasilitasi eksplorasi terhadap hasil analisis global ini, pada **bab 8** saya memperkenalkan basis data BiG-FAM, yang menawarkan antarmuka yang interaktif dan mudah digunakan untuk mendukung penelitian lebih lanjut terhadap sekitar 29,000 GCF yang telah dipetakan oleh BiG-SLiCE.

Akhirnya, pada **bab 9** saya menutup pembahasan utama dari disertasi ini dengan diskusi pendek tentang dampak dan potensi dari kakas, basis data, maupun hasil analisis baru yang terdapat dalam disertasi ini bagi perkembangan ilmu maupun teknologi terkait pencarian senyawa alami untuk obat-obatan.

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: **Satria Ardhe Kautsar**
Date: **25 May 2021**
Group: **Bioinformatics**
University: **Wageningen University & Research**

1) Start-Up Phase	date	cp
▶ **First presentation of your project**		
A computational framework for rapid discovery and engineering of biosynthetic pathways	2 Feb 2017	1.5
▶ **Writing or rewriting a project proposal**		
▶ **MSc courses**		
Subtotal Start-Up Phase		1.5

2) Scientific Exposure	date	cp
▶ **EPS PhD student days**		
EPS PhD student days 'Get2Gether', Soest (NL)	15-16 Feb 2018	0.6
EPS PhD student days 'Get2Gether', Soest (NL)	11-12 Feb 2019	0.6
▶ **EPS theme symposia**		
EPS theme 4 'Genome Biology', Wageningen University & Research	16 Dec 2016	0.3
EPS theme 4 'Genome Biology', Wageningen University & Research	13 Dec 2019	0.3
EPS theme 3 'Metabolism and Adaptation', Virtual	30 Oct 2020	0.2
▶ **Lunteren Days and other national platforms**		
Annual Meeting 'Experimental Plant Sciences', Lunteren (NL)	10-11 Apr 2017	0.6
Annual Meeting 'Experimental Plant Sciences', Lunteren (NL)	9-10 Apr 2018	0.6
4th Dutch Bioinformatics and Systems Biology Conference (BioSB), Lunteren (NL)	15-16 May 2018	0.6
▶ **Seminars (series), workshops and symposia**		
B-Wise Seminar: Ben Oyserman & Bastian Hornung, Wageningen (NL)	4 Oct 2016	0.2
B-Wise Seminar: Johan Willemsen & Siavash Sheikhizadehanari, Wageningen (NL)	1 Nov 2016	0.2
B-Wise Seminar: Berend Snel & Sander Rodenburg, Wageningen (NL)	6 Dec 2016	0.2
B-Wise Seminar: Judith Risse & Ruben van Heck, Wageningen (NL)	10 Jan 2017	0.2
B-Wise Seminar: Richard Notebaart & Fleur Gawehns-Bruning, Wageningen (NL)	7 Feb 2017	0.2
B-Wise Seminar: Egon Willighagen & Sabrina Simon, Wageningen (NL)	7 Mar 2017	0.2
B-Wise Seminar: Hesham Gibriel & Eduardo Saccenti, Wageningen (NL)	11 Apr 2017	0.2
B-Wise Seminar: Yang Li & Satria Kautsar, Wageningen (NL)	2 May 2017	0.1
B-Wise Seminar: Berend Snel, Wageningen (NL)	6 Jun 2017	0.1
B-Wise Seminar: Jens Allmer & Jesse van Dam, Wageningen (NL)	5 Sep 2017	0.2
B-Wise Seminar: Katy Wolstencroft & Dennis van Muijen, Wageningen (NL)	3 Oct 2017	0.2
B-Wise Seminar: Purva Kulkarni & Twan America, Wageningen (NL)	7 Nov 2017	0.2
B-Wise Seminar: Mathijs Nieuwenhuis & Jorge Navarro, Wageningen (NL)	5 Dec 2017	0.2
B-Wise Seminar: Anton Feenstra & Ehsan Motazedi, Wageningen (NL)	9 Jan 2018	0.2
B-Wise Seminar: Justin van der Hooft & Victor Carrion, Wageningen (NL)	6 Feb 2018	0.2
B-Wise Seminar: Martijn Derks & Rik Kooke, Wageningen (NL)	6 Mar 2018	0.2
B-Wise Seminar: Jeroen de Ridder & Miguel Correa, Wageningen (NL)	3 Apr 2018	0.2
B-Wise Seminar: Sumanth Mutte & Hernando Suarez Duran, Wageningen (NL)	1 May 2018	0.2
B-Wise Seminar: Joana Gonçalves & Jasper Depotter, Wageningen (NL)	5 Jun 2018	0.2
B-Wise Seminar: Gurnoor Singh & Janani Durairaj, Wageningen (NL)	4 Sep 2018	0.2
B-Wise Seminar: Christian Gilissen & Mohammad Alanjary, Wageningen (NL)	2 Oct 2018	0.2
B-Wise Seminar: Erik van den Bergh & Willem Kruijer, Wageningen (NL)	6 Nov 2018	0.2
B-Wise Seminar: Rachel Cavill & Mehmet Akdel, Wageningen (NL)	4 Dec 2018	0.2
B-Wise Seminar: Rik van Rosmalen & Sevgin Demirci, Wageningen (NL)	8 Jan 2019	0.2
B-Wise Seminar: Gerben Hermes & Pariya Behrouzi, Wageningen (NL)	5 Feb 2019	0.2
B-Wise Seminar: Martijn Huijnen & Mark Sterken, Wageningen (NL)	5 Mar 2019	0.2
B-Wise Seminar: Sven Warris & Vittorio Tracanna, Wageningen (NL)	2 Apr 2019	0.2
B-Wise Seminar: Jorge Roel & Victoria Pascal Andreu, Wageningen (NL)	4 Jun 2019	0.2
B-Wise Seminar: Veronika Laine & Raul Wijfjes, Wageningen (NL)	3 Sep 2019	0.2
B-Wise Seminar: Eliana Papoutsoglou & Roeland Voorrips, Wageningen (NL)	1 Oct 2019	0.2
B-Wise Seminar: Jingyuan Fu & Catarina Sales e Santos Loureiro, Wageningen (NL)	5 Nov 2019	0.2
B-Wise Seminar: Simon Rogers & Barbara Terlouw, Wageningen (NL)	3 Dec 2019	0.2
B-Wise Seminar: Mario Calus & Eef Jonkheer, Wageningen (NL)	7 Jan 2020	0.2
B-Wise Seminar: Chaozhi Zheng & Carlos de Lannoy, Wageningen (NL)	4 Feb 2020	0.2
B-Wise Seminar: Age Smilde & Cristina Furlan, Wageningen (NL)	3 Mar 2020	0.2
Indonesia Youth Symposium, Wageningen (NL)	28 Oct 2018	0.2
SCIEX's Virtual Podium 2020 Session 5: Microbiome (Virtual)	24 Apr 2020	0.2
Indonesian Webinar: Role of Bioinformatics in Drug Discovery and Development (Virtual)	8 May 2020	0.2
▶ **Seminar plus**		
▶ **International symposia and congresses**		
Wageningen Indonesian Scientific Expose Symposium, Wageningen (NL)	8 Mar 2017	0.3
DECHEMA European Conference on Natural Products, Frankfurt (DE)	2-5 Sep 2018	0.9
ISMB European Conference on Computational Biology (Virtual)	13-16 Jul 2020	1.2

►	**Presentations**		
	Talk: PlantiSMASH - Tools for the Identification and Prediction of Specialized Metabolite Biosynthetic Gene Clusters in Plants, Wageningen (NL)	16 Dec 2016	1.0
	Talk: The Plant Secondary Metabolite Analysis Shell (PlantiSMASH), Wageningen (NL)	2 May 2017	1.0
	Talk: Charting the Secondary Metabolic Diversity of 209,211 Microbial Genomes and Metagenome-assembled Genomes, ISMB2020 (Virtual)	24 Apr 2020	1.0
	Talk: Pendekatan genomik untuk bioprospecting antibiotik dan antiviral (Virtual)	8 May 2020	1.0
	Talk: Charting the Secondary Metabolic Diversity of 209,211 Microbial Genomes and Metagenome-assembled Genomes (Virtual)	15 Jul 2020	1.0
	Talk: Mapping the Secondary Metabolic Diversity of 209,211 Microbial Genomes and MAGs (introducing BiG-SLiCE: a new, ultra-scalable BGC clustering tool), ActinoBase e-Seminars (Virtual)	23 Jul 2020	1.0
	Poster: MAINFRAME - Computational Framework for Rapid Discovery & Engineering of Biosynthetic Pathways, Wageningen (NL)	8 Mar 2017	1.0
	Poster: From Genes to Compounds, Lunteren (NL)	10-11 Apr 2017	1.0
	Poster: BiG-FAM: the Biosynthetic Gene Cluster Families Database, Frankfurt (DE)	15-16 May 2018	1.0
	Poster: Charting the Secondary Metabolic Diversity of 209,211 Microbial Genomes and Metagenome-assembled Genomes, ISMB2020 (Virtual)	23 Jul 2020	1.0
►	**3rd year interview**		
►	**Excursions**		
	Subtotal Scientific Exposure		23.6

	3) In-Depth Studies	date	cp
►	**Advanced scientific courses & workshops**		
	High Performance Computing (HPC) Basic Course, Wageningen (NL)	7 Jun 2017	0.3
	AntiSMASH v4 "Hackathon" Workshop, Copenhagen (DK)	8-9 Nov 2016	0.6
	AntiSMASH v6 "Hackathon" Workshop, Copenhagen (DK)	10-11 Mar 2020	0.6
►	**Journal club**		
	Bioinformatics Group Bi-weekly Literature Discussion, Wageningen (NL)	2017-2020	3.0
►	**Individual research training**		
	Period abroad at Tilmann Weber Group, DTU Copenhagen (DK)	8-12 Apr 2019	1.5
	Subtotal In-Depth Studies		6.0

	4) Personal Development	date	cp
►	**General skill training courses**		
	EPS Introduction Course, Wageningen (NL)	16 Feb 2017	0.3
	WGS PhD Competence Assessment, Wageningen (NL)	13 Jun 2017	0.3
	WGS PhD Workshop Carousel, Wageningen (NL)	7 Apr 2017	0.3
	Wageningen In'to Languages Improve Your English – a Self-help Guide, Wageningen (NL)	2 Oct 2019	0.1
►	**Organisation of meetings, PhD courses or outreach activities**		
	Technical committee for Wageningen Indonesian Scientific Expose Symposium, Bogor (IDN)	5-6 Jul 2018	1.0
	Scientific committee for Indonesia Youth Symposium, Wageningen (NL)	28 Oct 2018	1.0
►	**Membership of EPS PhD Council**		
	Subtotal Personal Development		3.0

	5) Teaching & Supervision Duties	date	cp
►	**Courses**		
	BIF-30806 Advanced Bioinformatics	2016	1.0
	BIF-30806 Advanced Bioinformatics	2017	1.0
	BIF-30806 Advanced Bioinformatics	2018	1.0
	BIF-30806 Advanced Bioinformatics	2019	0.0
►	**Supervision of BSc/MSc students**		
	MSc project: "Discovery of gene sub-clusters encoding natural product substructures", Joris Louwen	2019	1.0
	MSc project: "Development of a Novel Tool for the Discovery of Plant Cyclopeptides", Floris de Waal	2019	1.0
	BSc project: "Comparative Analysis of Gene Cluster Family (GCF) Calling Algorithms", Erik IJland	2020	1.0
	Subtotal Teaching & Supervision Duties		6.0

TOTAL NUMBER OF CREDIT POINTS*	**40.1**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

Acknowledgments

Having failed my first bachelor's study and almost failed my master's study, it has never occurred to me that I was going to do a Ph.D. (and to actually complete it), especially in one of the best universities in the world. While I did work days and nights to be able to keep up with the workloads and standards to attain this Ph.D. degree, none of this would ever happen only because of my own individual effort. In fact, most of the credits should go to everyone in this acknowledgment section, who played a significant role for me to be able to get through it all while not losing my mind in the process.

First and foremost, I want to humbly thank 'الله سبحان و تعالى, for it is only through His grace and might that I can be where am I today, where I will go tomorrow, and everywhere in between. I always believed that nothing can ever happen except by His will, and through these following people (and many others left unmentioned) that He extended His blessing to me and my family.

Marnix, one-of-a-kind human being who not only was my daily supervisor, but is also my savior, intellectual inspiration, and forever role-model. Thank you for giving this foreign student with an unremarkable academic record an opportunity of his lifetime to ignite his true passion for science and going to places he never imagined before. Thank you for providing me with the best mentoring experience for my Ph.D., patiently guiding and encouraging me in my quest to achieve true scientific independence. As always, I will never be able to put into words how impactful your actions and examples were, and I hope all the best for your future.

I would also thank **Dick**, which has always been so supportive from start to finish, but especially during the latter, time-pressed phase of the thesis writing, upon which he provided tremendous support and rapid feedback that allows me to finish this thesis on time. One of the most valuable lessons I learned from interacting with Dick (and other *Dutch* colleagues) is to be more assertive and mentally-tough during intellectual discussions. This has transformed me from someone who was always unconfident of speaking his mind into someone who can stand his ground for what he strongly believes.

I want to send many thanks to all of our papers' **collaborators**, whose contributions have been instrumental in making all the science happen. **Hernando**. **Anne** and **Ancheng**. **Kai**, **Tilmann** and **Simon**. **Roger** and **Jeff**. **Michael** and **Nelly**. **Justin**, **Jorge**, **Mohammad**, **Victoria**, **Barbara**, **Ben**, **Serina**, along with all other **Medema group members**.

I am also very grateful to have tremendous support from the **Graduate School Experimental Plant Sciences' Ph.D. office**, especially **Susan**, who has been very helpful and supportive throughout this entire process.

I want to send my best regards to all of my thesis committees: **Eric Schranz**, **Jeroen de Ridder**, **Gilles van Wezel**, and **Mohamed Donia**. I deeply thank you for taking the time to thoroughly read my thesis amid your busy schedules, even further to attend my defense as an opponent.

I would also like to thank all **Bioinformatics department members** for all of the scientifically-engaging discussions and all the more casual talks during lunches and annual retreats (how I miss those days!). I want to specifically thank **Vittorio** and **Margi**, who agreed to be my Ph.D. bridesmaids and do all the dirty works for my thesis defense. Both of them have also been my person-in-charge in the group whenever I wanted to discuss science and life in general. In this case, I would also like to mention **Carlos**, who during my final year has been willingly answering my curiosities regarding long-reads sequencing. Finally, it will be ungrateful of me not to mention **Raul**, who has always been the most resourceful person in the group. Without him, the roof would surely fall down upon us.

I want to thank all **Indonesian natural product researchers** with whom I have extensive discussions for the past couple of years. These have given me a new perspective for my future academic career goals after my Ph.D. People in the **MABBI** and **INBIO** whatsapp group, especially **Mbak Tika**, **Pak Kholis**, **Pak Agustinus**, **Bu Ari**, **Bu Ema**, **Mas Danang**, **Mbak Aninditya**, **Mas Rahadian**, **Mbak Riyanti**, **Mas Afif**, **Mas Indra**, **Mas Mada**, **Mas Jekmal**, **Mbak Linda**, **Sausan**, **Mas Didik**, **Pak Wisnu**, **Pak Deden**, and many others. I believe that together, we can take Indonesian natural product science into the world stage.

Looking back, I want to also send my deepest thanks (and apology) to everyone at the **Electrical Engineering Department, University of Lampung**, for having to leave my post to pursue this one-of-a-lifetime opportunity of mine. I especially want to mention **Bu Yessi** and **Pak Ardian**, both of them who humbly understood my decision to follow what I think is best for me and my family. I also wanted to personally thank **Om Admi** for all the support during my stay at Unila. Further, I would also mention **Pak Saiful**, my M.Sc. mentor in the **Bandung Institute of Technology**, who pushed me to finish my thesis and not to repeat the same mistake I did during my bachelor days.

Not to forget in this journey are my colleague friends and their startups, who were among the first to resuscitate my future through multiple projects they enlisted me into. **Adit**, **Bayu**, and the **Nuesto** team, **Ewa** and the **GITS** team, **Lastiko** and **Ivan**, along with all fellow overnight-fighters from the **ARB project**.

All of that being said, I strongly believe that science only contributes to 50% of the requirements for finishing a Ph.D. For the remaining 50%, you will need consistent mental, emotional, and spiritual support that keeps you going through all the ups and downs of your Ph.D. journey. Therefore, I would close this by thanking all individuals that stand as my closest support circle throughout this journey.

The first on the list, I would address my deepest thanks to the love of my life, **Fitri Susana**. You changed my life, for the better, ever since the first day we met each other. In fact, none of this would ever happen if الله سبحان و تعالى didn't meet us together on that fateful day. We went through a lot during the first five years of our marriage: having our first daughter, moving to the Netherlands, trying to finish our studies, having our second son, and finally starting our new life in the United States. Throughout all these changes and circumstances, I could always depend on you as my partner in crime and as my shoulder to lean on whenever life throws its sourest lemons at me. I am always 100% grateful to have you, **Maitra**, and **Yusuf** in my life. You all gave me purpose, motivation, and support to put my all in chasing our dreams in life.

Of course, I would also thank **my parents** for everything. **Papa** dan **Mama**, terima kasih atas semua yang telah Papa Mama berikan: kasih sayang, dukungan moril maupun materil, serta pelajaran berharga tentang hidup dan kehidupan. Terima kasih telah membesarkan Ardhe dengan penuh kesabaran, atas segala kekurangan dan kekecewaan yang telah Ardhe berikan kepada Papa dan Mama selama ini. Setelah merasakan menjadi orang tua, Ardhe sekarang benar-benar paham bagaimana sulitnya dan luar biasanya kasih sayang orang tua dalam membesarkan anak-anaknya. Sampai akhir hayat pun kami tidak akan pernah bisa membalas semua yang telah kalian berikan. In this case, I want to also extend my deepest thanks to my little sister, **Putri**, and brother, **Adin**, for all the support and understanding of my situation. I’m sorry that I can’t be physically present to hold the fort of our family, and I thank you all for all the unconditional love you give to Maitra, Yusuf, and also Fitri. I love you all to the moon and back.

I also wanted to send all the love to my late grandpa and grandma. **Akas** and **Ombay**, I really wish you could see me from up there, making both of you proud.

Finally, living in a foreign country is never easy for everyone. Thankfully, we got tremendous support from all fellow **Indonesians living in Wageningen**. **Mas Anto** and **Mbak Nila**, **Mas Pebri** and **Mbak Ita**, **Mas Mugni** and **Mbak Dian**, **Mas Renato** and **Mbak Ningrum**, **Mas Iman** and **Mbak Tia**, **Mas Darmanto** and **Mbak Nadya**, **Uda Zukri** dan **Uni Elie**, **Pak Dikky** and **Bu Aulia**, **Mas Alim** and **Mbak Ayu**, **Mas Ayusta** and **Mbak Rovan**, **Mas Hardi** and **Mbak Eva**, **Mas Sahri** and **Mbak Ami**, **Mas Dani** and **Mbak Rika**, **Mas Fahmi** and **Mbak Metta**, **Margi**, **Iqbal** and **Putri**, **Rizal** and **Isna**, **Pak Eko** and **Bu Andra**, **Mas Fariz** and **Mbak Vivi**, **Mbak Saffiera**, along with many others that I failed to mention individually. All of you definitely make our home feel like it's 11,333 kilometers closer. I would also like to give a special mention to our Japanese friends, **Nori** and **Yuko**, for the fun, memorable few months that we had together back in Wageningen.

Last but not least, I want to send my deepest gratitude and love to our closest familial support circle in the Netherlands: **Opa Eddy**, **Oma Brenda**, **Bryan** and **Sanne**, **Elaine,** and **Marten**. Thank you for all the warmth and love that you extend to us during our stay in the Netherlands. I hope for good health for every one of you, and I hope that we can meet each other again soon. We missed you all.

Jupiter, 22 March 2021

The research described in this thesis was financially supported by the Graduate School Experimental Plant Sciences.

Cover design by Satria A Kautsar
Printed by Proefschriften, Ede.