GigaScience, 10, 2021, 1–17

doi: 10.1093/gigascience/giaa154
TECHNICAL NOTE

TECHNICAL NOTE

BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters

Satria A. Kautsar, Justin J. J. van der Hooft, Dick de Ridder and Marnix H. Medema*

Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB, Wageningen, The Netherlands

*Correspondence address. Marnix H. Medema, Bioinformatics Group, Wageningen University, Droevendaalsesteeg 1, 6708PB, Wageningen, The Netherlands. Tel: +31317484706; E-mail: marnix.medema@wur.nl http://orcid.org/0000-0002-2191-2821

Abstract

Background: Genome mining for biosynthetic gene clusters (BGCs) has become an integral part of natural product discovery. The >200,000 microbial genomes now publicly available hold information on abundant novel chemistry. One way to navigate this vast genomic diversity is through comparative analysis of homologous BGCs, which allows identification of cross-species patterns that can be matched to the presence of metabolites or biological activities. However, current tools are hindered by a bottleneck caused by the expensive network-based approach used to group these BGCs into gene cluster families (GCFs). **Results:** Here, we introduce BiG-SLiCE, a tool designed to cluster massive numbers of BGCs. By representing them in Euclidean space, BiG-SLiCE can group BGCs into GCFs in a non-pairwise, near-linear fashion. We used BiG-SLiCE to analyze 1,225,071 BGCs collected from 209,206 publicly available microbial genomes and metagenome-assembled genomes within 10 days on a typical 36-core CPU server. We demonstrate the utility of such analyses by reconstructing a global map of secondary metabolic diversity across taxonomy to identify uncharted biosynthetic potential. BiG-SLiCE also provides a "query mode" that can efficiently place newly sequenced BGCs into previously computed GCFs, plus a powerful output visualization engine that facilitates user-friendly data exploration. **Conclusions:** BiG-SLiCE opens up new possibilities to accelerate natural product discovery and offers a first step towards constructing a global and searchable interconnected network of BGCs. As more genomes are sequenced from understudied taxa, more information can be mined to highlight their potentially novel chemistry. BiG-SLiCE is available via https://github.com/medema-group/bigslice.

Keywords: biosynthetic gene cluster; gene cluster family; biosynthetic diversity; natural product discovery; microbial genomics; clustering analysis

Background

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 have 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 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 the chances to achieve future medicinal breakthroughs. More fundamentally, by learning about microbes and the compounds that 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 dis-

Received: 26 August 2020; **Revised:** 29 October 2020; **Accepted:** 29 November 2020

ease [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]. Alternatively, 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, 25]. BGCs responsible for the production of retimycin A [27], tambromycin [25], 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 the ability of soil to suppress 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 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, 24] 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 2 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 that they hold similar average numbers of BGCs, the ∼180,000 bacterial genomes in the NCBI RefSeq database [37] may yield >1,000,000 BGCs when processed with tools like antiSMASH [19] (or other BGC prediction tools like PRISM [38], EvoMining [39], ClusterFinder [24], and DeepBGC [40]).

To handle a dataset this large, even the currently fastest tool (a tool that 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 technologies continue to improve [41]. Here, we introduce BiG-SLiCE (Biosynthetic Genes Super-Linear Clustering Engine), which projects BGCs into Euclidean space to enable the use of a partitional clustering algorithm running in a near-linear [$\sim O(n)$] time complexity. Using this approach facilitates analyzing large datasets of BGCs orders of magnitude faster, finally allowing truly global GCF analyses on all available microbial genomes.

Methods and Implementation

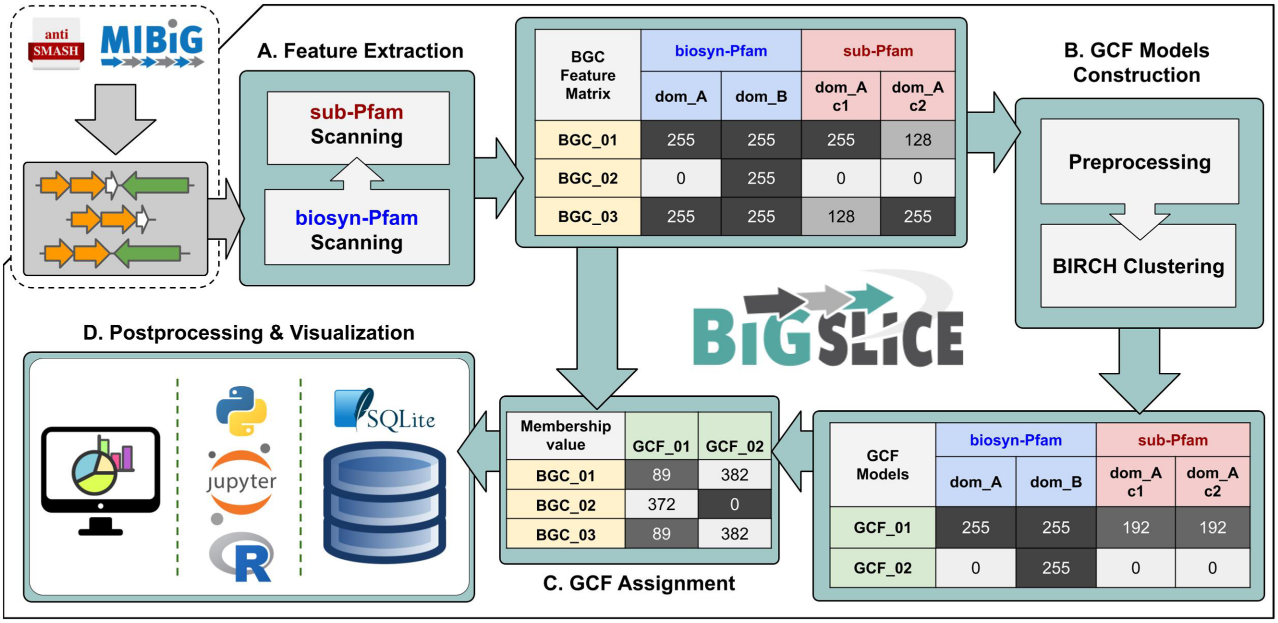
The BiG-SLiCE workflow starts at the vectorization (feature extraction) step (Fig. 1A), converting input BGCs into vectors of numerical features based on the absence/presence and bitscores of hits obtained from querying BGC gene sequences against a library of curated profile hidden Markov models (pHMMs). Those features are then processed by a super-linear clustering algorithm (Fig. 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 (Fig. 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 (Fig. 1D).

BGC feature extraction

In BiG-SCAPE, the (shared) occurrence and synteny (order) of Pfam [43] 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 [44]. 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 [45] or BIRCH [46], it is necessary to transform BGCs into numerical feature vectors (commonly known as quantization or vectorization).

Several approaches have been previously developed to perform (multi-)protein vectorization by adapting the Word2Vec [47] natural language processing algorithm. ProtVec applies an n-mer amino acid residue embedding to model sequence identity as a continuous multi-dimensional vector that can be used for protein family classification. While theoretically it is possible to aggregate features from multiple proteins to generate a BGC vector, its applicability might be limited because 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 [48], assign putative functions to unknown Pfam domains [49], and predict new BGC classes [40]. However, the construction of these models typically involves extensive hyperparameter tuning [48, 50], 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).



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 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 [42]), which can then be used **(D)** to perform further analysis (via external scripts) or to visualize the result in a user-interactive application.

Feature Set 1: biosynthetic domain absence/presence matrix (biosynthetic-Pfam)

Domain hits (retrieved using hmmscan [51] with the gathering threshold) obtained for a reduced list of Pfam version 32 [43] pHMM models (Fig. 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 [52] (which allows filtering for domains related to enzymatic functions) and manual filtering based on each domain's full description (Supplementary Table S1). This filtering was performed 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 genes encoding transporters and regulators). A library of 250 pHMM models from antiSMASH [19] was also included because they include 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 (Fig. 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 interclass diversity. BGCs of the same class typically share a limited set of “core” enzymes that determine the end product's scaffold based on the combination of their specificity and/or copy number variation. For example, the chemical scaffold produced by a Type I polyketide BGC is largely determined by the number and substrate specificities of its modules containing acyltransferase (AT) and ketosynthase (KS) domains [53]. To cover this sequence-level protein diversity, we constructed alignments of 9,451,490 representative protein sequences in the RP15 database (Release 2020_01) [54] to our preselected 293 core biosynthetic domain pHMMs (Supplementary Table S2). We performed hierarchical clustering analysis to group similar aligned sequences into clades and then built sublevel protein family pHMMs from the sequences of each clade (Fig. 2B). This approach resulted in a distinct set of 3,889 sublevel Pfam (sub-Pfam) models (10–100 clades per core domain). In BiG-SLiCE, 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—e.g., 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 × 2/3), and A-c2 = 85 (255 × 1/3). 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 (Fig. 2D).

GCF models construction

To efficiently group BGC features into GCFs, BiG-SLiCE uses a clustering method based on the Python scikit-learn [55] implementation of the BIRCH [46] 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 the “–complete” parameter). The next step in the pipeline is a distance sampling test 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 Xth-percentile (default $X = 1$) of Euclidean pairwise distances between 100 ×

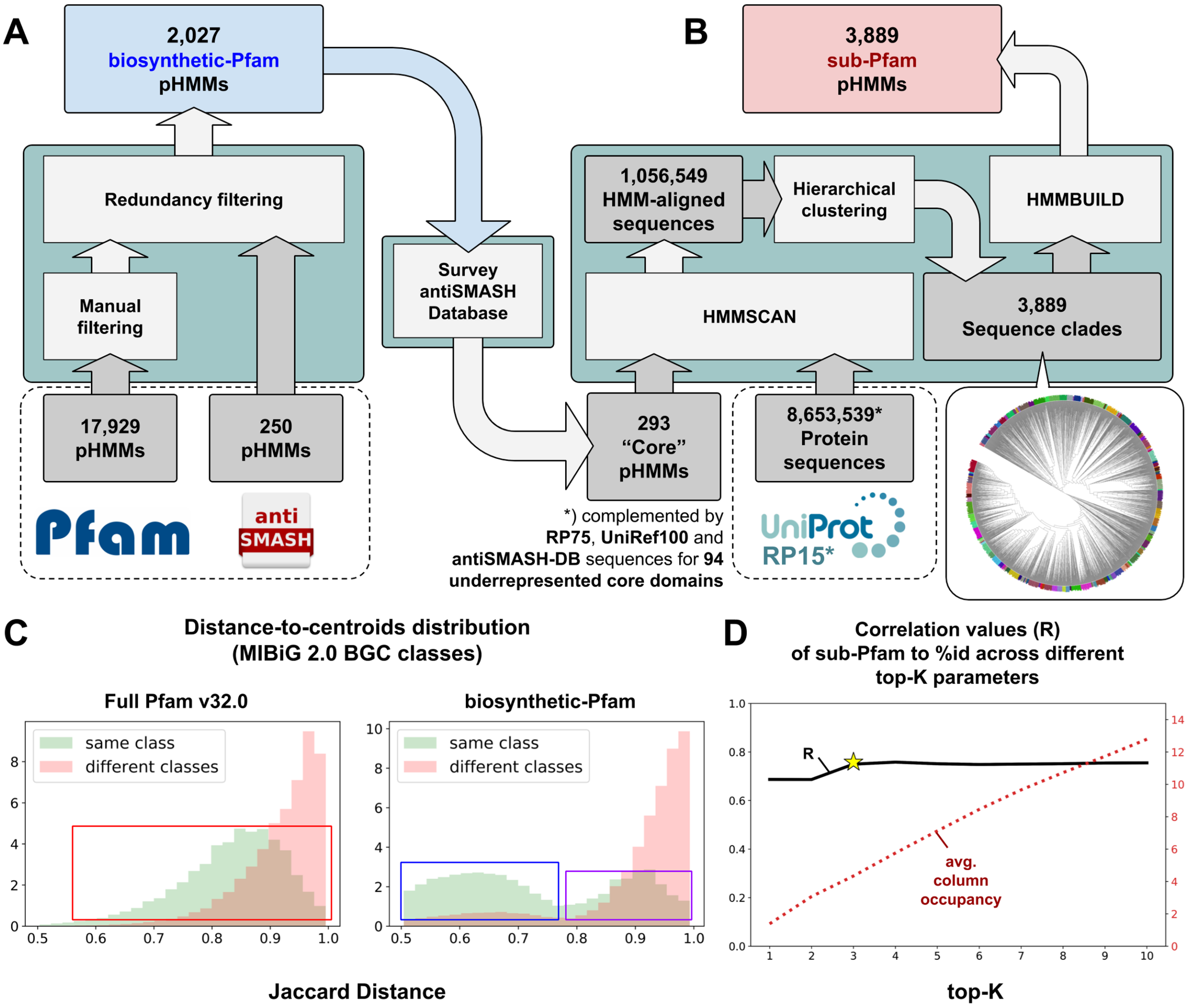


Figure 2: **(A)** Construction of biosynthetic-Pfam features and **(B)** sublevel Pfam (sub-Pfam) features. **(C)** Effect of Pfam model filtering on the discriminatory power of domain-presence Jaccard distance (JI index in BiG-SCAPE) measurements to separate MIBiG v2.0 generic classes (polyketide, NRP, 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 ≥2 distinct classes (e.g., NRPS-PKS, PKS-terpene-saccharide). **(D)** Pearson correlation values between protein sequence similarity (percent 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 (mean 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 (red dotted 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).

1,000 randomly sampled features from the input data. Afterwards, a flat-tree BIRCH (branching_factor > = n_samples) [56] 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 are 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.

Comparison against manually curated GCFs

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 article. Several different threshold parameters T were tested (300–1,500) and corresponding results were compared to the reference groups. We calculated the V-score [57] 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 1 cluster) of a clustering result when matched to a (manually defined) target reference (Fig. 3A), and plotted it alongside the difference of GCF counts (ΔGCF) 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 (Fig. 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; Supplementary Fig. S1) due to its Affinity Propagation [44] 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. BiG-SLiCE, however, 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 distances between BGCs within vs between the curated groups (Fig. 3C) and the visualized feature heat map of the most challenging groups (Fig. 3D).

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 Fig. S2) implemented as a file-based SQLite data store [42]. Processed input (including all metadata), supporting data, and clustering results are systematically stored in the database tables. Using this set-up, 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 (Fig. 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) (Fig. 4B) to perform even more complex analyses, e.g., to study the diversity of biosynthetic domains across samples and across taxonomy (Fig. 4C). As a matter of fact, all analyses performed in this study (see Results and Discussion) 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 particularly 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 an SQLite database with a mini web application written using Python's Flask library [58]. This allowed us to implement a feature-rich visualization "software" that can be deployed and run with minimal 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, 59], or even incorporating additional useful information such as the presence/absence of antibiotic-resistant genes [60] and regulatory domains [61] 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 <5,000 bp ("–minlength 5000"), and used the respective taxonomy options wherever applicable ("–taxon bacteria" for bacterial and archaeal genomes and "–taxon fungi" for fungal ones).

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 S1). Furthermore, we collected and processed 20,584 MAGs from previously published studies [62–66], 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 (e.g., there are many other notable recent publications [64, 67–77] not covered by this initial effort, not to mention the huge number of shotgun metagenomic studies publishing only contig-level assemblies of unassigned bins), the ∼20,000 MAGs presented here may already give us a glimpse of the untapped biosynthetic diversity of uncultured microbes. Finally, we incorporated all 1,910 entries from MIBiG v2.0 [78] 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 reported in Table 1.

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 Genome Taxonomy Database (GTDB), which was 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) 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

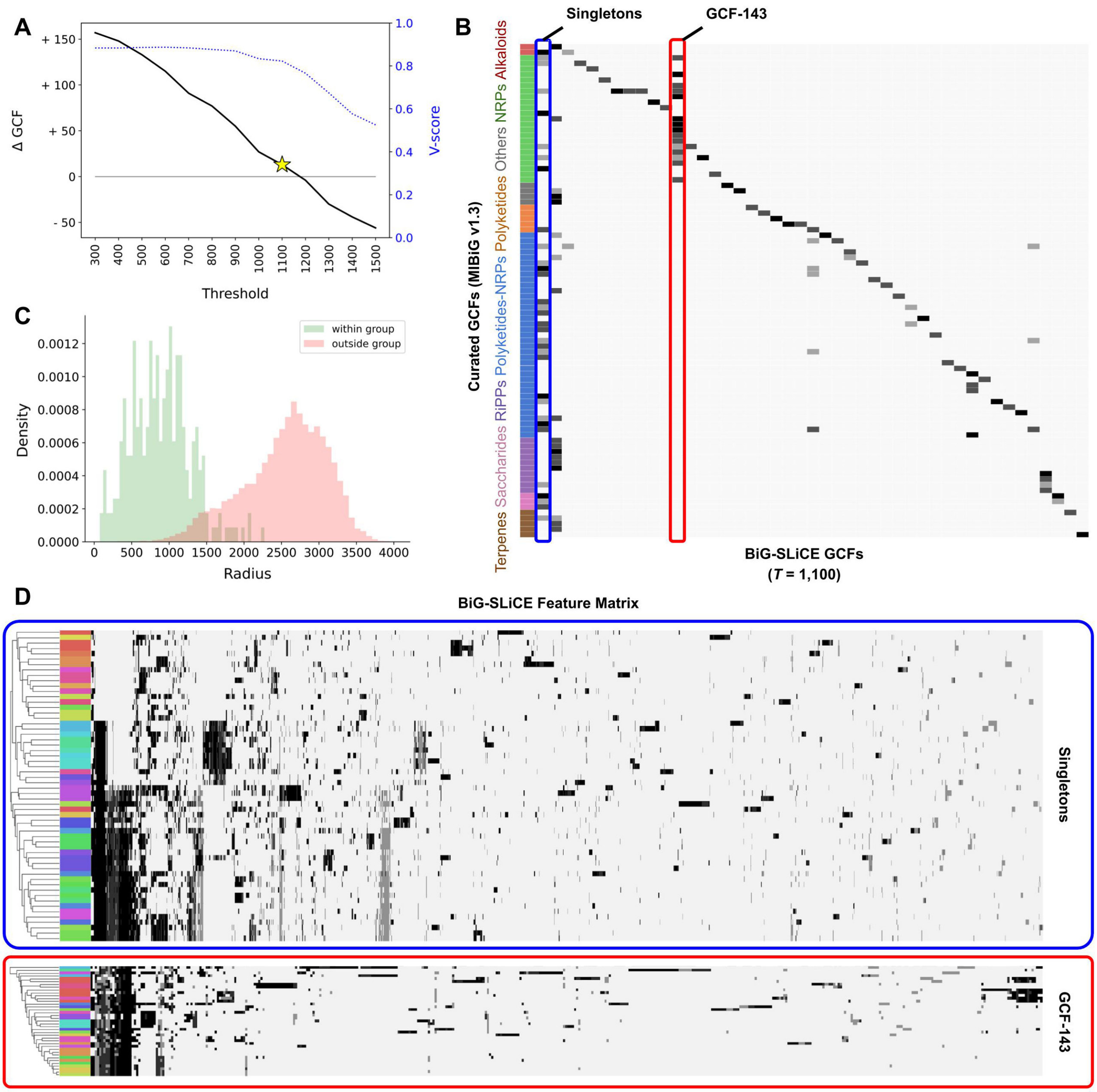
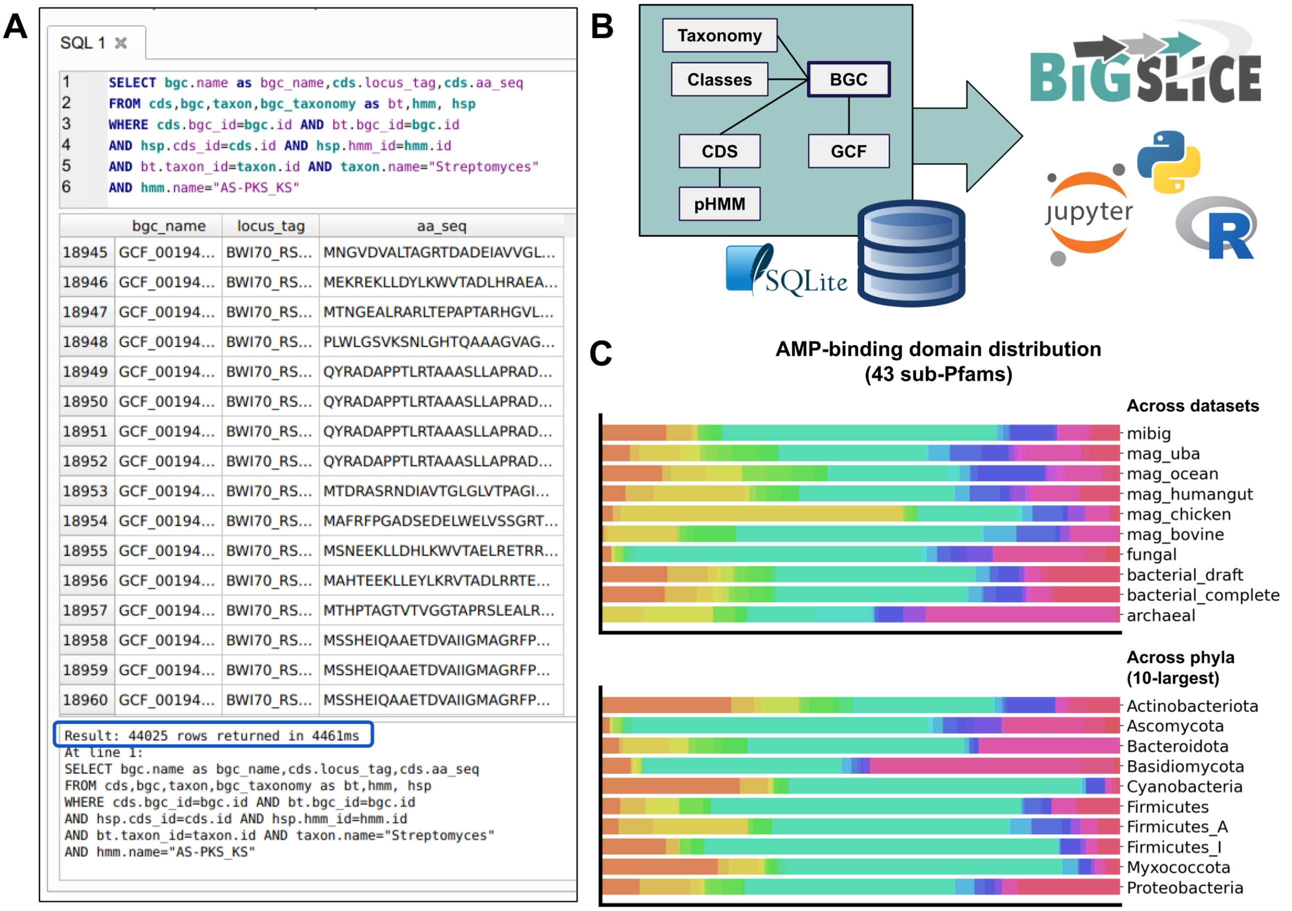


Figure 3: **(A)** BiG-SLiCE analysis results for a range of threshold values, as measured by the difference of GCF counts (ΔGCF) and the level of clustering agreement (V-score of 1.0 for perfect clustering) compared to MIBiG curated groups. A single threshold result with the lowest ΔGCF while maintaining a V-score > 0.8, $T = 1{,}100$, was used 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 heat map of the collapsed singleton group and GCF-143. Colored bars on the left indicate manually curated groups. In both cases, hierarchical clustering analysis (Euclidean-based, average-linkage) shows 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.

toolkit [82], a bioinformatics pipeline that integrates several tools [51, 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 S3).

Large-scale homology analysis of 1.2 Million BGCs

We then performed BiG-SLiCE clustering analyses over the merged datasets using a 36-core, 252 GB 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



```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 4: **(A)** An example SQL query for all protein sequences harboring ≥1 ketosynthase (AS-PKS_KS) domain from streptomycete BGCs. Here, the search performed against the total of ∼29 million coding sequences (CDSs) and >101 million domain hits in the database was completed in <5 seconds, returning 44,025 CDSs 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 ∼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 Dataset [123].

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 to map the fragmented BGCs onto their corresponding GCFs based on the calculated membership values *d* and a user-defined threshold *T*. For this analysis, we arbitrarily categorized GCF-to-BGC relationships into "core" ($d \leq T$), "putative" ($T < d \leq 2T$), 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 S4). The first run ($T = 300$), which carries the full workflow load (from feature extraction to membership assignment), was finished in ∼240 hours (10 days), or >150× faster than the estimated runtime of BiG-SCAPE (Supplementary Fig. S3). 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 (Fig. 5A). Subsequent runs ($T = 600$–$1{,}500$) reused the pre-calculated features, taking only an average of ∼4 hours runtime for each run (Fig. 5B).

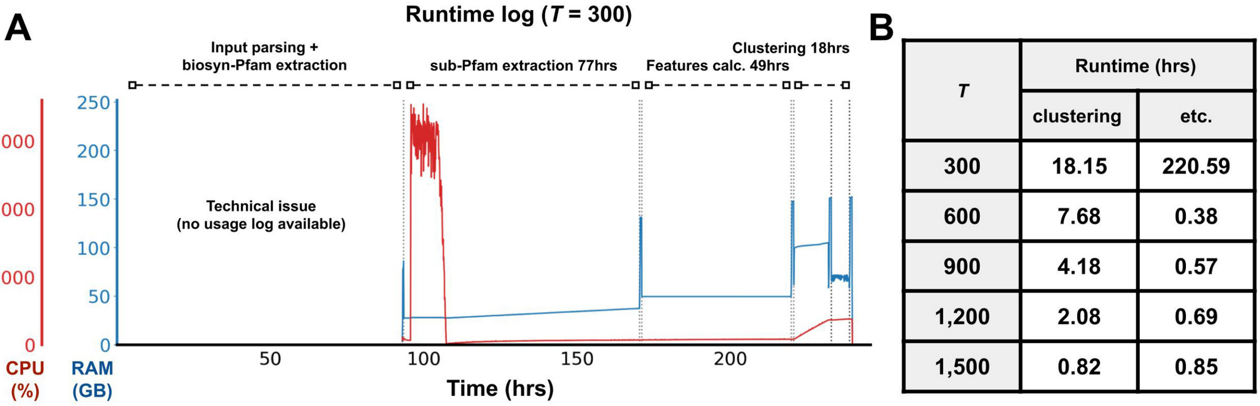
Charting a global map of BGC diversity

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 the Supplementary Dataset for details). Another round of membership assignment was performed to match the full set of 1.2 million 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 Dataset) and plotted via the iTOL web server [88]. By annotating this tree with various types of quantitative information (Supplementary Table S5), the resulting phylogram pictures a generic, "bird's-eye view" on the entire set of 1.2 million BGCs (Fig. 6).

Table 1: Numbers of genomes and BGCs in all datasets included for the large-scale diversity analysis.

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 et al. 2017 (uncultivated bacteria and archaea MAGs)	[62]	7,280 (7,280)	15,829	0 (0)	0	623 (623)	756
Tully et al. 2018 (Tara Ocean MAGs)	[63]	2,283 (2,326)	4,829	0 (0)	0	344 (301)	518
Almeida et al. 2019 (unified human gut MAGs)	[64]	4,616 (4,616)	4,766	0 (0)	0	28 (28)	25
Stewart et al. 2019 (cow's rumen MAGs)	[65]	4,815 (4,815)	8,380	0 (0)	0	126 (126)	589
Glendinning et al. 2020 (chicken's caecum MAGs)	[66]	469 (469)	481	0 (0)	0	0 (0)	0
MIBiG v2.0	[78]	0 (0)	15,94	0 (0)	276	0 (0)	40
Total		200,984 (200,970)	10,964,073	5,939 (5,905)	124,092	2,283 (2,331)	4,506

Numbers inside parentheses indicate the total number of genomes assigned to each kingdom based on the subsequent taxonomy analysis. The "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 S3.



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 5: (A) Runtime breakdown of the full run ($T = 300$) on a 36-core CPU, 262 GB RAM server. Owing 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 1 CPU core, giving room for further improvement, e.g., via SQL parallelization. Spikes in the RAM usage (peak = ~150 GB) came from the periodic "dumping" of the in-memory database (used 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 feature extraction. Here, runtimes are separately shown for both the clustering (GCF model construction + membership assignment) and other steps (input parsing, hmmscanning, and feature extraction).

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 (e.g., there are 21 bins made up of a single BGC [Supplementary Table S5], which can happen when the randomly placed initial centroid hits an outlier/singleton in the dataset). This is analogous to taking a 2D 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 3D structure of the object will always remain constant. While the map shown in Fig. 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–a4) have quite distinct features that discriminate them from the rest of the (mostly bacterial) datasets. Clades a1–a3 contain mostly nonribosomal peptide (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 2 simple NRP-encoding fungal BGCs from the MIBiG dataset, encoding the biosynthesis of the proteasome inhibitor fellutamide B [89] (BGC0001399) and aspergillic

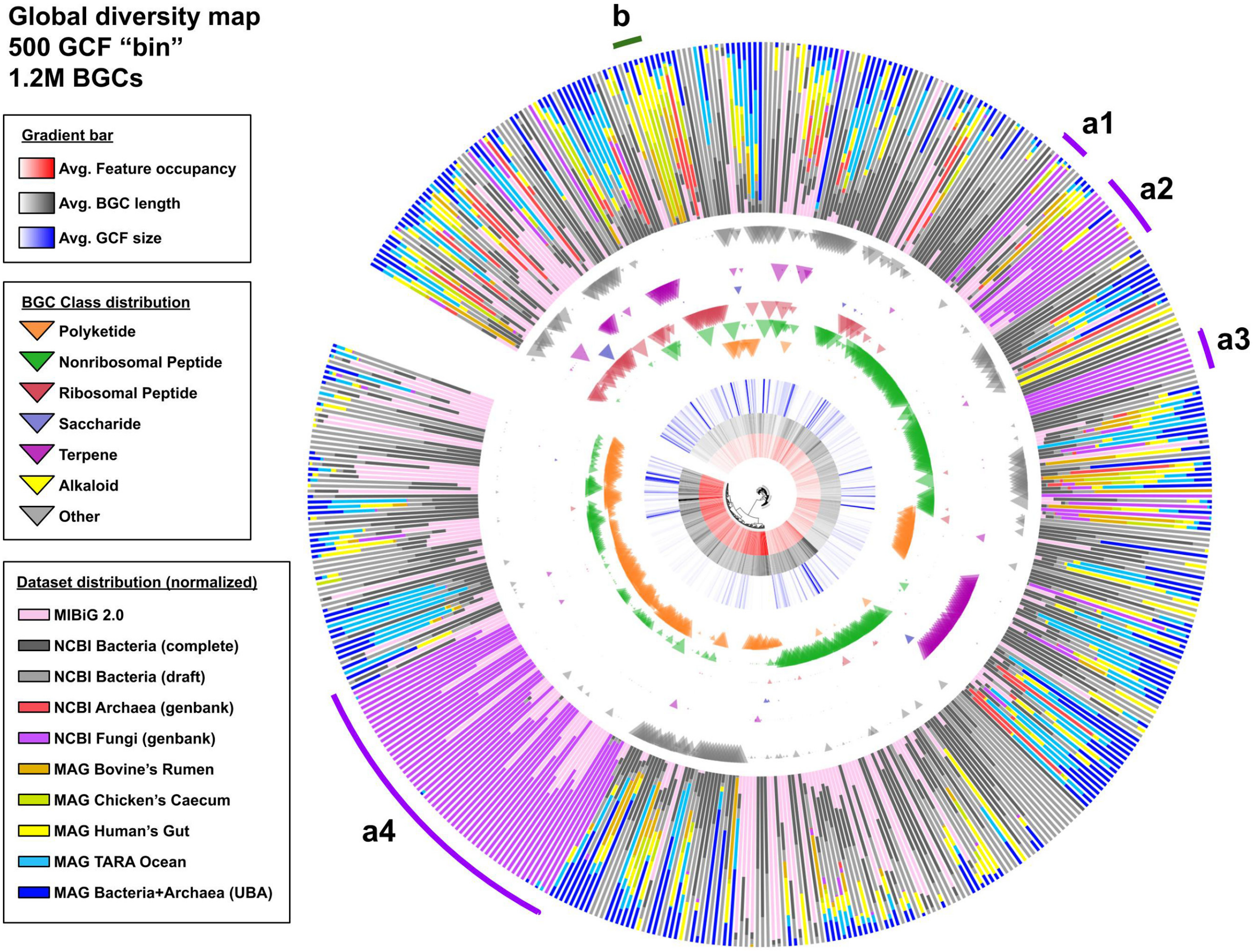


Figure 6: A phylogram created via the hierarchical clustering analysis of 500 GCF bins. The phylogram was rooted on a null (all zeros) dummy feature matrix. For each node, the raw dataset distribution values (Supplementary Table S5) were double-normalized, first against the number of BGCs each dataset has in total, giving the fraction values, then against all fraction values of other datasets in the bin. Furthermore, some notably interesting clades are manually highlighted (a1–a4, b) for follow-up discussion (see main text). UBA: Uncultivated Bacteria and Archaea.

acid [90] (BGC0001516) from *Aspergillus* (and on the bacterial side: 4 MIBiG BGCs including another simple proteasome inhibitor, livipeptin [91, 92], of which the production is encoded by BGC0001168 from *Streptomyces lividans*). Clade a2 contains a major portion (50 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 β-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 >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 1 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 Dataset.

At the opposite side of the phylogram, 42,716 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 AT, KS, 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 polyketides (PKSs) [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-PKSs 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 ribosomally synthesized and post-translationally modified peptide (RiPP) BGCs from the "gut" metagenome datasets (bovine's rumen, chicken's caecum, human gut). Aside from the 2,546 (17.88% of the 3 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 provides an important clue: 99.68% of the BGCs belong to the sactipeptide RiPP subclass as annotated by antiSMASH, and seem 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 is 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. in solventogenic *Clostridia* [102] suggested that these RiPPs might play a role in the quorum-sensing system and in controlling cellular metabolism of such organisms.

Next, by looking at how the pink (innermost) bar is distributed all across the phylogram, we can infer that despite holding ≤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 knowledge on experimentally characterized BGCs that are also present in MIBiG. This means that the 1.2 million BGCs captured from those 209,000 genomes are all evolutionarily related, albeit distantly, to ≥1 MIBiG BGC. To go beyond these canonical pathways, several unsupervised but "lower-confidence" alternative algorithms [40] 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 in 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 apparent that there is a distinct separation between two major groups of GCF bins: a high feature counts group (more intense red bars) consisting of mostly of domain-rich polyketide (and some 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 Fig. S4A). This causes a large dichotomy in GCF sizes (Supplementary Fig. S4B) due to the limitation of the single-threshold clustering method of BIRCH as described previously. 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 Fig. S4C) because 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 Fig. S4D). 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]).

Measuring the "hidden iceberg" of microbial secondary metabolism

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 potential, we took the 29,955 GCFs calculated from $T = 900$, measured the distance of every GCF model against their closest MIBiG BGC features (Supplementary Table S6), and then plotted a histogram from the data (Fig. 7A).

Indeed, it is immediately clear from Fig. 7A that almost all (96.63%) 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 because 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 (Fig. 7B). This also means that many BGCs with very low feature counts 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.

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 be used 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 S7). To obtain an estimate of the α-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 (Fig. 8A). In this scenario, 3 Firmicutes (*Bacillus velezensis*, *Bacillus thuringiensis*, *Streptococcus pneumoniae*) and 5 Proteobacteria (*Escherichia flexneri*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Escherichia coli, Burkholderia ubonensis*) dropped out of the top-30 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]).

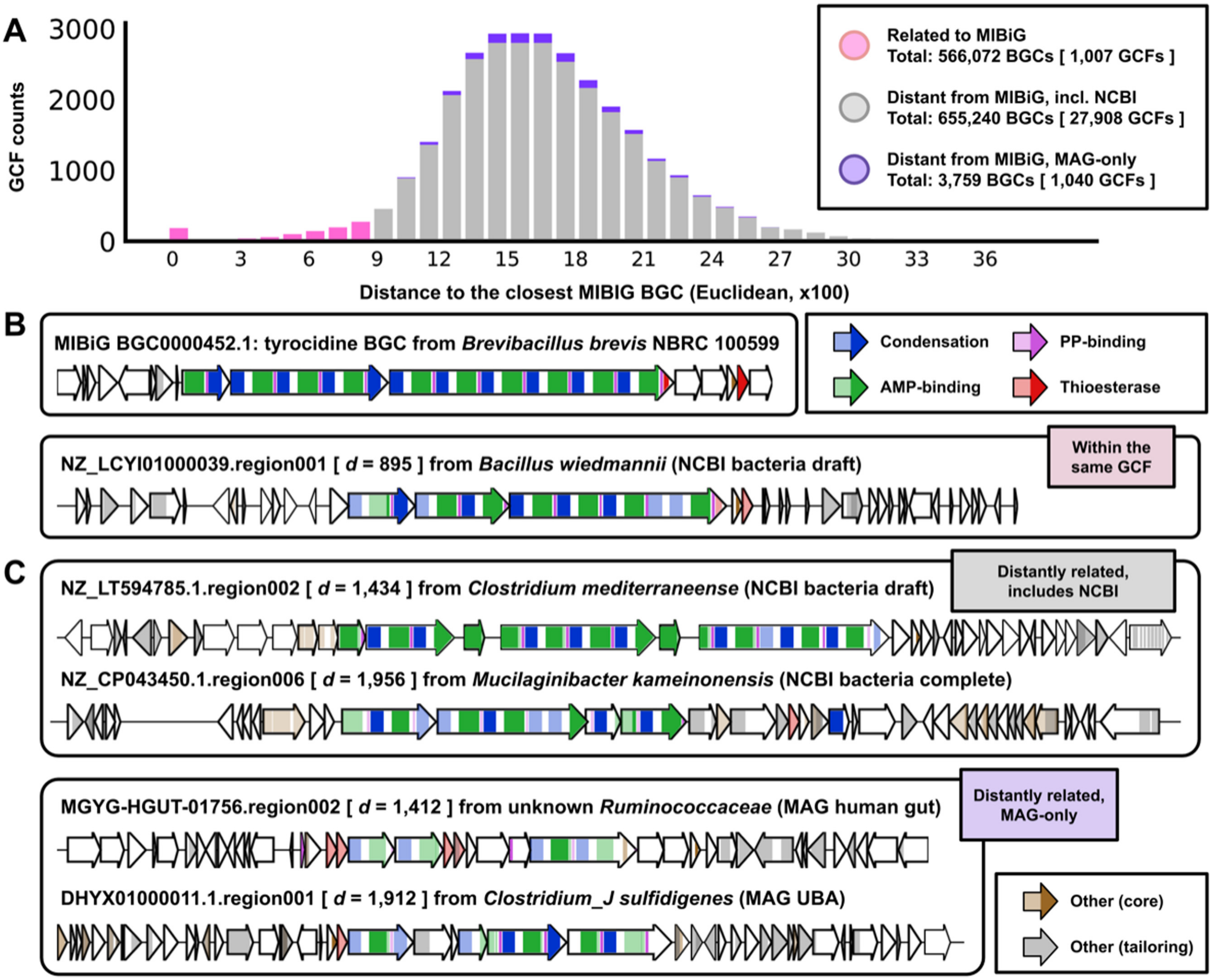


Figure 7: (A) Histogram of Euclidean distances (x-axis) of GCF models to their closest BGC from the MIBiG 2.0 dataset. Here, all GCFs having $d \leq 900$ were denoted as "related to MIBiG" and "distant from MIBiG" if otherwise, particularly highlighting those coming only from the MAG datasets. **(B)** Selected anecdotal example of a MIBiG BGC and 1 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.

Four *Streptomyces* species made it into the selected list of 19 species that consistently ranked top-30 in both runs (Fig. 8B) despite having relatively few genomes (24–78) 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 4 genomes in the dataset (723 species from 486 genera; Supplementary Table S7) 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* (mean 30 GCFs from 2 species), and *Amycolatopsis* (mean 29 GCFs from 3 species).

The rest of the bacterial species (1 actinobacterium, 3 firmicutes, and 3 proteobacteria) that made it into the top-19 are mainly composed of pathogens that have had many of their genomes sequenced (183–4,838 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 a 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 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

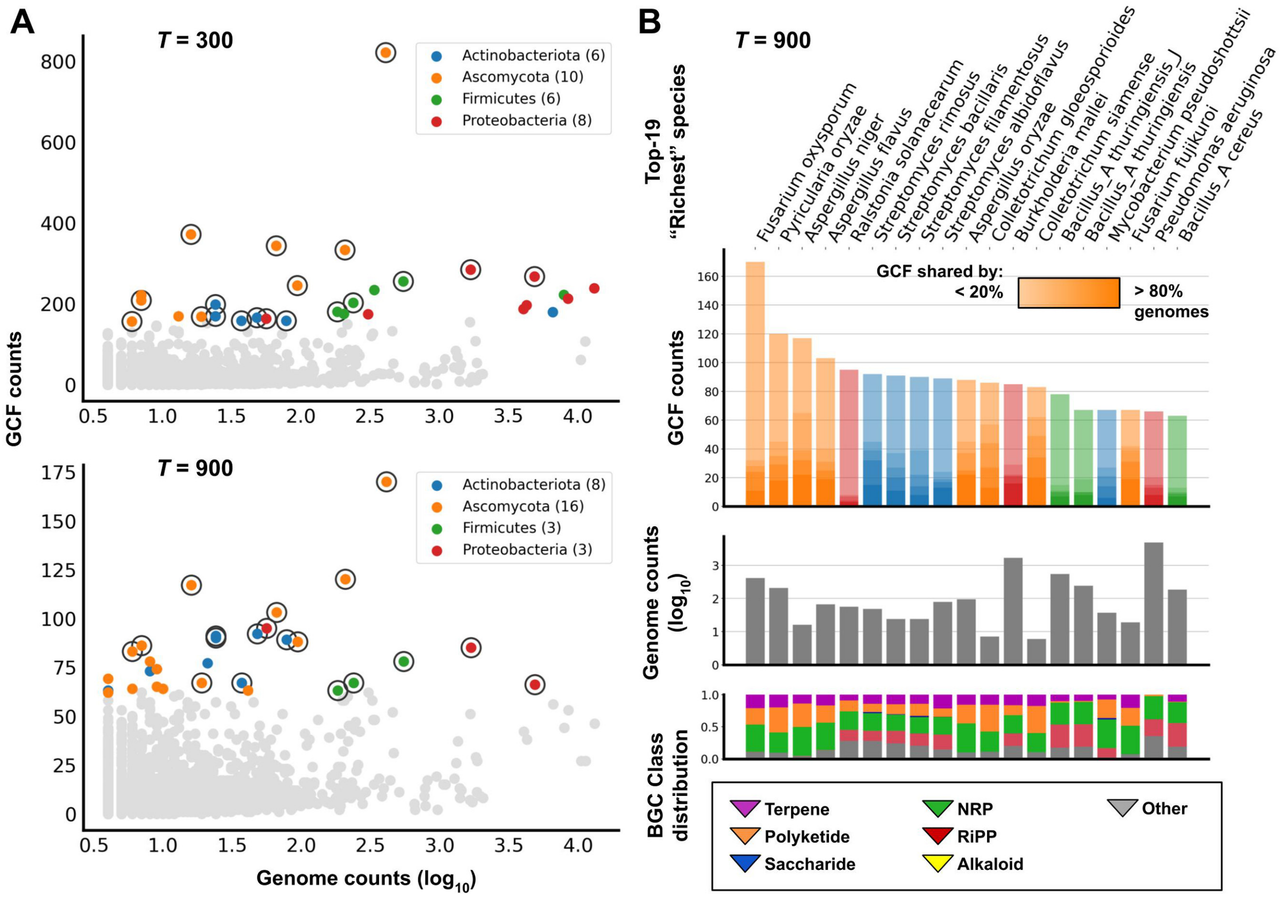


Figure 8: (A) Distribution of GCF counts across species having 4or more genomes in the dataset. Two plots showing results at the most stringent ($T = 300$) and a fairly lenient ($T = 900$) threshold, each highlighting 30 species with the highest GCF counts (colored dots). Nineteen species present in the top-30 of both thresholds are marked with black circles. (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 S7). Additionally, the total distribution of BGC classes per species is also measured (Supplementary Table S8).

observation from this species' data is how little overlap occurred between the BGCs from different strains: 87 of the 95 (91.6%) GCFs are shared only between <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 their 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 have come 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 ∼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 17 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. Because fungi and bacteria seem to frequently compete with each other in the wild [119], it may be logical to expand search efforts for new antibacterial compounds from this nemesis of bacteria, complementary to bacterial genome mining.

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 on 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 and DeepBGC.

Furthermore, the sub-Pfam approach that we introduced here could have potential uses beyond GCF construction. By using it in place of the more generic Pfam models, it would be possible to 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, one can also use sub-Pfams 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 of as a global target-directed genome mining approach [60, 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 that 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 resources 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.

Availability of Supporting Source Code and Requirements

Project name: BiG-SLiCE
Project home page: https://github.com/medema-group/bigslice
RRID:SCR_019130
BiotoolsID: big_slice
Operating system(s): Linux/UNIX-based OS, output web app can be viewed on any modern Internet browser
Programming language: Python
Other requirements: Python 3.6 or higher
License: GNU Affero General Public License v3.0

Data Availability

Input BGCs, analysis results, and Python scripts used to generate all figures and tables in this study, and all supplementary texts and figures are 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].

Additional Files

Supplementary Table S1. List of biosynthetic-Pfam pHMMs used by BiG-SLiCE.
Supplementary Table S2. List of "core" biosynthetic-Pfam and the respective sub-Pfam pHMM models.
Supplementary Table S3. List of genomes per dataset along with the total count of BGCs predicted by antiSMASH and their assigned taxonomy.
Supplementary Table S4. Summary of 5 different run parameters on the full dataset of 1.2M BGCs.
Supplementary Table S5. Calculated statistics of the 266 GCFs that were used to annotate the global phylogram map of biosynthetic diversity.
Supplementary Table S6. BGC counts per dataset of 29,955 GCFs from the $T = 900$ run and the calculated distance to the closest matching MIBiG BGC.
Supplementary Table S7. Unique GCF counts of species having ≥4 strain genomes in the full dataset.
Supplementary Table S8. BGC class absence/presence distribution of species in the full dataset. Hybrid BGCs had each of their classes counted separately, meaning the sum of the numbers will not be equal to the total number of BGCs per species.
Supplementary Figure S1. Confusion heatmap of BiG-SCAPE result compared to the curated set of MIBiG BGCs. The result was generated using BiG-SCAPE version 1.0.1, using a cutoff threshold of 0.75 and hybrid mode turned off, as specified in the original paper. A "vertical band" is highlighted in blue, comprising BGCs unintentionally assigned as singletons due to the strictness of the cutoff parameter being used.
Supplementary Figure S2. An Entity-Relationship Diagram (ERD) of the SQLite3 database used in BiG-SLiCE v1.0.0 (this study). The ERD was generated using SchemaSpy version 6.1.0 (http://schemaspy.org/).
Supplementary Figure S3. Runtime comparison between BiG-SCAPE and BiG-SLiCE. Runs were performed on a 36-cores CPU using subsets of randomly sampled BGCs from the dataset (a single subset will be used for both compared runs and will also be included for subsequent runs with larger subsets). Using data points from the sampled runs, a curve was fitted to estimate the runtime of an input size of 1,225,071 BGCs for BiG-SCAPE, while the real runtime taken from the full run log of T = 300 is used for BiG-SLiCE.
Supplementary Figure S4. A. Distribution of features count (calculated by the total feature values divided by 255) across different BGC classes. Here, the distribution of BGCs having less than 50 features is highlighted, showing that some BGC classes tend to have much fewer features than others. B. Distribution of GCF sizes from the T = 900, showing some GCFs having a significantly high number of BGCs, mainly due to the effect of low features count of the BGCs. C. Examples of BGCs having high copy numbers of the same domain, and D. BGCs relying on (or having) only a single biosynthetic domain as detected by BiG-SLiCE, thus resulting in a highly similar features matrix, leading them being grouped together into a single GCF.

Supplementary Text S1. NCBI query scripts (to be used in https://www.ncbi.nlm.nih.gov/assembly/advanced/) used to download all isolate genomes for this study.

Abbreviations

AMP: adenosine monophosphate; API: application programming interface; AT: acyltransferase; BGC: biosynthetic gene cluster; BiG-SLiCE: Biosynthetic Genes Super-Linear Clustering Engine; bp: base pairs; CPU: central processing unit; GCF: gene cluster family; GTDB: Genome Taxonomy Database; JI: Jaccard Index; KS: ketosynthase; MAG: metagenome-assembled genome; NCBI: National Center for Biotechnology Information; NRP: non-ribosomal peptide; NRPS: non-ribosomal peptide synthase; pHMM: protein hidden Markov model; PKS: polyketide synthase; RAM: random access memory; RiPP: ribosomally translated post-translationally modified peptide.

Competing Interests

M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. The authors declare that they have no other competing interests.

Funding

The work of S.A.K. was supported by the Graduate School for Experimental Plant Sciences (EPS), The Netherlands. J.J.J.v.d.H. and M.H.M. acknowledge funding by the Netherlands eScience Center (NLeSC) Accelerating Scientific Discoveries Grant (ASDI.2017.030).

Authors' Contributions

S.A.K. and M.H.M. conceived the study. S.A.K. designed and wrote the BiG-SLiCE software. S.A.K. collected and processed all input data. S.A.K. performed all analyses with help and input from all other authors. J.J.J.v.d.H. and M.H.M. provided input on the biochemical perspective of the study. D.D.R., J.J.J.v.d.H., and M.H.M. provided input on the computational parts of the clustering algorithm. S.A.K. wrote the initial draft of the manuscript. All authors contributed to writing and editing the final version of the manuscript.

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.

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**:3063–82.
4. Mukherjee AK, Das K. Microbial surfactants and their potential applications: an overview. Adv Exp Med Biol 2010;**672**:54–64.
5. No Time to Wait: Securing the future from drug-resistant infections. World Health Organization, 2019.Accessed Feb 18, 2020; Available from: 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**:5970–5.
7. Larsen BB, Miller EC, Rhodes MK, et al. Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life. Q Rev Biol 2017;**92**:229–65.
8. Li S, Hu X, Li L, et al. 1-hydroxy-7-oxolavanducyanin and $\Delta^{7'',8''}$-6″-hydroxynaphthomevalin from *Streptomyces* sp. CPCC 203577. J Antibiot (Tokyo) 2020;**73**:324–8.
9. Nguyen HT, Pokhrel AR, Nguyen CT, et al. *Streptomyces* sp. VN1, a producer of diverse metabolites including non-natural furan-type anticancer compound. Sci Rep 2020;**10**:1756.
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, et al. Two new glutarimide antibiotics from *Streptomyces* sp. HS-NF-780. J Antibiot (Tokyo) 2019;**72**:241–5.
12. Han Y, Wang Y, Yang Y, et al. 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, et al. Streptovertimycins A–H, new fasamycin-type antibiotics produced by a soil-derived *Streptomyces morookaense* strain. J Antibiot (Tokyo) 2020;**73**:283–9.
14. Eckburg PB, Gill SR, Costello EK, et al. The Integrative Human Microbiome Project. Nature 2019;**569**:641–8.
15. Mendes R, Kruijt M, de Bruijn I, et al. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 2011;**332**:1097–100.
16. Amos GCA, Awakawa T, Tuttle RN, 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**:E11121–30.
17. Du C, van Wezel GP. Mining for microbial gems: integrating proteomics in the postgenomic natural product discovery pipeline. Proteomics 2018;**18**:1700332.
18. Rochfort S. Metabolomics reviewed: a new “omics” platform technology for systems biology and implications for natural products research. J Nat Prod 2005;**68**:1813–20.
19. Blin K, Shaw S, Steinke K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 2019;**47**:W81–7.
20. Christopher T, Walsh MAF. Natural Products Version 2.0: Connecting genes to molecules. J Am Chem Soc 2010;**132**:2469.
21. Fondi M, Emiliani G, Fani R. Origin and evolution of operons and metabolic pathways. Res Microbiol 2009;**160**:502–12.
22. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 2020;**16**:60–8.
23. Doroghazi JR, Albright JC, Goering AW, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol 2014;**10**:963–8.
24. Cimermancic P, Medema MH, Claesen J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 2014;**158**:412–21.
25. Goering AW, McClure RA, Doroghazi JR, et al. Metabologenomics: correlation of microbial gene clusters with metabo-

lites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. ACS Cent Sci 2016;**2**:99–108.
26. Moghaddam JA, Crüsemann M, Alanjary M, et al. Analysis of the genome and metabolome of marine myxobacteria reveals high potential for biosynthesis of novel specialized metabolites. Sci Rep 2018;**8**:16600.
27. Duncan KR, Crüsemann M, Lechner A, 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**:460–71.
28. Nielsen JC, Grijseels S, Prigent S, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. Nat Microbiol 2017;**2**, 17044.
29. McClure RA, Goering AW, Ju K-S, et al. Elucidating the rimosamide-detoxin natural product families and their biosynthesis using metabolite/gene cluster correlations. ACS Chem Biol 2016;**11**:3452–60.
30. Parkinson EI, Tryon JH, Goering AW, et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. ACS Chem Biol 2018;**13**:1029–37.
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**:e00387–19.
32. Olm MR, Bhattacharya N, Crits-Christoph A, et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria. Sci Adv 2019;**5**:eaax5727.
33. Carrión VJ, Perez-Jaramillo J, Cordovez V, et al. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 2019;**366**:606–12.
34. The long view on sequencing. Nat Biotechnol 2018;**36**:287.
35. Blin K, Pascal Andreu V, de los Santos ELC, et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Res 2019;**47**:D625–30.
36. Palaniappan K, Chen I-MA, Chu K, et al. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res 2020;**48**:D422–30.
37. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;**44**:D733–45.
38. Skinnider MA, Merwin NJ, Johnston CW, et al. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res 2017;**45**:W49–54.
39. Sélem-Mojica N, Aguilar C, Gutiérrez-García K, et al. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. Microb Genom 2019;**5**:e000260.
40. Hannigan GD, Prihoda D, Palicka A, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res 2019;**47**:e110.
41. Papageorgiou L, Eleni P, Raftopoulou S, et al. Genomic big data hitting the storage bottleneck. EMBnet J 2018;**24**:e910.
42. SQLite Home Page. https://www.sqlite.org/index.html. Accessed 27 January 2020.
43. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. Nucleic Acids Res 2019;**47**:D427–32.
44. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;**315**:972–6.
45. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett 2010;**31**:651–66.
46. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. SIGMOD Rec 1996;**25**:103–14.
47. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv 2013:1301.3781v3.
48. Viehweger A, Krautwurst S, Parks DH, et al. An encoding of genome content for machine learning. BioRxiv 2019, doi:10.1101/524280.
49. Buchan DWA, Jones DT. Learning a functional grammar of protein domains using natural language word embedding techniques. Proteins 2020;**88**:616–24.
50. Caselles-Dupré H, Lesaint F, Royo-Letelier J. Word2vec applied to recommendation: hyperparameters matter. In: Proceedings of the 12th ACM Conference on Recommender Systems. New York, NY, USA: ACM; 2018:352–6.
51. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011;**7**:e1002195.
52. Alborzi SZ, Devignes M-D, Ritchie DW. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics 2017;**18**:107.
53. Katz L. Manipulation of modular polyketide synthases. Chem Rev 1997;**97**:2557–76.
54. Chen C, Natale DA, Finn RD, et al. Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation. PLoS One 2011;**6**:e18910.
55. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;**12**:2825–30.
56. Lorbeer B, Kosareva A, Deva B, et al. Variations on the clustering algorithm BIRCH. Big Data Res 2018;**11**:44–53.
57. Rosenberg A, Hirschberg J. V-Measure: a conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007:410–20.
58. Flask. Pallets. https://palletsprojects.com/p/flask/. Accessed 27 January 2020.
59. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, et al. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. Genome Biol Evol 2016;**8**:1906–16.
60. Mungan MD, Alanjary M, Blin K, et al. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Res 2020;**48**:W546–52.
61. Krause J, Handayani I, Blin K, et al. Disclosing the potential of the SARP-type regulator PapR2 for the activation of antibiotic gene clusters in streptomycetes. Front Microbiol 2020;**11**:225.
62. Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;**2**:1533–42.
63. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 2018;**5**:170203.
64. Almeida A, Nayfach S, Boland M, 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.

65. Stewart RD, Auffret MD, Warr A, et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol 2019;**37**:953–61.
66. Glendinning L, Stewart RD, Pallen MJ, et al. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol 2020;**21**(1):34.
67. Hervé V, Liu P, Dietrich C, 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, 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, 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, doi:10.1101/2020.05.19.105130.
70. Youngblut ND, de la Cuesta-Zuluaga J, Reischer GH, et al. Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. mSystems 2020;**5**:e01045–20.
71. Pamela Engelberts J, Robbins SJ, de Goeij JM, et al. Characterization of a sponge microbiome using an integrative genome-centric approach. ISME J 2020;**14**(5):1100–10.
72. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 2020;**38**:701–7.
73. Liang R, Lau MCY, Saitta ET, et al. Genome-centric resolution of novel microbial lineages in an excavated *Centrosaurus* dinosaur fossil bone from the Late Cretaceous of North America. Environ Microbiome 2020;**15**:4724.
74. Eze MO, Lütgert SA, Neubauer H, et al. Metagenome assembly and metagenome-assembled genome sequences from a historical oil field located in Wietze, Germany. Microbiol Resour Announc 2020;**9**:e00333–20.
75. Newberry E, Bhandari R, Kemble J, et al. Genome-resolved metagenomics to study co-occurrence patterns and intraspecific heterogeneity among plant pathogen metapopulations. Environ Microbiol 2020;**22**:2693–708.
76. Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 2019;**176**:649–62.
77. Nayfach S, Shi ZJ, Seshadri R, et al. New insights from uncultivated genomes of the global human gut microbiome. Nature 2019;**568**:505–10.
78. Kautsar SA, Blin K, Shaw S, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 2020;**48**:D454–8.
79. Martínez-Romero E, Rodríguez-Medina N, Beltrán-Rojel M, et al. Genome misclassification of *Klebsiella variicola* and *Klebsiella quasipneumoniae* isolated from plants, animals and humans. Salud Publica Mex 2017;**60**:56–62.
80. Ciufo S, Kannan S, Sharma 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, et al. Phylogenomics reveals clear cases of misclassification and genus-wide phylogenetic markers for *Acinetobacter*. Genome Biol Evol 2019;**11**:2531–41.
82. Chaumeil P-A, Mussig AJ, Hugenholtz P, et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 2019;**36**:1925–7.
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, et al. 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, et al. 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, 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**:W256–9.
89. Yeh H-H, Ahuja M, Chiang Y-M, 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**:2275–84.
90. Lebar MD, Cary JW, Majumdar R, 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, inventors; Instituto Politecnico Nacional Centro de Investigacion Y Estudio, assignee. Genetic system for producing a proteases inhibitor of a small peptide aldehyde type. US Patent 10,414,796 (2019).
92. Cruz-Morales P, Vijgenboom E, Iruegas-Bocardo F, 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**:1165–75.
93. Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC Evol Biol 2010;**10**:26.
94. Simpson TJ, Cox RJ. Polyketide biosynthesis: Fungi. In: Begley TP . Wiley Encyclopedia of Chemical Biology. Hoboken, NJ, USA: Wiley; 2008:380.
95. Chen H, Du L. Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. Appl Microbiol Biotechnol 2016;**100**:541–57.
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, 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 , eds. Polyketides. 2007:154–66. ACS Publications, ACS Symposium Series; Vol. 955.
98. Liu W, Christenson SD, Standage S, et al. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science 2002;**297**:1170–3.
99. Li X, Lei X, Zhang C, 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**:2745–55.
101. Hudson GA, Burkhart BJ, DiCaprio AJ, 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**: 8228–38.
102. Chen Y, Yang Y, Ji X, et al. The SCIFF-derived ranthipeptides participate in quorum sensing in solventogenic clostridia. Biotechnol J 2020;**15**:2000136.
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**:4169–82.
104. Tietz JI, Schwalen CJ, Patel PS, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol 2017;**13**:470–8.
105. Walker MC, Eslami SM, Hetrick KJ, et al. 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, et al. RRE-Finder: a genome-mining tool for class-independent RiPP discovery. mSystems 2020;**5**:e00267–20.
107. Baltz RH. Gifted microbes for genome mining and natural product discovery. J Ind Microbiol Biotechnol 2017;**44**:573–88.
108. Pye CR, Bertin MJ, Lokey RS, et al. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A 2017;**114**:5601–6.
109. Park CJ, Smith JT, Andam CP. Horizontal gene transfer and genome evolution in the phylum Actinobacteria. In: Villa TG, Viñas M , eds. Horizontal Gene Transfer. Cham: Springer; 2019:155–74.
110. McDonald BR, Currie CR. Lateral gene transfer dynamics in the ancient bacterial genus *Streptomyces*. MBio 2017;**8**:e00644–17.
111. Tidjani A-R, Lorenzi J-N, Toussaint M, et al. Massive gene flux drives genome diversity between sympatric *Streptomyces* conspecifics. MBio 2019;**10**:e01533–19.
112. Procópio RE L, Silva IR, Martins MK, et al. Antibiotics produced by *Streptomyces*. Braz J Infect Dis 2012;**16**:466–71.
113. Rhodes MW, Kator H, McNabb A, 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**:1139–47.
114. Kreutzer MF, Kage H, Gebhardt P, et al. Biosynthesis of a complex yersiniabactin-like natural product via the *mic* locus in phytopathogen *Ralstonia solanacearum*. Appl Environ Microbiol 2011;**77**:6117–24.
115. Spraker JE, Sanchez LM, Lowe TM, et al. *Ralstonia solanacearum* lipopeptide induces chlamydospore development in fungi and facilitates bacterial entry into fungal tissues. ISME J 2016;**10**:2317–30.
116. Prior P, Ailloud F, Dalsing BL, et al. Genomic and proteomic evidence supporting the division of the plant pathogen *Ralstonia solanacearum* into three species. BMC Genomics 2016;**17**:90.
117. Remenant B, Coupat-Goutaland B, Guidot A, et al. Genomes of three tomato pathogens within the *Ralstonia solanacearum* species complex reveal significant evolutionary divergence. BMC Genomics 2010;**11**:379.
118. Hayward AC. Characteristics of *Pseudomonas solanacearum*. J Appl Bacteriol 1964;**27**:265–77.
119. Bahram M, Hildebrand F, Forslund SK, et al. Structure and function of the global topsoil microbiome. Nature 2018;**560**:233–7.
120. Del Carratore F, Zych K, Cummings M, et al. 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**:1426–37.
122. Kautsar SA, Blin K, Shaw S, et al. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res 2020;**D1**:D490–D497
123. Kautsar SA, van der Hooft JJJ, Ridder D, et al. Supporting data for "BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters." GigaScience Database 2020. http://dx.doi.org/10.5524/100826.
124. Kautsar SA. medema-group/bigslice: Version 1.0.0. Zenodo. 2020. http://dx.doi.org/10.5281/zenodo.3975432.