BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters

Satria A. Kautsar[1], Justin J. J. van der Hooft[1], Dick de Ridder[1], Marnix H. Medema[1]✉

[1]Bioinformatics Group, Wageningen University, the Netherlands

✉ correspondence email: marnix.medema@wur.nl

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 suffer from 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 (MAGs) within ten days on a typical 36-cores 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, searchable interconnected network of BGCs. As more genomes get sequenced from understudied taxa, more information can be mined to highlight their potentially novel chemistry. BiG-SLiCE is available via https://github.com/medema-lab/bigslice.

Keywords: *Biosynthetic Gene Cluster, Gene Cluster Family, Biosynthetic Diversity, Natural Products 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 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 (WHO) 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]. Due to the sheer size of microbial and enzymological biodiversity, there exists a vast repertoire of potentially useful compounds remains to be unearthed. 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 [23,25]. 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 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 [23,24] 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 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.

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 [37]. 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 ($\sim O(n)$) 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.

Methods and Implementation

The BiG-SLiCE workflow starts at the vectorization (feature extraction) step [Figure 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 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 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 1D].

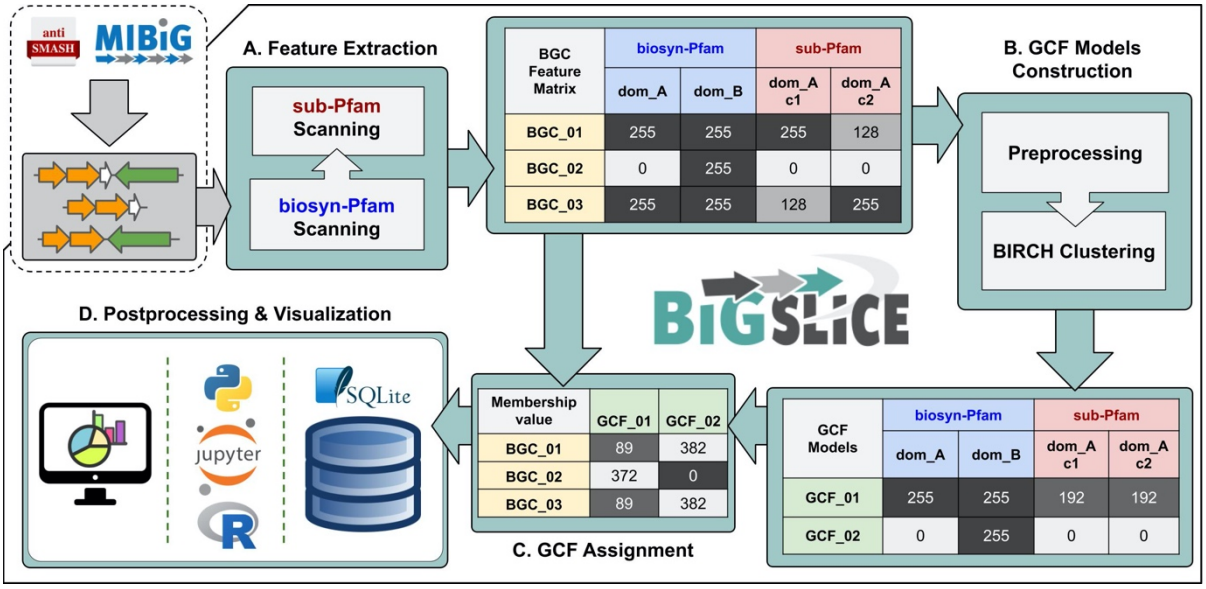


Figure 1. An overview of BiG-SLiCE's GCF analysis workflow. Taking an input of region/cluster GenBank files from antiSMASH and MIBiG, **A.** BiG-SLiCE converts BGCs into numerical feature vectors, which are used to **B.** construct the GCF models (cluster centroids) and **C.** calculate BGC-to-GCF membership values. Processed data and results are all stored in a file-based SQL database (using SQLite3 [38]), which can then be used **D.** to perform further analysis (via external scripts) or to visualize the result in a user-interactive application.

BGC feature extraction

In BiG-SCAPE, the (shared) occurrence and synteny (order) of Pfam [39] 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 [40]. 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 [41] or BIRCH [42], we need to transform BGCs into numerical feature vectors (commonly known as quantization). We do this using a combination of two approaches: 1) biosynthetic domain (biosynthetic-pfam) absence/presence matrix construction and 2) signature domain (sub-Pfam) fingerprinting.

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

Domain hits (retrieved using hmmscan [43] with the gathering threshold) obtained for a reduced list of Pfam version 32 [39] pHMM models [Figure 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 [44] (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 2C].

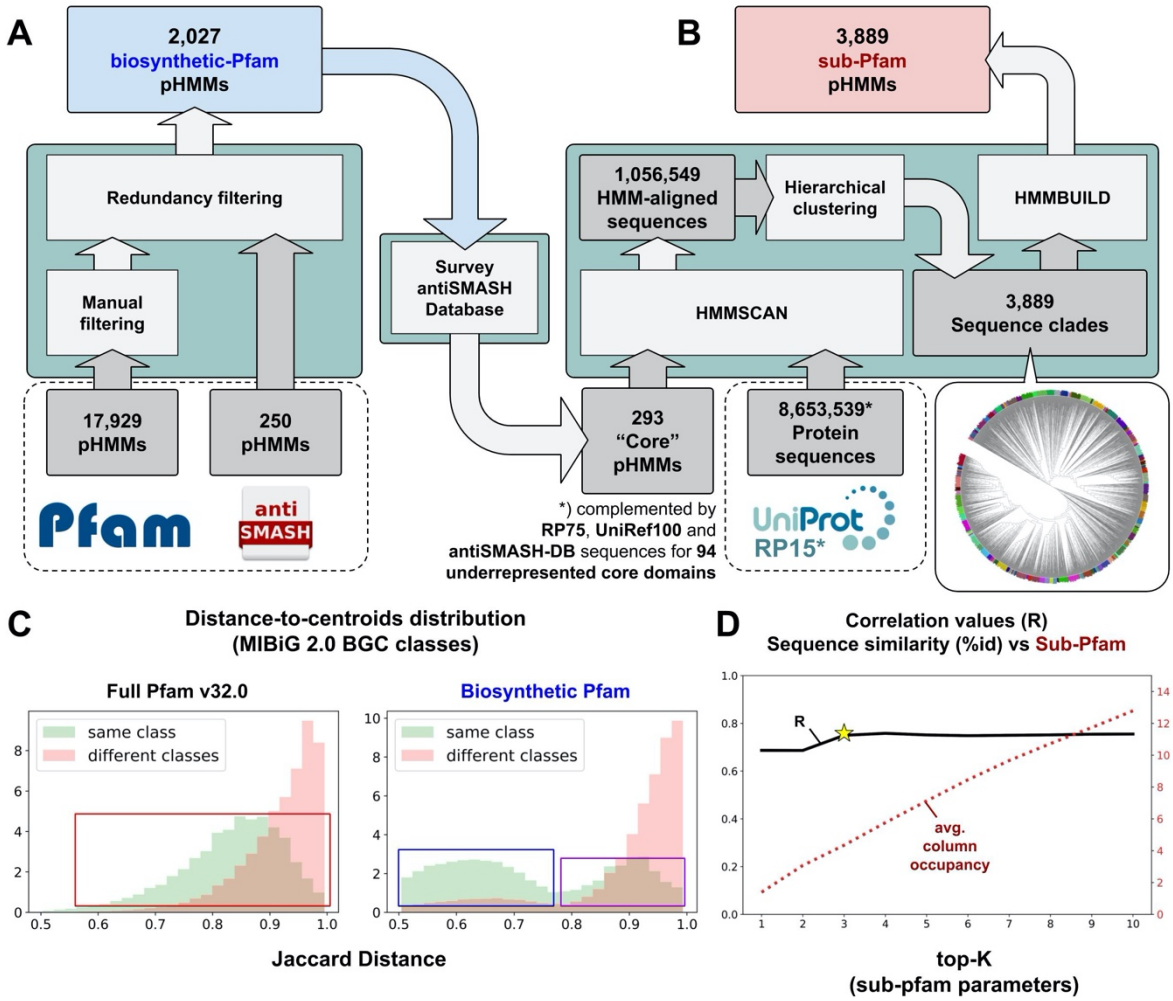


Figure 2. A. Construction of biosynthetic-Pfam features and **B.** Sub-level Pfam (sub-Pfam) features. **C.** Effect of Pfam model filtering on the discriminatory power of domain-presence Jaccard distance (JI index in BiG-SCAPE) measurements to separate MIBiG v2.0 generic classes (Polyketide, NRP, RiPP, 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 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 [45]. To cover this sequence-level protein diversity, we constructed alignments of 9,451,490 representative protein sequences in the RP15 database (Release 2020_01) [46] 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 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 2D].

GCF models construction

To efficiently group BGC features into GCFs, BiG-SLiCE uses a clustering method based on the python scikit-learn [47] implementation of the BIRCH [42] 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 *Xth*-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*) [48] 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.

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 - 1,500) and corresponding results were compared to the reference groups. We calculated the V-score [49] 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 (*Δ*GCF) between the two. We found that BiG-SLiCE produces a generally agreeable result at the selected "optimal" threshold (V-score = 0.81 on *T* = 1,100), but is not able to capture the "perfect" clustering denoted by the reference groups [Figure 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 Figure 1]) due to its Affinity Propagation [40] 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 [Figure 3C] and the visualized feature heatmap of the most challenging groups [Figure 3D].

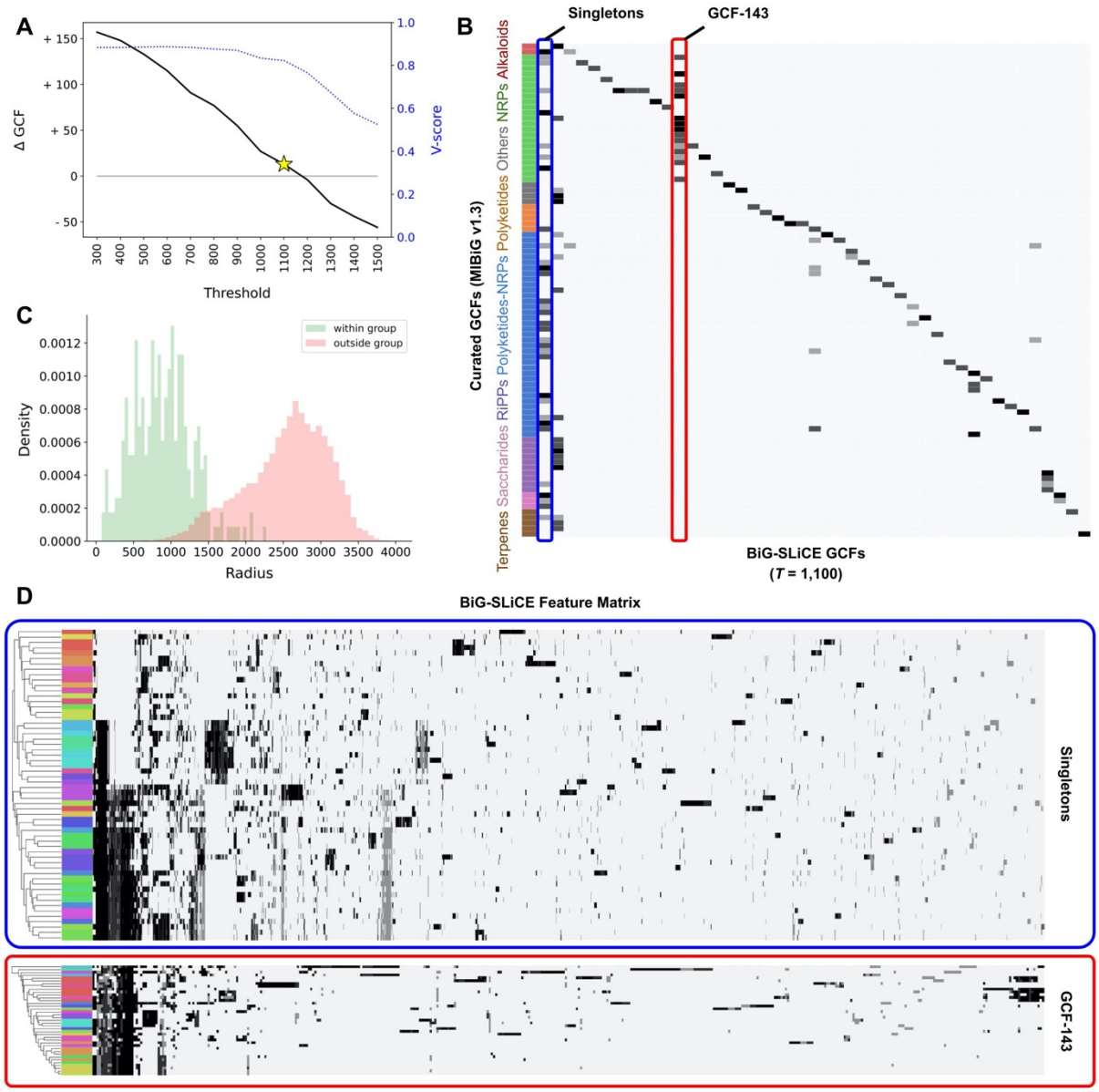


Figure 3. A. BiG-SLiCE analysis results for a range of threshold values, as measured by the difference of GCF counts (ΔGCF) and the level of clustering agreement (V-score of 1.0 for perfect clustering) compared to MIBiG curated groups. A single threshold result with the lowest ΔGCF while maintaining a V-score > 0.8, T = 1,100, was taken 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. 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.

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 [38]. Processed input (including all metadata), supporting data and clustering results are systematically stored in the database tables. 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 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 4B] to perform even more complex analyses, for example to study the diversity of biosynthetic domains across samples and across taxonomy [Figure 4C]. As a matter of fact, all analyses performed in this study (see Results & Discussion) heavily benefitted from (and relied on) the data-wrangling convenience provided by BiG-SLiCE's SQLite database.

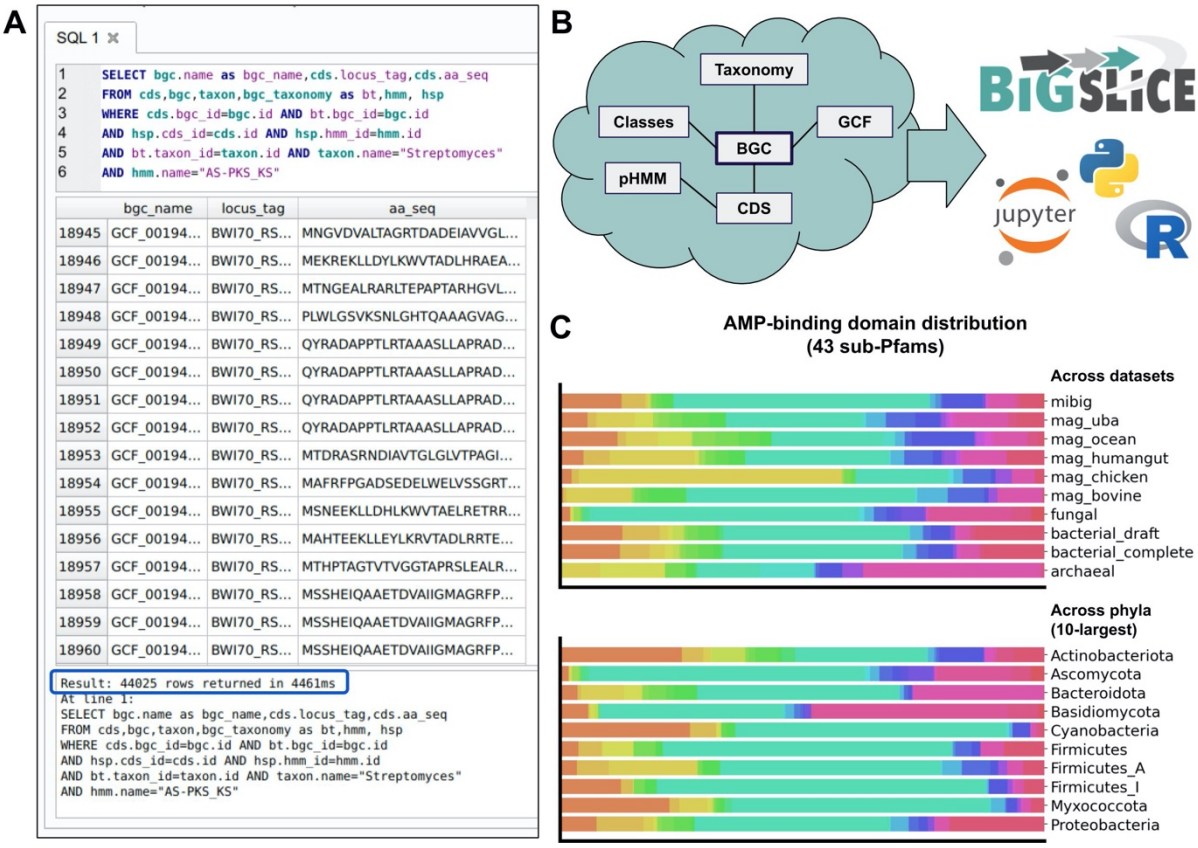


Figure 4. **A.** An example SQL query for all protein sequences harboring at least one Ketosynthase (AS-PKS_KS) domain from streptomycete BGCs. Here, the search performed against the total of ~29 million CDSes and >101 million domain hits in the database was completed in under five seconds, returning 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 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.

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 [50]. 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 (https://github.com/medema-group/bigslice#user-interactive-output). 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 [22,51], or even incorporating additional useful information such as the presence/absence of antibiotic-resistant genes [52] and regulatory domains [53] 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).

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]). Furthermore, we collected and processed 20,584 MAGs from previously published studies [54–58], 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 [59–70] 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 [71] 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 1.

Table 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. 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 3].

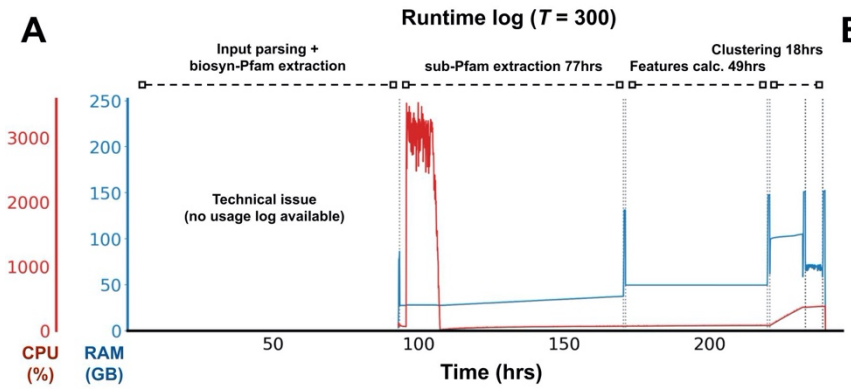
Dataset Name	Study	Counts (Genomes, BGCs)					
		Bacterial		Fungal		Others	
RefSeq complete bacteria	-	19,169 (19,166)	101,531	0 (0)	0	0 (3)	0
RefSeq draft bacteria	-	162,352 (162,297)	959,061	0 (0)	0	0 (55)	346
GenBank fungi	-	0 (0)	0	5,939 (5,905)	123,816	0 (34)	123
GenBank archaea	-	0 (1)	2	0 (0)	0	1,162 (1,161)	2,109
Parks 2017 (Uncultivated Bacteria and Archaea MAGs)	[54]	7,280 (7,280)	15,829	0 (0)	0	623 (623)	756
Tully 2018 (TARA ocean MAGs)	[55]	2,283 (2,326)	4,829	0 (0)	0	344 (301)	518
Almeida 2019 (Unified Human Gut MAGs)	[56]	4,616 (4,616)	4,766	0 (0)	0	28 (28)	25
Stewart 2019 (Cow's rumen MAGs)	[57]	4,815 (4,815)	8,380	0 (0)	0	126 (126)	589
Glendinning 2020 (Chicken's caecum MAGs)	[58]	469 (469)	481	0 (0)	0	0 (0)	0
MIBiG v2.0	[71]	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

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 [72–74]. 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 [75]. 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 [75], a bioinformatics pipeline that integrates several tools [43,76–80], 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]).

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 (10 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 5A]. Subsequent runs (T = 600 - 1,500) reused the precalculated features, taking only an average of ~4 hours runtime for each run [Figure 5B].



B

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, 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 + membership assignment) and other steps (input parsing, hmmscanning and features extraction).

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/) [81]. 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 6].

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 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).

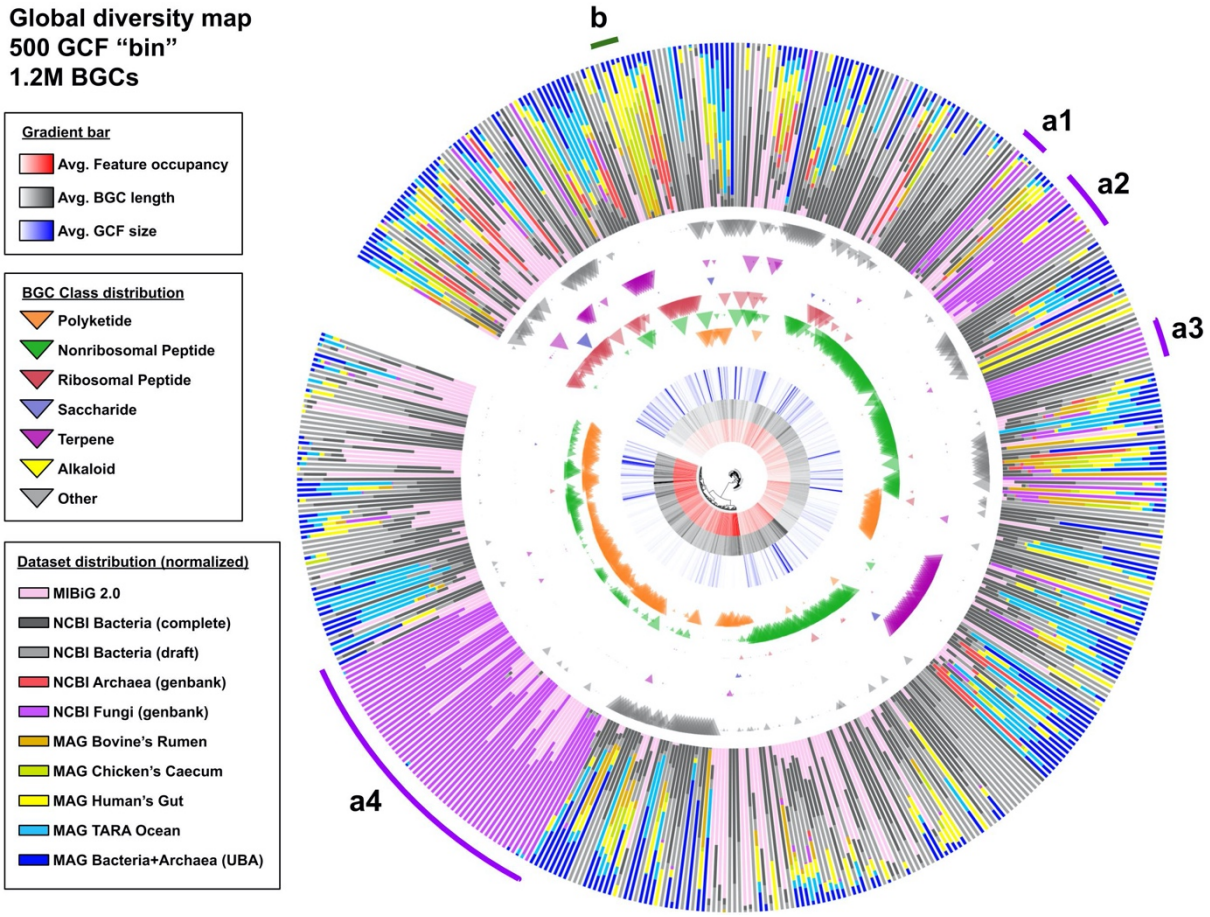


Figure 6. A phylogram created via the hierarchical clustering analysis of 500 GCF bins. The phylogram was rooted on a null (all zeros) dummy feature matrix. For each node, the raw dataset distribution values [Supplementary table 5] were double-normalized, first against the number of BGCs each dataset 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).

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 [82] (BGC0001399) and aspergillic acid [83] (BGC0001516) from *Aspergillus* (and on the bacterial side: four MIBiG BGCs including another simple proteasome inhibitor livipeptin [84,85] 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 [86]. 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 [87]. 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 [88–90]. However, no bacterial BGC from MIBiG, including those of known iterative type [91,92], 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 [93] (recently proposed to be reclassified as ranthipeptides [94]), 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. in solventogenic *Clostridia [95]* 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 [24,96] 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 [97] (which can, for example, be very powerful for analyzing RiPP precursor peptide variation [98–100]).

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 [23,24,101,102]. 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 7A].

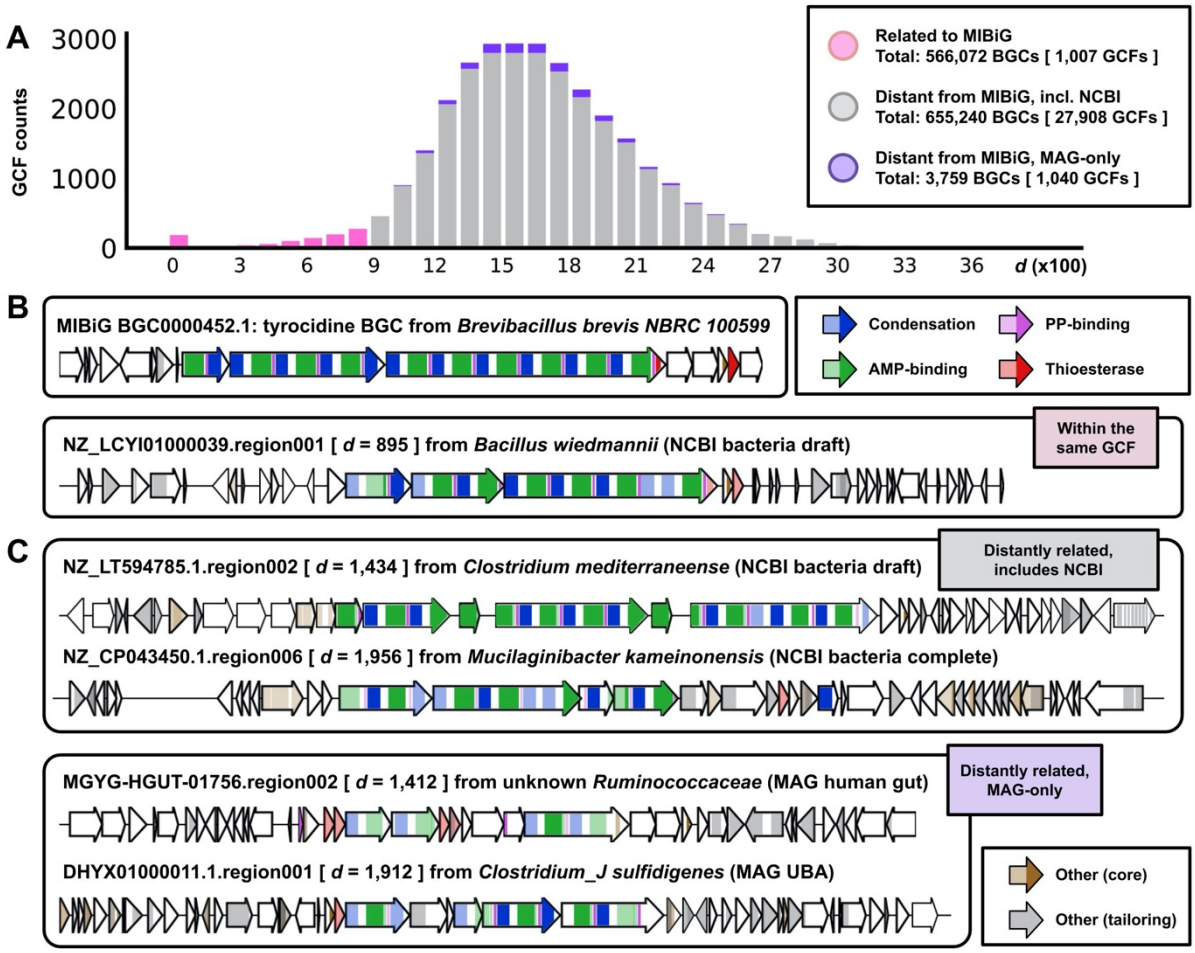


Figure 7. A. Histogram of Euclidean distances (x-axis) of GCF models to their closest BGC from the MIBiG 2.0 dataset. Here, all GCFs having *d* <= 900 were denoted as "Related to MIBiG" and "Distant from MIBiG" if otherwise, particularly highlighting those coming only from the MAG datasets. **B.** Selected 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.

Indeed, it is immediately clear from [Figure 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 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.

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 8A]. In this scenario, three *Firmicute*s (*Bacillus velezensis*, *Bacillus thuringiensis*, *Streptococcus pneumoniae*) and five *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 [103–105]).

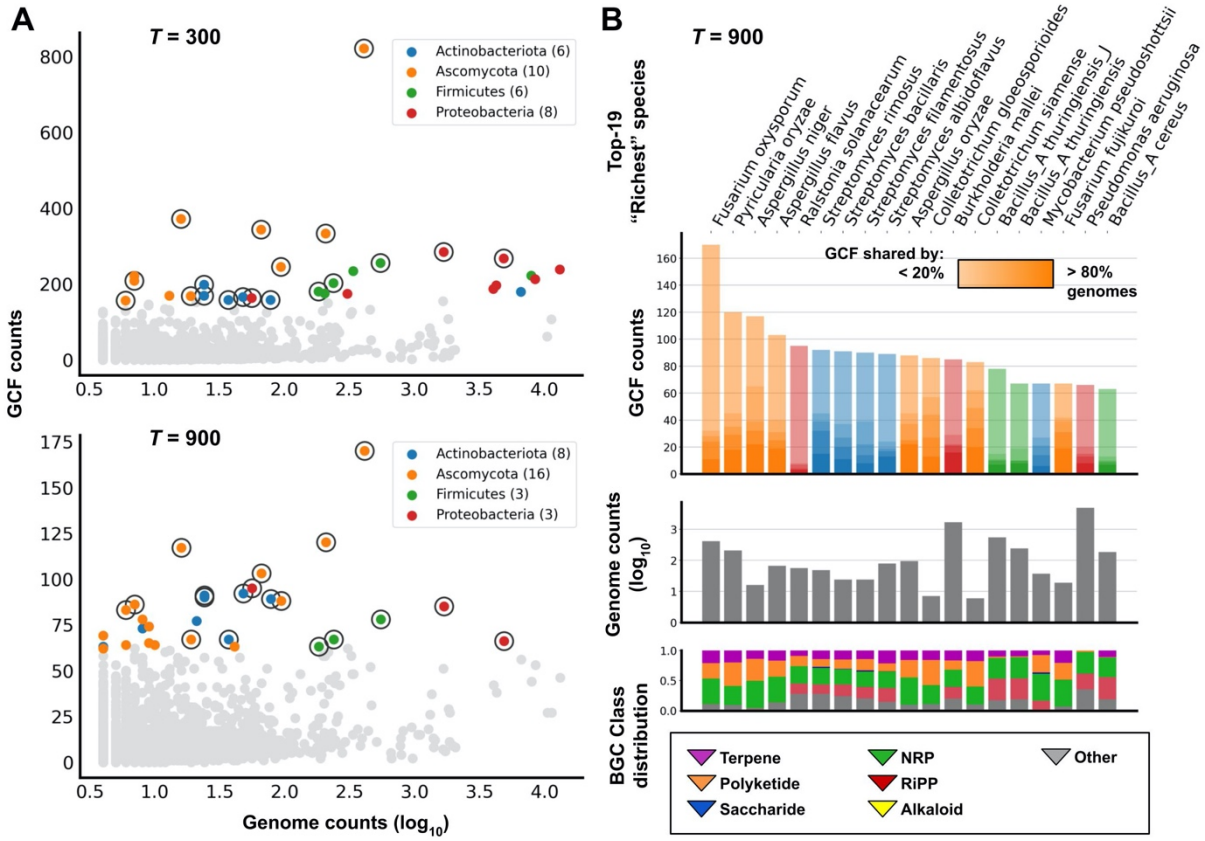


Figure 8. A. Distribution of GCF counts across species having four or more genomes in the dataset. Two plots showing results at the most stringent (*T* = 300) and a fairly lenient (*T* = 900) threshold, each highlighting 30 species with the highest GCF counts (colored dots). 19 species present in the top-30 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].

Four *Streptomyces* species made it into the selected list of 19 species that consistently ranked top-30 in both runs [Figure 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 [1,106]. 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).

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 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* [107] 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 [108] and a fungi-colonizing agent ralsolamycin [109] from a tomato-associated strain GMI1000, were deposited in MIBiG under accessions BGC0001014 and BGC0001363/1754), but several comparative genomic analyses [110,111] have linked this highly divergent metabolic capacities with their unusual ability to attack a vast range of plant species [112].

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 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. As fungi and bacteria seem to frequently compete with each other in the wild [113], it may be logical to increase the search 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.

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 plan to convert the precalculated global analysis result into a publicly accessible "reference" GCF database, 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 Data and Materials

Project name: BiG-SLiCE

Project home page: https://github.com/medema-lab/bigslice

Operating system(s): Linux / UNIX-based OS, output web app can be viewed on any modern Internet browsers

Programming language: Python

Other requirements: Python 3.6 or higher

License: GNU Affero General Public License v3.0

Input BGCs, analysis results and python scripts used to generate all figures and tables in this study is available at https://bioinformatics.nl/~kauts001/ltr/bigslice/paper_data/. An archived v1.0.0 release of the BiG-SLiCE software including the pHMM models used for this study can be downloaded from Zenodo [114].

Supplementary texts

All supplementary texts are available via BioRxiv.

Supplementary figures

Supplementary Figure 1. 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 2. 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 3. 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 4. 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 tables

Supplementary table 1. List of biosynthetic-Pfam pHMMs used by BiG-SLiCE.

Supplementary table 2. List of "core" biosynthetic-Pfam and the respective sub-Pfam pHMM models.

Supplementary table 3. List of genomes per dataset along with the total count of BGCs predicted by antiSMASH and their assigned taxonomy.

Supplementary table 4. Summary of five different run parameters on the full dataset of 1.2M BGCs.

Supplementary table 5. Calculated statististics of the 266 GCFs that were used to annotate the global phylogram map of biosynthetic diversity.

Supplementary table 6. 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 7. Unique GCF counts of species having at least 4 strain genomes in the full dataset.

Supplementary table 8. BGC class absence/presence distribution of species in the full dataset. Hybrid BGCs will have each of their classes counted separately, meaning the sum of the numbers will not be equal to the total number of BGCs per species.

Abbreviations

AMR: antimicrobial resistant; BGC: biosynthetic gene cluster; GCF: gene cluster family; GTDB: Genome Taxonomy Database; MAG: metagenome-assembled genome; NCBI: National Center for Biotechnology Information; NRPS: non-ribosomal peptide synthase; pHMM: protein hidden markov model; PKS: polyketide synthase; RiPP: ribosomally translated post-translationally modified peptide.

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.

Competing Interests

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

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. acknowledges funding by the Netherlands eScience Center (NLeSC) Accelerating Scientific Discoveries Grant [ASDI.2017.030].

Author's Contributions

SAK and MHM conceived the study. SAK designed and wrote the BiG-SLiCE software. SAK collected and processed all input data. SAK performed all analyses with the help and input from all other authors. JJJVDH and MHM provided input on the biochemical perspective of the study. DDR, JJJVDH and MHM provided input on the computational parts of the clustering algorithm. SAK wrote the initial draft of the paper. All authors contributed to writing and editing the final version of the manuscript.

References

1. Demain AL. Importance of microbial natural products and the need to revitalize their discovery. J Ind Microbiol Biotechnol. Springer Berlin Heidelberg; 2014;41:185–201.

2. Tanaka Y, Omura S. Agroactive compounds of microbial origin. Annu Rev Microbiol. 1993;47:57–87.

3. A review of soluble microbial products (SMP) in wastewater treatment systems. Water Res. Pergamon; 1999;33:3063–82.

4. Mukherjee AK, Das K. Microbial Surfactants and Their Potential Applications: An Overview. Biosurfactants. Springer, New York, NY; 2010. p. 54–64.

5. WHO | No Time to Wait: Securing the future from drug-resistant infections. World Health Organization; 2019 [cited 2020 Feb 18]; 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. National Academy of Sciences; 2016;113:5970–5.

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;92:229–65.

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 . Nature Publishing Group; 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. Nature Publishing Group; 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. Multidisciplinary Digital Publishing Institute; 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 . Nature Publishing Group; 2019;72:241–5.

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. Multidisciplinary Digital Publishing Institute; 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 . Nature Publishing Group; 2020;1–7.

14. Eckburg PB, Gill SR, Costello EK, Hsiao EY, Gopalakrishnan V, Matson V, et al. The Integrative Human Microbiome Project. Nature. Nature Publishing Group; 2019;569:641–8.

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. American Association for the Advancement of Science; 2011;332:1097–100.

16. pubmeddev, van Wezel GP DCA. Mining for Microbial Gems: Integrating Proteomics in the

Postgenomic Natural Product Discovery Pipeline. - PubMed - NCBI [Internet]. [cited 2020 Jan 29]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29708658

17. 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. National Academy of Sciences; 2017;114:E11121–30.

18. pubmeddev, Rochfort S. Metabolomics reviewed: a new “omics” platform technology for systems biology and implications for natural products research. - PubMed - NCBI [Internet]. [cited 2020 Jan 29]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16378385

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. Narnia; 2019;47:W81–7.

20. Christopher T. Walsh MAF. Natural Products Version 2.0: Connecting Genes to Molecules. J Am Chem Soc. NIH Public Access; 2010;132:2469.

21. Origin and evolution of operons and metabolic pathways. Res Microbiol. Elsevier Masson; 2009;160:502–12.

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. Nature Publishing Group; 2019;16:60–8.

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. Nature Publishing Group; 2014;10:963–8.

24. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell. Cell Press; 2014;158:412–21.

25. pubmeddev, Goering AW E al. Metabologenomics: Correlation of Microbial Gene Clusters with Metabolites Drives Discovery of a Nonribosomal Peptide with an Unusual Amino Acid Mon... - PubMed - NCBI [Internet]. [cited 2020 Jan 27]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27163034

26. 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. Nature Publishing Group; 2018;8:1–14.

27. Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. Chem Biol. Cell Press; 2015;22:460–71.

28. 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. Nature Publishing Group; 2017;2:1–9.

29. 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;11:3452–60.

30. pubmeddev, Parkinson EI E al. Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. - PubMed - NCBI [Internet]. [cited 2020 Jan 27]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29510029

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 [Internet]. American Society for Microbiology Journals; 2019 [cited 2020 Feb 3];4. Available from: https://msystems.asm.org/content/4/4/e00387-19.abstract

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. American Association for the Advancement of Science; 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. American Association for the Advancement of Science; 2019;366:606–12.

34. The long view on sequencing. Nat Biotechnol. Nature Publishing Group; 2018;36:287–287.

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. Narnia; 2019;47:D625–30.

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. Narnia; 2020;48:D422–30.

37. Papageorgiou L, Eleni P, Raftopoulou S, Mantaiou M, Megalooikonomou V, Vlachakis D. Genomic big data hitting the storage bottleneck. EMBnet.journal [Internet]. NIH Public Access; 2018 [cited 2020 Jan 29];24. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958914/

38. SQLite Home Page [Internet]. [cited 2020 Jan 27]. Available from: https://www.sqlite.org/index.html

39. 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. Narnia; 2019;47:D427–32.

40. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315:972–6.

41. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31:651–66.

42. Zhang T, Ramakrishnan R, Livny M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. 1996 [cited 2020 Jan 27]; Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.2504

43. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195.

44. Alborzi SZ, Devignes M-D, Ritchie DW. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics. 2017;18:107.

45. Katz L. Manipulation of Modular Polyketide Synthases. Chem Rev. 1997;97:2557–76.

46. 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. Public Library of Science; 2011;6:e18910.

47. 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;12:2825–30.

48. Variations on the Clustering Algorithm BIRCH. Big Data Research. Elsevier; 2018;11:44–53.

49. 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. p. 410–20.

50. Flask [Internet]. Pallets. [cited 2020 Jan 27]. Available from: https://palletsprojects.com/p/flask/

51. 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;8:1906–16.

52. 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:W546–52.

53. 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 [Internet]. Frontiers; 2020 [cited 2020 Aug 2];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00225/pdf

54. 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. Nature Publishing Group; 2017;2:1533–42.

55. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Scientific Data. Nature Publishing Group; 2018;5:1–8.

56. 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 [Internet]. bioRxiv. 2019 [cited 2020 Feb 18]. p. 762682. Available from: https://www.biorxiv.org/content/10.1101/762682v1.abstract

57. 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. Nature Publishing Group; 2019;37:953–61.

58. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. BioMed Central; 2020;21:1–16.

59. 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. PeerJ Inc.; 2020;8:e8614.

60. 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 [Internet]. bioRxiv. 2020 [cited 2020 Jul 28]. p. 2020.05.12.088096. Available from: https://www.biorxiv.org/content/10.1101/2020.05.12.088096v1.abstract

61. Anderson CL, Fernando SC. Insights into rumen microbial biosynthetic gene cluster diversity

through genome-resolved metagenomics [Internet]. bioRxiv. 2020 [cited 2020 Jul 28]. p. 2020.05.19.105130. Available from: https://www.biorxiv.org/content/10.1101/2020.05.19.105130v1.abstract

62. 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 [Internet]. bioRxiv. 2020 [cited 2020 Jul 28]. p. 2020.06.05.135962. Available from: https://www.biorxiv.org/content/10.1101/2020.06.05.135962v1.abstract

63. 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. Nature Publishing Group; 2020;1–11.

64. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol. Nature Publishing Group; 2020;38:701–7.

65. 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.

66. 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 [Internet]. 2020;9. Available from: http://dx.doi.org/10.1128/MRA.00333-20

67. 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;22:2693–708.

68. 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 [Internet]. 2020; Available from: http://dx.doi.org/10.1038/s41587-020-0603-3

69. 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;176:649–62.e20.

70. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568:505–10.

71. 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. Narnia; 2020;48:D454–8.

72. 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.

73. 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. Microbiology Society; 2018;68:2386.

74. 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. Oxford Academic; 2019;11:2531–41.

75. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics [Internet]. 2019 [cited 2020 Mar 4]; Available from: https://academic.oup.com/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz848/31199158/btz848.pdf

76. 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.

77. 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.

78. 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.

79. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.

80. 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.

81. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:W256–9.

82. 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;11:2275–84.

83. 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.

84. Cruz Morales P, Barona Gómez F, Ramos Aboites HE. GENETIC SYSTEM FOR PRODUCING A PROTEASES INHIBITOR OF A SMALL PEPTIDE ALDEHYDE TYPE [Internet]. World Patent. 2016 [cited 2020 Aug 6]. Available from: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016097957

85. 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;5:1165–75.

86. Bushley KE, Turgeon BG. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC Evol Biol. 2010;10:26.

87. Begley TP, editor. Polyketide Biosynthesis: Fungi. Wiley Encyclopedia of Chemical Biology. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. p. 380.

88. Chen H, Du L. Iterative polyketide biosynthesis by modular polyketide synthases in bacteria. Appl Microbiol Biotechnol. 2016;100:541–57.

89. Fisch KM. Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS. RSC Adv. 2013;3:18228.

90. 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. p. 154–66.

91. Liu W, Christenson SD, Standage S, Shen B. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science. 2002;297:1170–3.

92. 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.

93. 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.

94. 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;141:8228–38.

95. 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.

96. 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.

97. 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.

98. 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;13:470–8.

99. 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.

100. 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.

101. Baltz RH. Gifted microbes for genome mining and natural product discovery. J Ind Microbiol Biotechnol. 2017;44:573–88.

102. 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;114:5601–6.

103. 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. p. 155–74.

104. McDonald BR, Currie CR. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. MBio [Internet]. American Society for Microbiology; 2017 [cited 2020 Jul 29];8.

Available from: https://mbio.asm.org/content/8/3/e00644-17.abstract

105. 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 [Internet]. American Society for Microbiology; 2019 [cited 2020 Jul 29];10. Available from: https://mbio.asm.org/content/10/5/e01533-19.abstract

106. Antibiotics produced by Streptomyces. Braz J Infect Dis. Elsevier; 2012;16:466–71.

107. 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;55:1139–47.

108. 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;77:6117–24.

109. 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;10:2317–30.

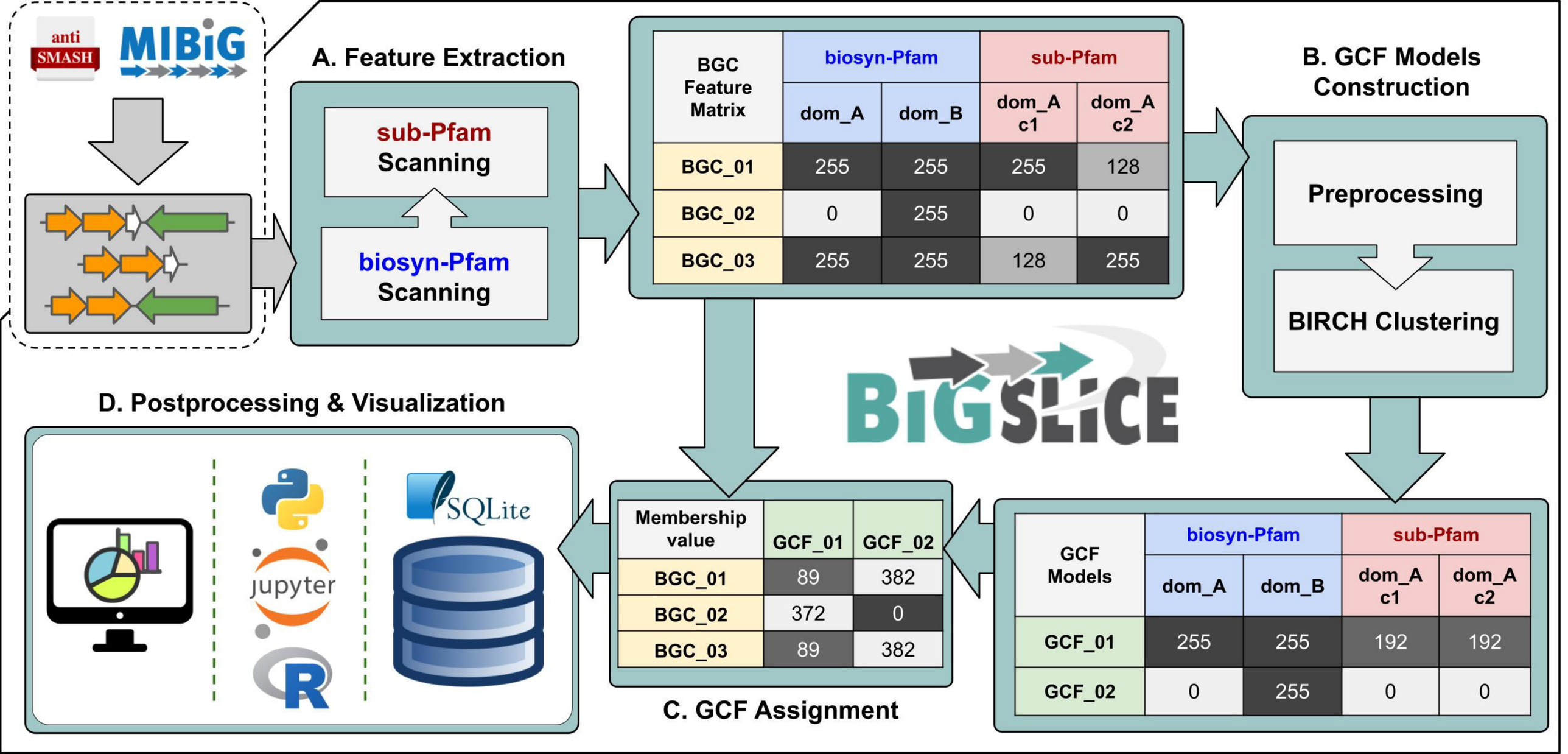
110. Prior P, Ailloud F, Dalsing BL, Remenant B, Sanchez B, Allen C. Genomic and proteomic evidence supporting the division of the plant pathogen Ralstonia solanacearum into three species. BMC Genomics. BioMed Central; 2016;17:1–11.

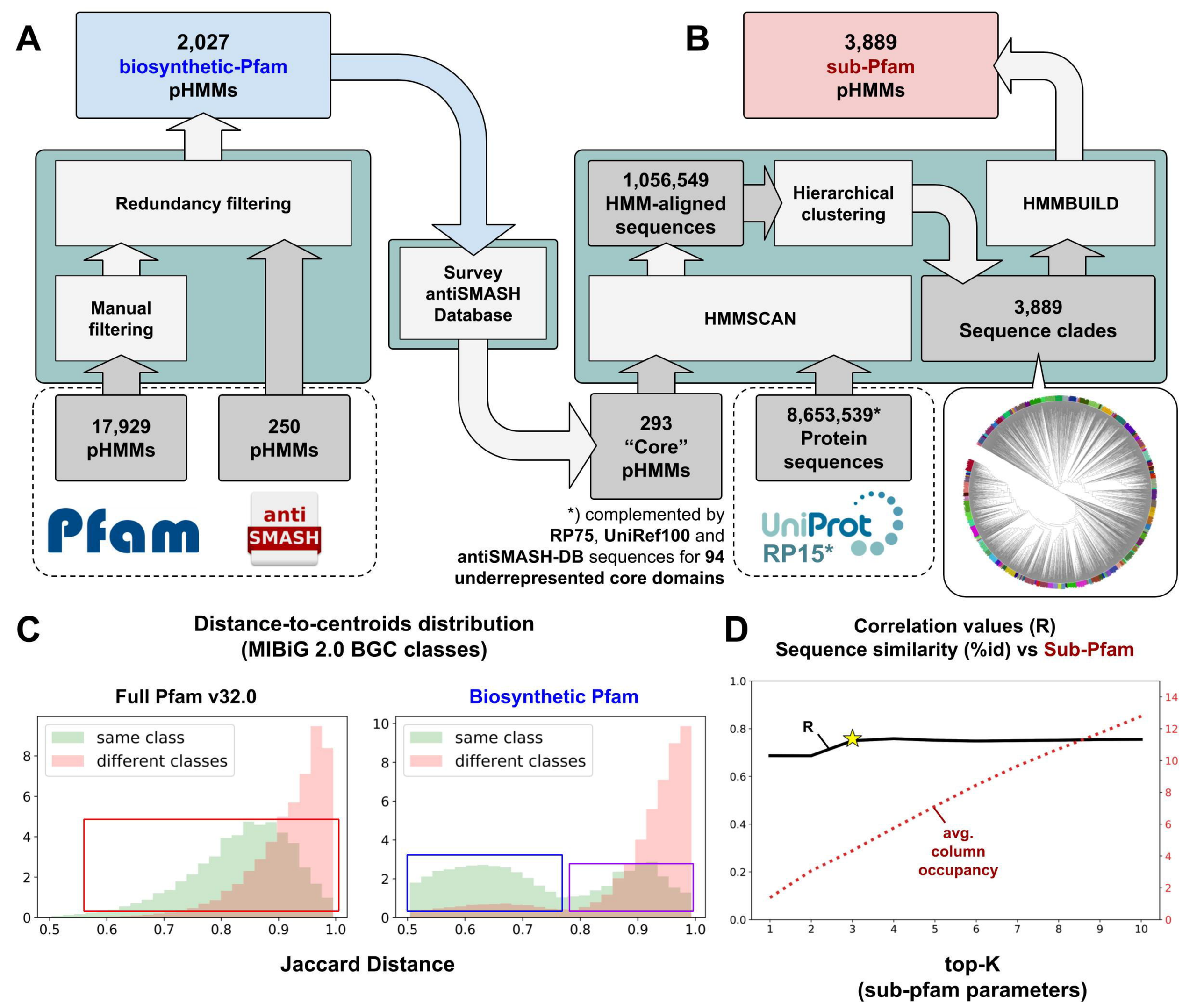
111. Remenant B, Coupat-Goutaland B, Guidot A, Cellier G, Wicker E, Allen C, et al. Genomes of three tomato pathogens within the Ralstonia solanacearum species complex reveal significant evolutionary divergence. BMC Genomics. BioMed Central; 2010;11:1–16.

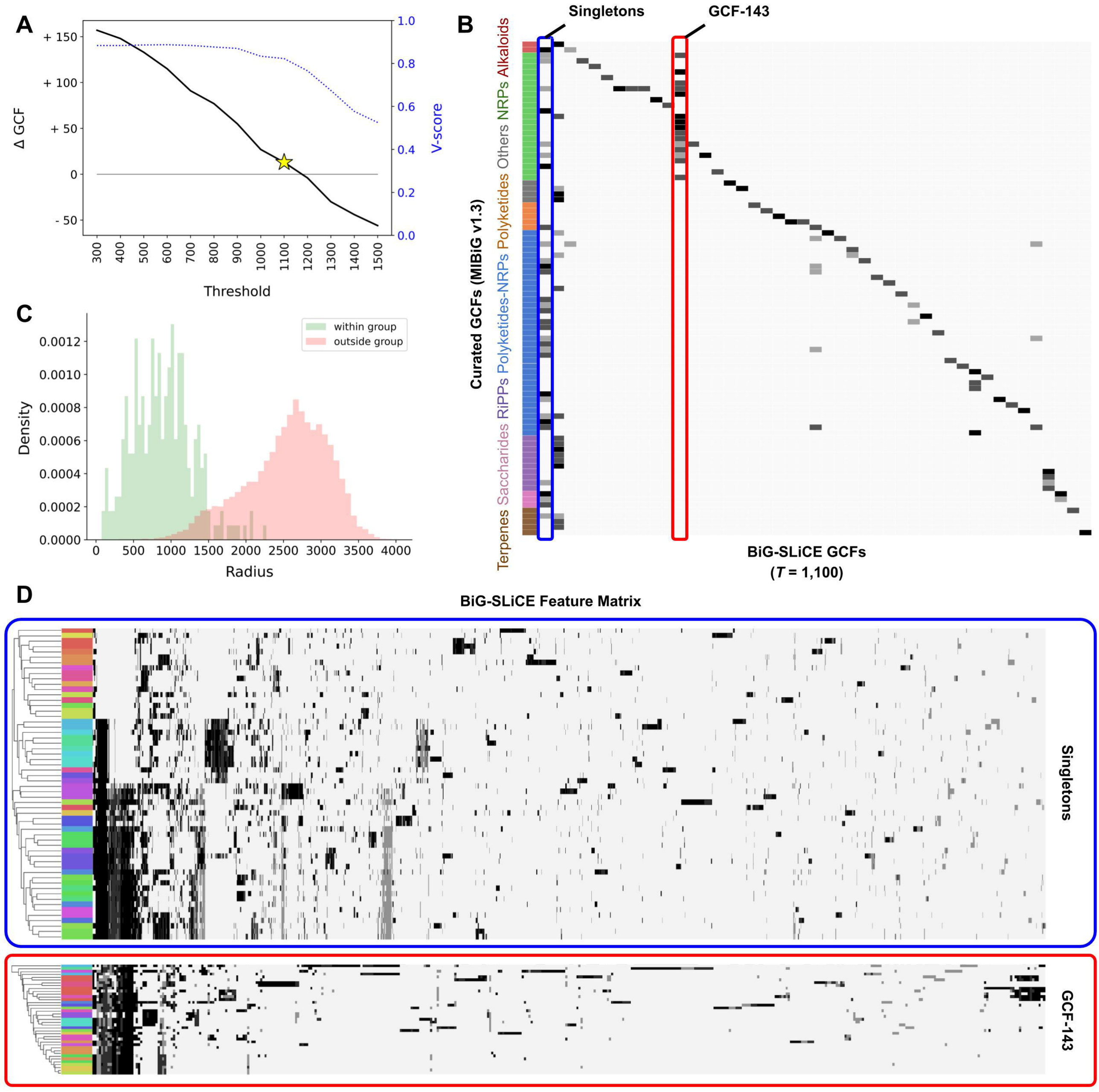
112. Hayward AC. Characteristics of Pseudomonas solanacearum. J Appl Bacteriol. 1964;27:265–77.

113. Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, et al. Structure and function of the global topsoil microbiome. Nature. 2018;560:233–7.

114. Kautsar SA. medema-group/bigslice: Version 1.0.0. 2020 [cited 2020 Aug 7]; Available from: https://zenodo.org/record/3975432


anti SMASH
MIBiG
A. Feature Extraction
sub-Pfam Scanning
biosyn-Pfam Scanning
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
B. GCF Models Construction
Preprocessing
BIRCH Clustering
BIGSLICE
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
C. GCF Assignment
D. Postprocessing & Visualization
jupyter
R
SQLite


A
2,027
biosynthetic-Pfam
pHMMs
Redundancy filtering
Manual filtering
17,929
pHMMs
250
pHMMs
Pfam
anti SMASH
Survey antiSMASH Database
293
"Core"
pHMMs
*) complemented by
RP75, UniRef100 and
antiSMASH-DB sequences for 94
underrepresented core domains
B
3,889
sub-Pfam
pHMMs
1,056,549
HMM-aligned
sequences
Hierarchical clustering
HMMBUILD
HMMSCAN
3,889
Sequence clades
8,653,539*
Protein
sequences
UniProt
RP15*
C
Distance-to-centroids distribution
(MIBiG 2.0 BGC classes)
Full Pfam v32.0
Biosynthetic Pfam
same class
different classes
same class
different classes
Jaccard Distance
D
Correlation values (R)
Sequence similarity (%id) vs Sub-Pfam
R
avg.
column
occupancy
top-K
(sub-pfam parameters)


A
Δ GCF
V-score
Threshold
B
Singletons
GCF-143
Alkaloids
NRPs
Others
Polyketides
Polyketides-NRPs
RiPPs
Saccharides
Terpenes
Curated GCFs (MIBiG v1.3)
BiG-SLiCE GCFs
(T = 1,100)
C
within group
outside group
Density
Radius
D
BiG-SLiCE Feature Matrix
Singletons
GCF-143

A

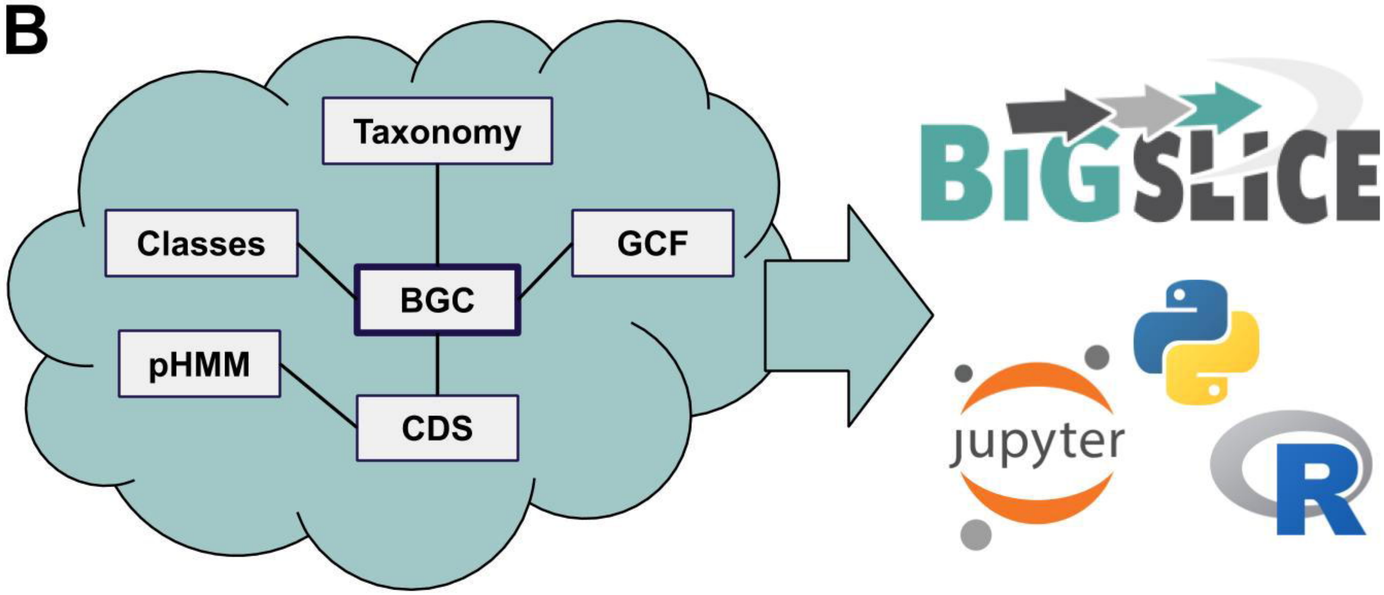
SQL 1

```
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"
```

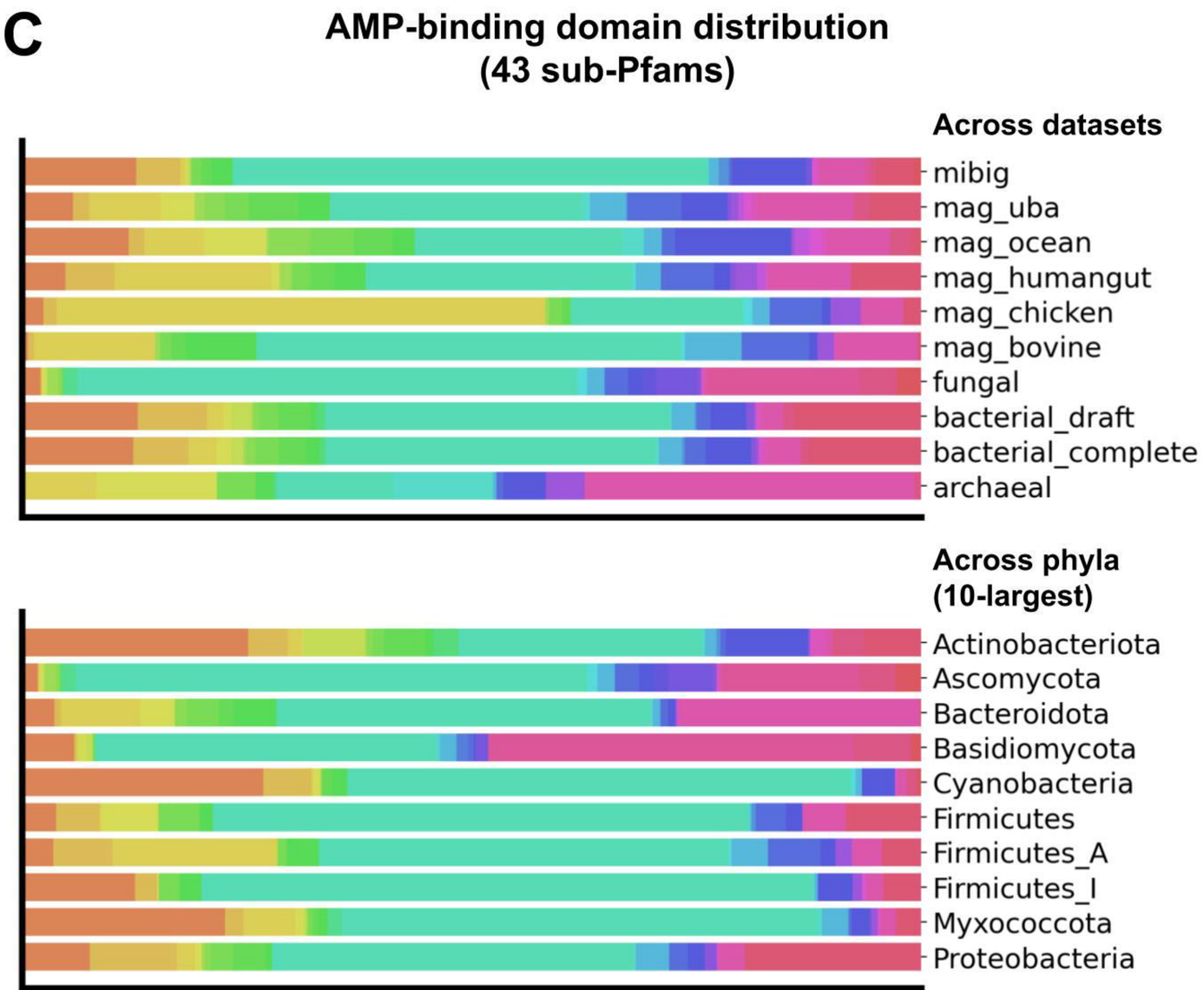
	bgc_name	locus_tag	aa_seq
18945	GCF_00194...	BWI70_RS...	MNGVDVALTAGRTDADEIAVVGL...
18946	GCF_00194...	BWI70_RS...	MEKREKLLDYLKWVTADLHRAEA...
18947	GCF_00194...	BWI70_RS...	MTNGEALRARLTEPAPTARHGVL...
18948	GCF_00194...	BWI70_RS...	PLWLGSVKSNLGHTQAAAGVAG...
18949	GCF_00194...	BWI70_RS...	QYRADAPPTLRTAAASLLAPRAD...
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"
```

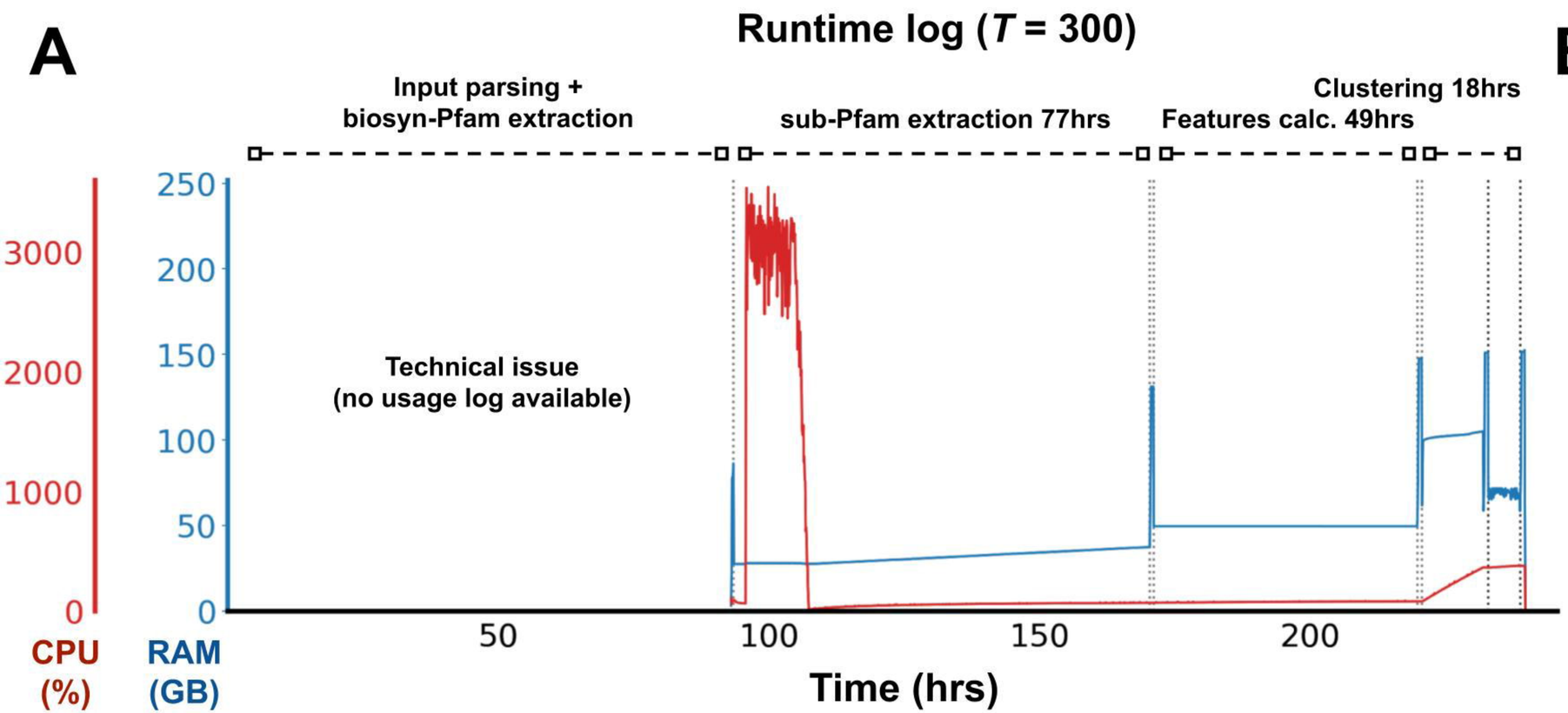
B



C



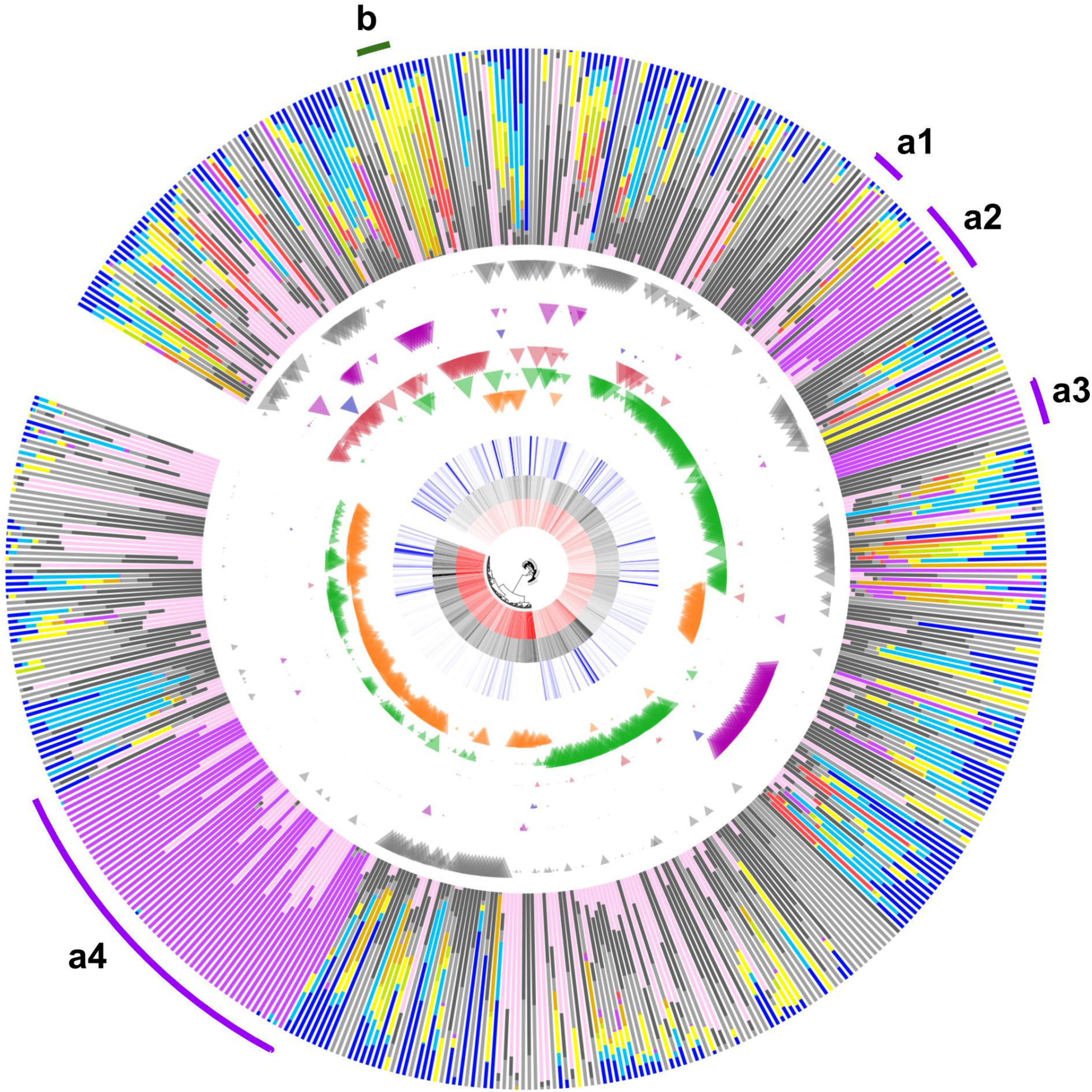
A



B

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

Global diversity map
500 GCF "bin"
1.2M BGCs



Gradient bar

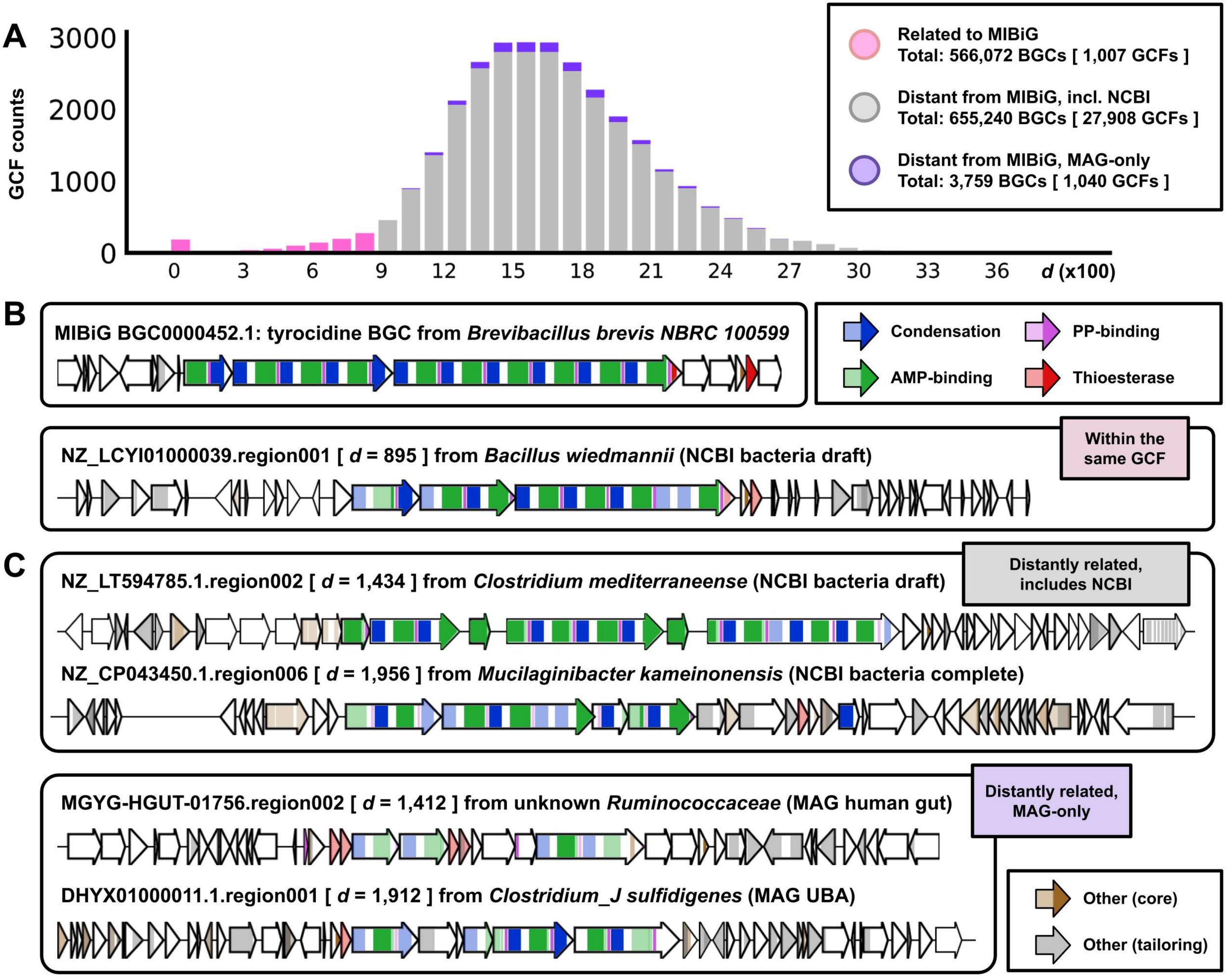
- Avg. Feature occupancy
- Avg. BGC length
- Avg. GCF size

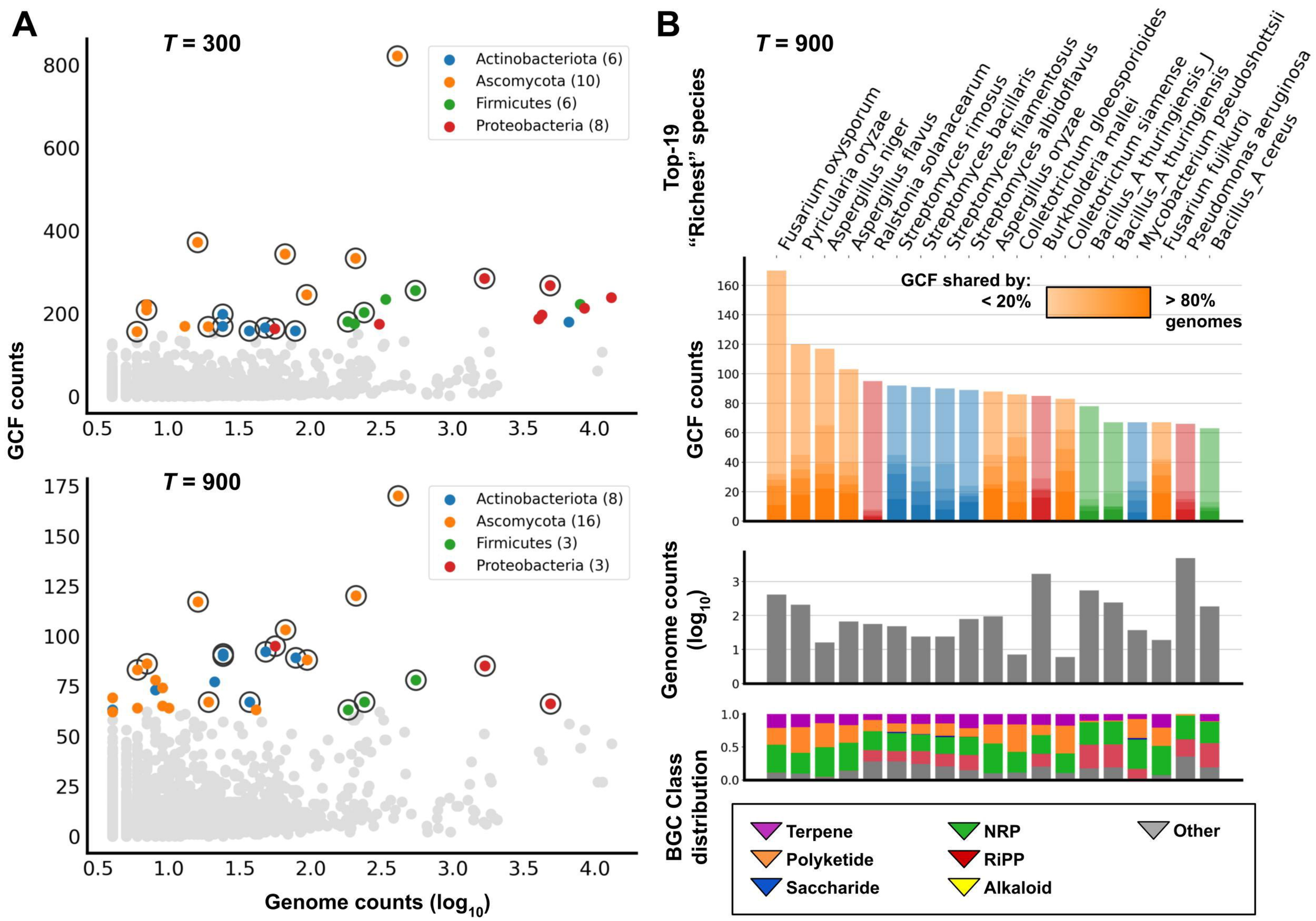
BGC Class distribution

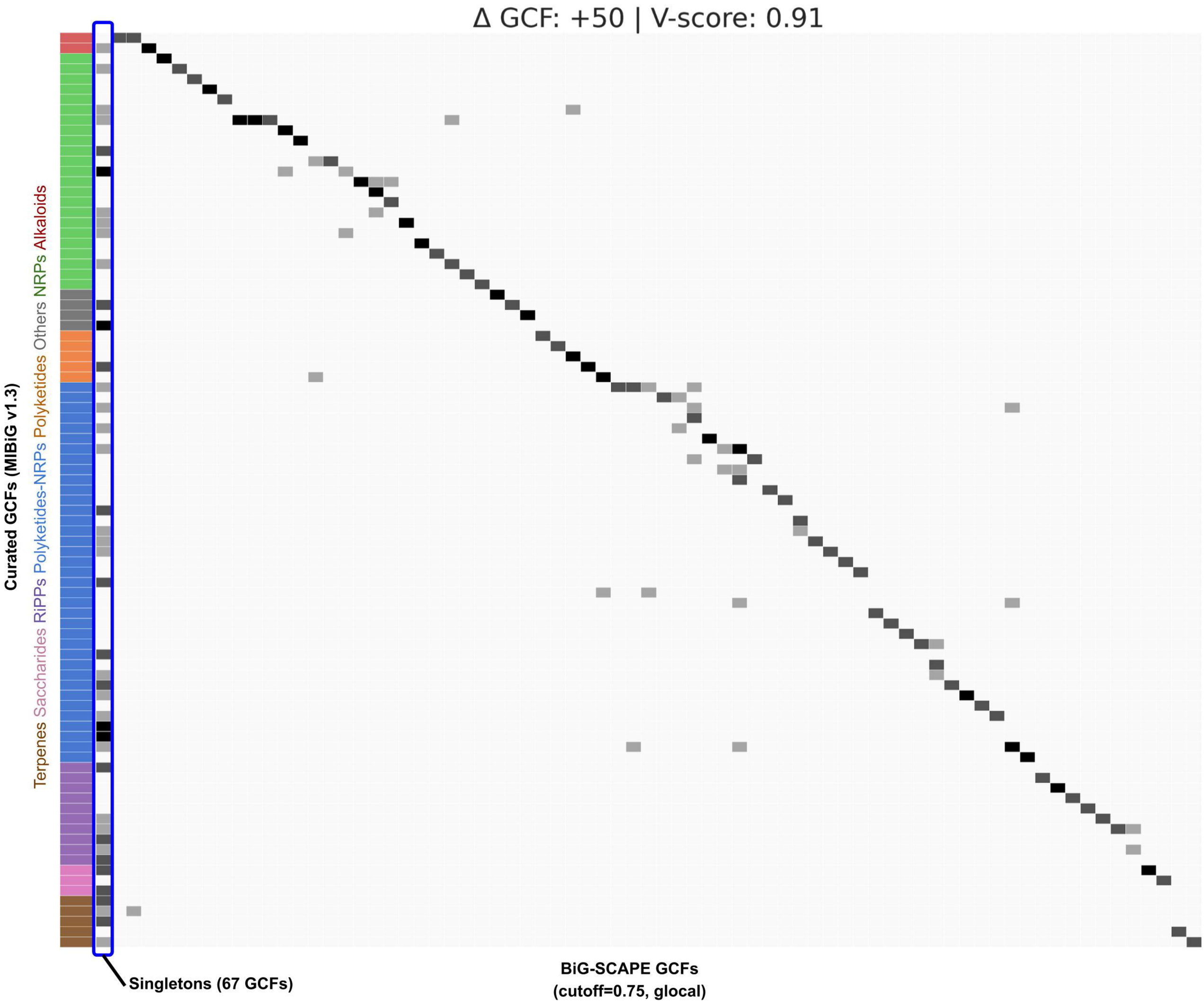
- Polyketide
- Nonribosomal Peptide
- Ribosomal Peptide
- Saccharide
- Terpene
- Alkaloid
- Other

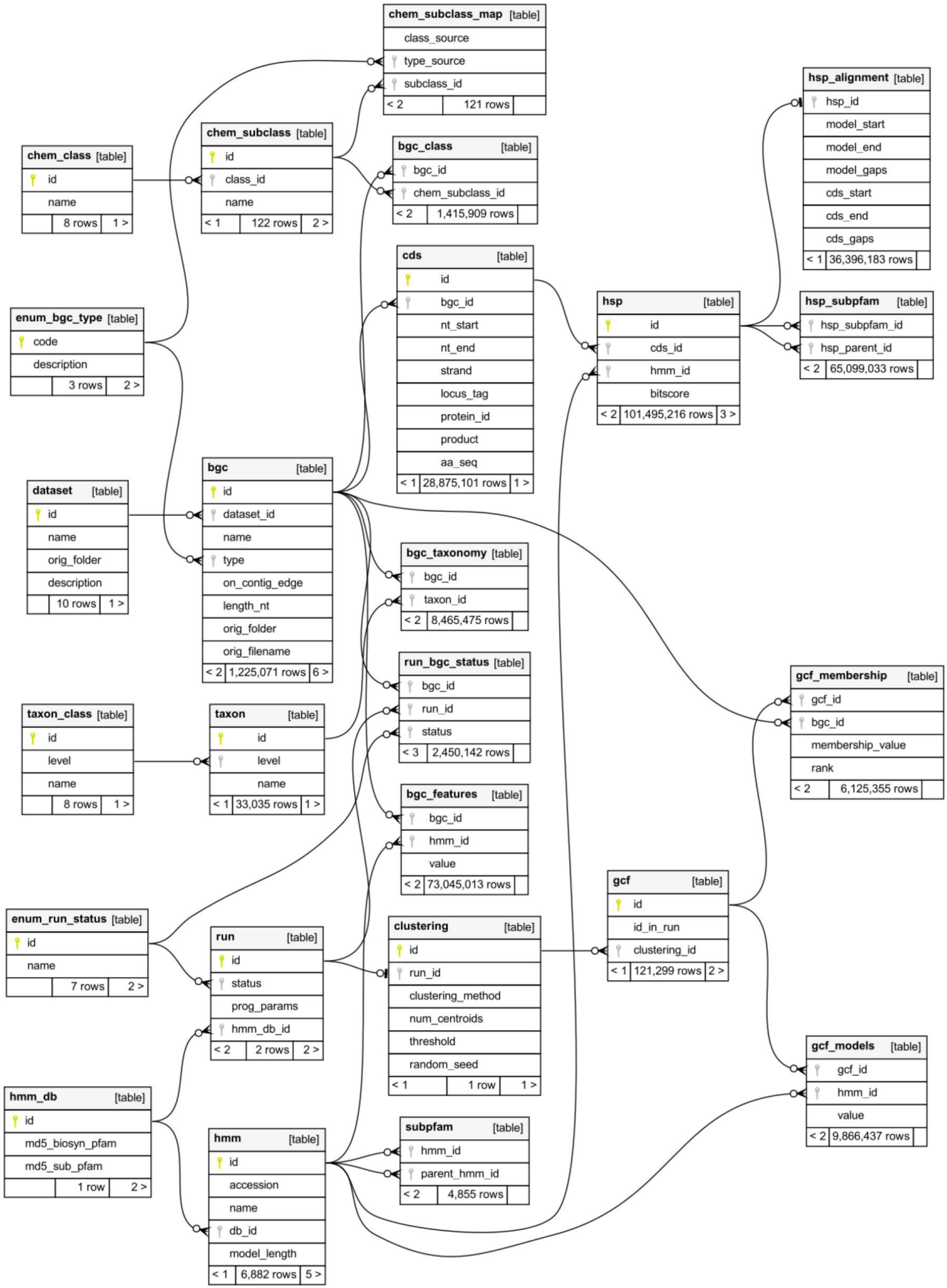
Dataset distribution (normalized)

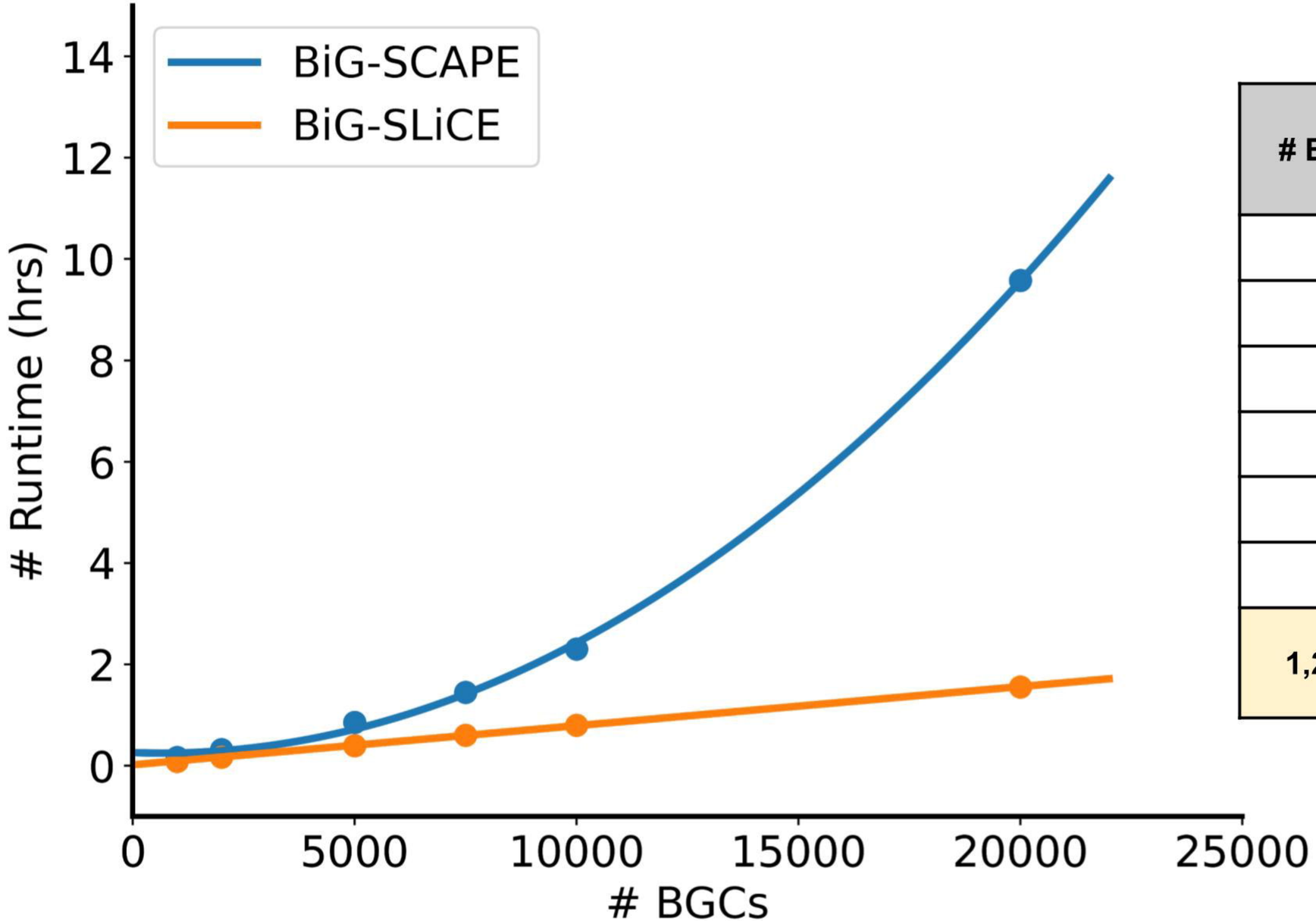
- MIBiG 2.0
- NCBI Bacteria (complete)
- NCBI Bacteria (draft)
- NCBI Archaea (genbank)
- NCBI Fungi (genbank)
- MAG Bovine's Rumen
- MAG Chicken's Caecum
- MAG Human's Gut
- MAG TARA Ocean
- MAG Bacteria+Archaea (UBA)


A
GCF counts
3000
2000
1000
0
0
3
6
9
12
15
18
21
24
27
30
33
36
d (x100)
Related to MIBiG
Total: 566,072 BGCs [1,007 GCFs]
Distant from MIBiG, incl. NCBI
Total: 655,240 BGCs [27,908 GCFs]
Distant from MIBiG, MAG-only
Total: 3,759 BGCs [1,040 GCFs]
B
MIBiG BGC0000452.1: tyrocidine BGC from Brevibacillus brevis NBRC 100599
Condensation
PP-binding
AMP-binding
Thioesterase
NZ_LCYI01000039.region001 [d = 895] from Bacillus wiedmannii (NCBI bacteria draft)
Within the same GCF
C
NZ_LT594785.1.region002 [d = 1,434] from Clostridium mediterraneense (NCBI bacteria draft)
Distantly related, includes NCBI
NZ_CP043450.1.region006 [d = 1,956] from Mucilaginibacter kameinonensis (NCBI bacteria complete)
MGYG-HGUT-01756.region002 [d = 1,412] from unknown Ruminococcaceae (MAG human gut)
Distantly related, MAG-only
DHYX01000011.1.region001 [d = 1,912] from Clostridium_J sulfidigenes (MAG UBA)
Other (core)
Other (tailoring)

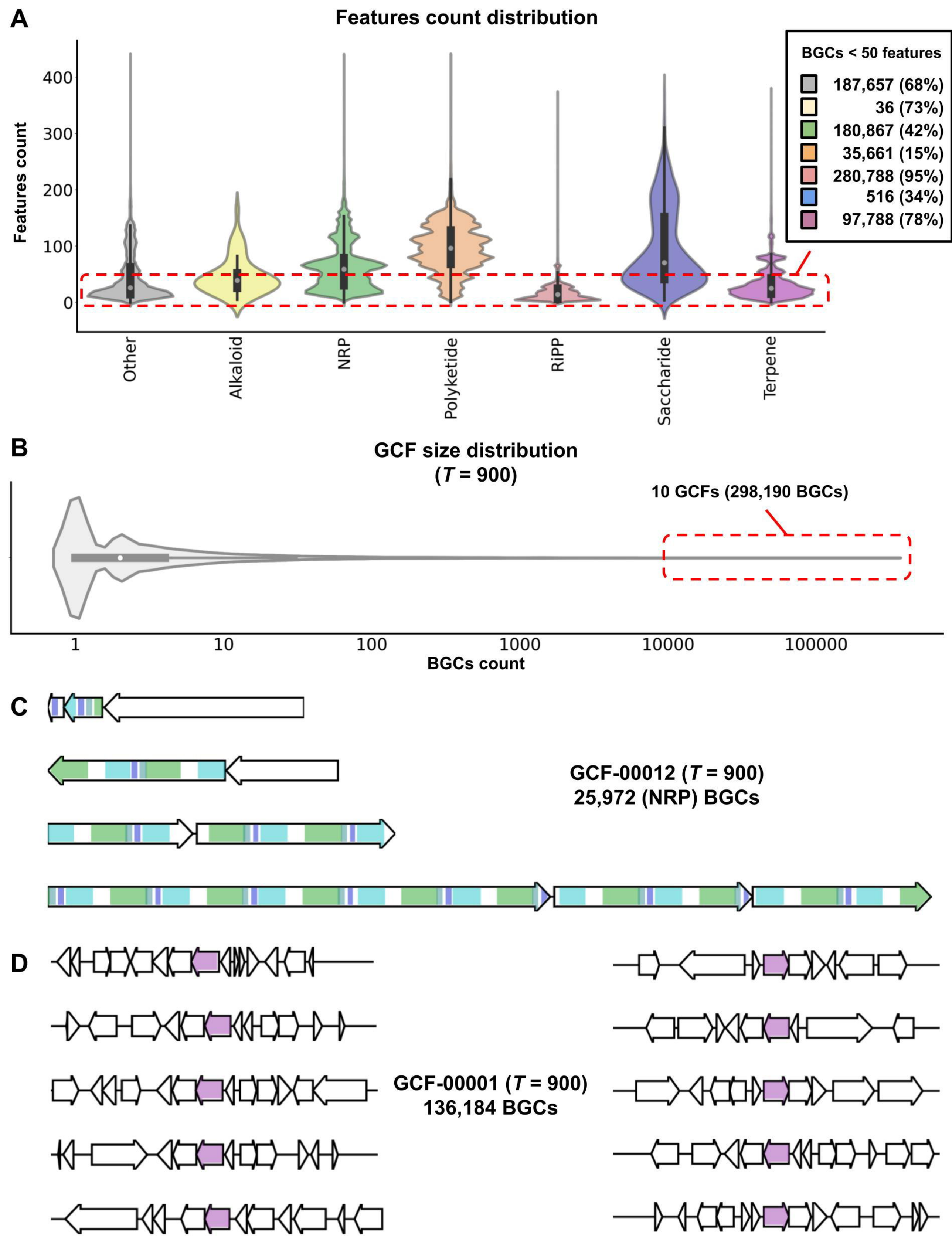

A
T = 300
Actinobacteriota (6)
Ascomycota (10)
Firmicutes (6)
Proteobacteria (8)
GCF counts
T = 900
Actinobacteriota (8)
Ascomycota (16)
Firmicutes (3)
Proteobacteria (3)
Genome counts (log10)
B
T = 900
Top-19 "Richest" species
Fusarium oxysporum
Pyricularia oryzae
Aspergillus niger
Aspergillus flavus
Ralstonia solanacearum
Streptomyces rimosus
Streptomyces bacillaris
Streptomyces filamentosus
Streptomyces albidoflavus
Aspergillus oryzae
Colletotrichum gloeosporioides
Burkholderia mallei
Colletotrichum siamense
Bacillus_A thuringiensis_J
Bacillus_A thuringiensis
Mycobacterium pseudoshottsii
Fusarium fujikuroi
Pseudomonas aeruginosa
Bacillus_A cereus
GCF shared by:
< 20%
> 80% genomes
GCF counts
Genome counts (log10)
BGC Class distribution
Terpene
Polyketide
Saccharide
NRP
RiPP
Alkaloid
Other


Δ GCF: +50 | V-score: 0.91
Curated GCFs (MIBiG v1.3)
Terpenes Saccharides RiPPs Polyketides-NRPs Polyketides Others NRPs Alkaloids
Singletons (67 GCFs)
BiG-SCAPE GCFs
(cutoff=0.75, glocal)


chem_subclass_map [table]
class_source
type_source
subclass_id
< 2 121 rows
chem_class [table]
id
name
8 rows 1 >
chem_subclass [table]
id
class_id
name
< 1 122 rows 2 >
bgc_class [table]
bgc_id
chem_subclass_id
< 2 1,415,909 rows
hsp_alignment [table]
hsp_id
model_start
model_end
model_gaps
cds_start
cds_end
cds_gaps
< 1 36,396,183 rows
cds [table]
id
bgc_id
nt_start
nt_end
strand
locus_tag
protein_id
product
aa_seq
< 1 28,875,101 rows 1 >
hsp [table]
id
cds_id
hmm_id
bitscore
< 2 101,495,216 rows 3 >
hsp_subpfam [table]
hsp_subpfam_id
hsp_parent_id
< 2 65,099,033 rows
enum_bgc_type [table]
code
description
3 rows 2 >
bgc [table]
id
dataset_id
name
type
on_contig_edge
length_nt
orig_folder
orig_filename
< 2 1,225,071 rows 6 >
dataset [table]
id
name
orig_folder
description
10 rows 1 >
bgc_taxonomy [table]
bgc_id
taxon_id
< 2 8,465,475 rows
run_bgc_status [table]
bgc_id
run_id
status
< 3 2,450,142 rows
gcf_membership [table]
gcf_id
bgc_id
membership_value
rank
< 2 6,125,355 rows
taxon_class [table]
id
level
name
8 rows 1 >
taxon [table]
id
level
name
< 1 33,035 rows 1 >
bgc_features [table]
bgc_id
hmm_id
value
< 2 73,045,013 rows
gcf [table]
id
id_in_run
clustering_id
< 1 121,299 rows 2 >
enum_run_status [table]
id
name
7 rows 2 >
run [table]
id
status
prog_params
hmm_db_id
< 2 2 rows 2 >
clustering [table]
id
run_id
clustering_method
num_centroids
threshold
random_seed
< 1 1 row 1 >
gcf_models [table]
gcf_id
hmm_id
value
< 2 9,866,437 rows
hmm_db [table]
id
md5_biosyn_pfam
md5_sub_pfam
1 row 2 >
hmm [table]
id
accession
name
db_id
model_length
< 1 6,882 rows 5 >
subpfam [table]
hmm_id
parent_hmm_id
< 2 4,855 rows



# BGCs	Runtime (hours)	
	BiG-SCAPE	BiG-SLiCE
1,000	0.16	0.09
2,000	0.32	0.17
5,000	0.86	0.40
7,500	1.45	0.60
10,000	2.30	0.80
20,000	9.58	1.55
1,225,071	**37,280.73** (fitted)	**238.74**


A
Features count distribution
Features count
400
300
200
100
0
BGCs < 50 features
187,657 (68%)
36 (73%)
180,867 (42%)
35,661 (15%)
280,788 (95%)
516 (34%)
97,788 (78%)
Other
Alkaloid
NRP
Polyketide
RiPP
Saccharide
Terpene
B
GCF size distribution
(T = 900)
10 GCFs (298,190 BGCs)
1
10
100
1000
10000
100000
BGCs count
C
GCF-00012 (T = 900)
25,972 (NRP) BGCs
D
GCF-00001 (T = 900)
136,184 BGCs