Published in final edited form as:
*Nat Chem Biol*. 2020 January 01; 16(1): 60–68. doi:10.1038/s41589-019-0400-9.

# A computational framework to explore large-scale biosynthetic diversity

**Jorge C. Navarro-Muñoz**[1,2,#], **Nelly Selem-Mojica**[3,#], **Michael W. Mullowney**[4,#], **Satria A. Kautsar**[1], **James H. Tryon**[4], **Elizabeth I. Parkinson**[5,$], **Emmanuel L.C. De Los Santos**[6], **Marley Yeong**[1], **Pablo Cruz-Morales**[3], **Sahar Abubucker**[7,9], **Arne Roeters**[1], **Wouter Lokhorst**[1], **Antonio Fernandez-Guerra**[8], **Luciana Teresa Dias Cappelini**[4], **Anthony W. Goering**[4], **Regan J. Thomson**[4], **William W. Metcalf**[5], **Neil L. Kelleher**[4,*], **Francisco Barona-Gomez**[3,*], **Marnix H. Medema**[1,*]

[1]Bioinformatics Group, Wageningen University, The Netherlands [2]Fungal Natural Products Group, Westerdijk Fungal Biodiversity Institute, The Netherlands [3]Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, México [4]Department of Chemistry, Northwestern University, Evanston, Illinois, United States [5]Carl R. Woese Institute for Genomic Biology and Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States [6]Warwick Integrative Synthetic Biology Centre, University of Warwick, Coventry, United Kingdom [7]Novartis Institutes for BioMedical Research, Cambridge, United States [8]Microbial Genomics and Bioinformatics, Max Planck Institute for Marine Microbiology, Bremen, Germany

## Abstract

Genome mining has become a key technology to exploit natural product diversity. While initially performed on a single-genome basis, the process is now being scaled up to mine entire genera, strain collections and microbiomes. However, no bioinformatic framework is currently available for effectively analyzing datasets of this size and complexity. Here, we provide a streamlined computational workflow consisting of two new software tools: The ‘Biosynthetic Gene Similarity Clustering And Prospecting Engine’ (BiG-SCAPE) facilitates fast and interactive sequence similarity network analysis of biosynthetic gene clusters and gene cluster families. ‘CORe



[*]Joint corresponding authors: .
[9]Present Address: Sanofi, Cambridge, United States
[#]Joint first authors
[$]Current address: Department of Chemistry, Purdue University, West Lafayette, Indiana, United States

**Author contributions**

RJT, WWM, NLK, FBG and MHM originally conceived of the research and coordinated the work. JCNM designed and developed BiG-SCAPE, with help of SAK, ELCdlS, MY, SA, AR, WL, AF and MHM. SAK designed the output visualisations with help of JCNM and ELCdlS. NSM designed and developed CORASON, with help of PCM and FBG. MWM, JHT, EIP, LTDC, AWG, RJT, WWM and NLK designed and performed the experimental research. JCNM, NSM, MWM, JHT, FBG and MHM wrote the first draft of the manuscript, and all authors participated in editing the manuscript.

**Conflicts of Interest**

MHM is on the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals. NLK, WWM and RJT are on the board of directors of MicroMGx.

Analysis of Syntenic Orthologues to prioritize Natural product gene clusters' (CORASON) elucidates phylogenetic relationships within and across these families. We validate BiG-SCAPE by correlating its output to metabolomic data across 363 actinobacterial strains and demonstrate the discovery potential of CORASON by comprehensively mapping biosynthetic diversity across a range of detoxin/rimosamide-related gene cluster families, culminating in the characterization of seven novel analogues.

## Introduction

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

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

Initially, genome mining was performed on a single-genome basis: a research group or consortium would sequence the genome of a single microbial strain and attempt to identify and characterize each of its BGCs one by one. This approach has revealed much about the metabolic capacities of model natural-product-producing organisms like *Streptomyces coelicolor*, *Sorangium cellulosum* and *Aspergillus nidulans*, and has provided clues regarding the discovery potential[10] from corresponding genera[11–13]. Computational tools for the identification of BGCs and the prediction of their products' chemical structures, such as antiSMASH[14–17] and PRISM[18–20], have played a key role in the success of genome mining. These *in silico* approaches have been strengthened by comparative analysis of identified BGCs with biochemical reference data, such as those provided by the MIBiG community effort[21].

Fueled by rapid developments in high-throughput sequencing, genome mining efforts are now expanding to large-scale pan-genomic mining of entire bacterial genera[4,22,23], strain collections[24] and metagenomic datasets from which thousands of metagenome-assembled genomes (MAGs) can be extracted at once[25–28]. Such studies pave the path towards systematic investigations of the biosynthetic potential of broad taxonomic groups of organisms, as well as entire ecosystems. These large-scale analyses easily lead to the identification of thousands of BGCs with varying degrees of mutual similarity, ranging from widely distributed homologues of gene clusters for the production of well-known molecules to rare or unique gene clusters that encode unknown enzymes and pathways.




Europe PMC Funders Author Manuscripts

To map and prioritize this complex biosynthetic diversity, several groups have devised methods to compare architectural relationships between BGCs in sequence similarity networks and group them into gene cluster families (GCFs), each of which contains BGCs across a range of organisms that are linked to a highly similar natural product chemotype[3,4,29,30]. Such GCFs can be matched to molecular families (MFs) identified from mass spectrometry (MS) data based on observed/predicted chemical features[31–33]. Alternatively, their presence or expression can be statistically correlated to the presence of MFs in MS data in a process termed metabologenomics[4,34–36]. However, current methods fail to correctly measure the similarity between complete and fragmented gene clusters (which frequently occur in metagenomes and large-scale pan-genome sequencing projects based on short-read technologies), do not consider the complex and multi-layered evolutionary relationships within and between GCFs, require lengthy CPU-time on supercomputers when processing large datasets, and lack a user-friendly implementation that interacts directly with other key resources. These shortcomings preclude adoption by the broader scientific community and impede substantial advances in natural product discovery.

Here, we provide a streamlined computational workflow that tightly integrates two new software tools, BiG-SCAPE and CORASON (https://bigscape-corason.secondarymetabolites.org/), with the gene cluster identification and empirical biosynthetic data comparison possible through antiSMASH[17] and MIBiG[21] (Fig. 1). BiG-SCAPE facilitates rapid calculation and interactive exploration of BGC sequence similarity networks; it accounts for differences in modes of evolution between BGC classes, groups gene clusters at multiple hierarchical levels, and introduces a ‘glocal’ alignment mode to handle fragmented BGCs. CORASON employs a phylogenomic approach to elucidate evolutionary relationships between gene clusters by computing high-resolution multi-locus phylogenies of BGCs within and across GCFs; additionally, it allows researchers to comprehensively identify all genomic contexts in which gene cassettes of interest (‘sub-clusters’ within larger BGCs) can be found. We confirm that metabologenomic correlations accurately connect GCFs to mass features across metabolomic data from 363 strains. Furthermore, we demonstrate the power of the combined workflow, together with the EvoMining algorithm[8], to comprehensively map biosynthetic diversity, by identifying three new families responsible for the biosynthesis of new detoxins.

## Results

### Large-scale network analysis and classification of BGCs

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

In previous studies[3, 4], two sets of distance metrics had been independently developed to measure the (dis)similarity of pairs of BGCs. In BiG-SCAPE, we aimed to combine the respective strengths of both approaches. The strength of the former approach was the elegant compression of gene clusters into strings of Pfam domains[37], combined with the Jaccard

index (JI) to measure domain content similarity (Fig. 2a). However, an informative index for synteny conservation had been missing. To this end, we added an Adjacency Index (AI), which measures how many pairs of adjacent domains are shared between gene clusters (Supplementary Note 1). Also, sequence identity is an important parameter, as Pfam domains are often very broad and frequently comprise a wide range of enzyme subfamilies with different catalytic activities or substrate specificities. Yet, previous approaches suffered from extremely long compute times when including sequence identity calculations, requiring the use of supercomputers that would preclude day-to-day use. The underlying issue is that comparing large numbers of protein sequences from many BGCs is an all-versus-all problem that scales quadratically when the size of the data increases. To mitigate this, we replaced all-versus-all calculations with all-versus-profile calculations, by aligning each protein domain sequence to its profile Hidden Markov Model from Pfam using the hmmalign tool from the HMMER suite (hmmer.org). This leads to a marked speed increase compared to conventional multiple sequence alignment using Muscle or MAFFT, especially for large numbers of sequences (see Online Methods). We implemented the profile-based alignment into the domain sequence similarity (DSS) index, which measures both Pfam domain copy number differences and sequence identity. The combination of JI, AI and DSS indices into a new combined metric constitutes a fast and informative method to calculate distances between BGCs. BiG-SCAPE obtains very similar results in a fraction of time when compared to the previously published method[4] (see Online Methods).

One notable limitation of a generic distance metric is that different classes of BGCs have different evolutionary dynamics. For example, the chemical structures of aryl polyenes have been shown to remain very stable across large evolutionary timescales, while the amino acid sequence identity between their key biosynthetic enzymes has become less than 30-40%[3]. On the other hand, the structures of rapamycin-family polyketides exhibit major differences even when sequence identities are as high as ~80%[38]. Although there is not enough information available to construct individual metrics for each specific natural product family, we did calibrate specific weights of the JI, AI and DSS indices for BGCs encoding eight different BiG-SCAPE classes (type I polyketide synthases (PKS), other PKSs, nonribosomal peptide synthetases (NRPS), PKS/NRPS hybrids, RiPPs, saccharides, terpenes and others; Fig. 2b) by choosing the weight combination that maximized the correlation between BGC and Compound distances for every pair of BGCs in the same class (Online Methods, Supplementary Note 2). In the BiG-SCAPE output, separate networks are generated for each BiG-SCAPE class, along with an optional overall network that combines BGCs from all classes (Supplementary Table 1).

Another problem of previous approaches for calculating distances between BGCs was how to handle comparisons between complete and partial BGCs (e.g., from fragmented genome assemblies), as well as comparisons with pairs of genomically adjacent BGCs that are merged by BGC identification tools. Both global similarity (used in all previous methods) and local similarity lead to artifacts in such cases. To compare the appropriate corresponding regions between BGCs, we introduced a new ‘glocal’ alignment mode, which first finds the longest common substring between the Pfam strings of a BGC pair, and then uses match/mismatch penalties to extend this alignment (Fig. 2c, Online Methods). Information about whether an antiSMASH-annotated cluster is located at the edge of a contig can also be used

Europe PMC Funders Author Manuscripts

to automatically select a third pairwise distance calculation mode that relies on global alignment for complete clusters and glocal alignment when at least one of the BGCs in a pair is fragmented.

BGC sequence similarity networks are then generated by applying a cutoff to the distance matrix calculated by BiG-SCAPE. Subsequently, two rounds of affinity propagation clustering[39] are performed to group BGCs into GCFs, and GCFs into "Gene Cluster Clans" (GCCs) (see Online Methods). While tighter (lower) cutoffs are more appropriate at grouping BGCs that produce identical compounds, looser (higher) cutoffs provide a broader perspective on related families of natural products. This process of categorization facilitates calculating metabologenomic correlations[35,40] at multiple levels.

## Validation using large-scale metabolomics data

To verify that BiG-SCAPE can group BGCs that are known to be related, we constructed a chemical similarity network from all products of BGCs in MIBiG (Online Methods), and used this to derive a curated set of 376 compounds, which were manually classified into 92 groups (e.g. 14-membered macrolides, benzoquinone ansamycins, quinomycin antibiotics etc.; see Supplementary Dataset) and 9 classes (e.g. polyketides, NRPs, RiPPs etc.). We then used BiG-SCAPE to group the corresponding BGCs into GCFs and observed good correspondence between manually curated families and those predicted by BiG-SCAPE (Supplementary Fig. 1).

Arguably, the greatest value of BiG-SCAPE lies in the practical utility of the predicted GCFs for discovery applications. Hence, we assessed the accuracy of correlations of BiG-SCAPE-predicted GCFs to MS ions from known natural products through metabologenomics[35]. First, we performed a BiG-SCAPE analysis of 74,652 BGCs from 3,080 actinobacterial genomes (see Online Methods), including 1,393 reference BGCs from MIBiG[21]. BiG-SCAPE grouped these BGCs into a total number of 17,718 GCFs and 801 GCCs using default parameters. Extracts from 363 actinomycete strains were analyzed using untargeted high-resolution LC-MS/MS[4,35,40,41]. Exploration of gene cluster networks and molecular networks[42] highlighted high diversity in both gene clusters and molecules; for example, 105 different BGCs were identified (at default <0.3 distance) related to known detoxin/rimosamide gene clusters (Fig. 3a,b), and 110 different molecules were identified related to detoxins and rimosamides (Fig. 3c). The GCF annotations for all 363 strains from two BiG-SCAPE modes (global and glocal) at two distance cutoffs (0.30 and 0.50) were used to generate and compare four rounds of metabologenomic correlations utilizing a binary scoring metric (Supplementary Fig. 2) as described previously[4,35]. BiG-SCAPE's gene cluster family annotations were then assessed against ion production patterns. A test dataset of nine known ion signals and their characterized gene clusters (for CE-108, benarthin, desertomycin, tambromycin, enterocin, tyrobetaine, chlortetracycline, rimosamide, and oxytetracycline), which were known to be present across multiple strains in the data, were manually tracked across the four correlation rounds. Based on the metabologenomic analysis of the four rounds, the glocal mode with a 0.3 distance cutoff (Fig. 3d) was chosen as default for BiG-SCAPE (Supplementary Tables 2 and 3). Using these parameters, the analysis showed that at least 6 out of these 9 molecule-GCF combinations ended up in the rightmost

tail of the distribution of all correlation values, which would indicate a possible/likely connection if it were used as a prediction (Fig. 3d, Online Methods).

## BGC phylogenies resolve evolutionary relationships

Genetic diversity of BGCs within GCFs is often directly related to structural differences between their molecular products, and even small chemical variations can lead to different biological activities[38]. Hence, mapping the evolutionary relationships between BGCs within and across GCFs is crucial for the discovery process. To this end, we introduce the 'CORe Analysis of Syntenic Orthologues to prioritize Natural product biosynthetic gene clusters' (CORASON) software, written in Perl and available open source (Fig. 4). Given a query gene inside a BGC of interest, the CORASON pipeline identifies other genomic loci that contain homologues of this gene and identifies the conserved core of these loci (Fig. 4a). Based on this core, a multi-locus approximate-maximum-likelihood phylogenetic tree[43] is constructed to reveal clades that may be responsible for the biosynthesis of different types of chemistry due to the association of specific types of additional enzyme-coding genes (Fig. 4b). This procedure can be performed for the 'core' enzyme-coding genes of a BGC, but also for e.g. tailoring genes, in order to reveal various GCFs likely to produce molecules with similar chemical modifications (Fig. 4c).

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

## An integrated workflow and interactive visualization

BiG-SCAPE and CORASON connect seamlessly with antiSMASH and MIBiG, as GenBank outputs of antiSMASH can be used directly as inputs for the workflow (Online Methods), and MIBiG reference data can be included in the analysis automatically. Although calculations on hundreds or thousands of genomes are too compute-intensive to provide them on a free public web server, results of each BiG-SCAPE run are still made available in an interactive HTML visualization that enables efficient exploration of biosynthetic diversity across large datasets for non-programmers. This can be viewed offline on any web browser or uploaded to the web to share results with other scientists. In a single view, the visualization displays BGC nodes colored by GCF in interactive sequence similarity networks, side-by-side with arrow visualizations of the gene clusters, which contain gene annotation and Pfam domain details that appear on mouse-over. Networks can be searched by the compound names of MIBiG reference clusters, Pfam domains of interest, or species names, with resulting matched nodes being instantly highlighted within the network. Each GCF has its own view panel, which shows the CORASON-based multi-locus phylogeny of the underlying BGCs and includes links to related families within the same gene cluster clan. Finally, an overview page is provided that displays statistics of the identified BGCs, as well as a GCF absence/presence heatmap of the most frequently occurring gene clusters within the input dataset.

To illustrate BiG-SCAPE/CORASON usage, we provide an example output of an analysis with antiSMASH-predicted BGCs from 103 complete *Streptomyces* genomes (see Online Methods), including as outgroups the genomes of *Catenulispora acidiphila* and *Salinispora arenicola* (interactive version of these results available at https://bigscape-corason.secondarymetabolites.org/streptomyces_example/). To connect the absence/presence map of GCFs across these genomes to species phylogeny, a high-resolution multilocus whole-genome phylogeny (Supplementary Fig. 4) was inferred from the *Streptomyces* conserved-core using CORASON, and the tree was decorated with the GCF absence/presence patterns (Supplementary Fig. 5). As has been observed before in other genera like *Salinispora* [30], this shows high conservation of some GCFs across a larger number of genomes (27 GCFs [~2%] occur across >10 genomes), combined with a large number of rare GCFs that are specific to one or a few genomes (1564 GCFs [92%] occur across three or fewer genomes).

## Case study: identification of novel detoxin analogues

To showcase the power of our workflow for the analysis of large data sets with high-resolution mapping of biosynthetic diversity, we focused on the detoxin and rimosamide GCFs[41] (Supplementary Figs. 6 and 7). Our analysis of the actinobacterial genomes revealed that these BGCs are taxonomically widespread and architecturally diverse. The conserved core (see Online Methods) of detoxin and rimosamide BGCs is composed of one NRPS, one NRPS/PKS hybrid, and one *tauD*-like gene. The rimosamide BGC differs from those of the detoxins by having an additional NRPS, which codes for an extension of the common detoxin/rimosamide core scaffold with isobutyrate and glycine[41].

The fact that the *tauD* gene was present across all members of this family but relatively unique within secondary metabolism caught our attention. The product of the *tauD* gene belongs to the Fe(II)/α-ketoglutarate-dependent hydroxylase enzyme superfamily and is named for the commonly encoded α-ketoglutarate-dependent taurine dioxygenase involved in the assimilation of sulphite by oxygenolytic release from the amino acid taurine[44]. Interestingly, this family also includes enzymes across fungi, bacteria and plants that catalyze hydroxylations, desaturations, ring expansions and ring formations, among other chemical transformations. To date, the role of *tauD* in detoxin and rimosamide biosynthesis is unknown, although it has been suggested to be responsible for the proline oxidation observed in some analogues[41].

An EvoMining[8] analysis of the TauD dioxygenase protein family showed specialized metabolism-related expansions of paralogues across genera such as *Streptomyces, Rhodococcus*, *Frankia* and *Amycolatopsis* (Supplementary Fig. 8). One expanded clade contained fifteen *tauD* homologues that belonged to experimentally characterized BGCs from MIBiG v1.3, as well as one within the rimosamide BGC (Supplementary Table 4).

Next, we investigated the genomic contexts of all *tauD* expansions (comprising 1175 BGCs) with the goal of identifying novel detoxin- and rimosamide-related BGCs. The BGCs were processed by CORASON using *tauD* as the query gene. Although ideally the detoxin/rimosamide BGC core would be defined as also containing the NRPS and NRPS-PKS hybrid genes, herein *tauD* was used as the sole member of the 'BGC core' to allow also

identifying fragmented BGCs. Gaps in the genome sequences were observed for some organisms including *Streptomyces humi* and *Amycolatopsis vancoresmycina* (Fig. 5). CORASON analysis revealed that the detoxin and rimosamide GCFs identified in BiG-SCAPE were part of a larger gene cluster clan related to peptide biosynthesis that also comprised unexplored clades across the phylum Actinobacteria (Fig. 5, Supplementary Fig. 9). Importantly, the high-resolution organization of BGC relationships enabled by the CORASON phylogeny revealed additional BGCs that were omitted by GCF clustering in BiG-SCAPE. This is because the fragmented nature of genome assemblies or the merger with adjacent BGCs by antiSMASH made these BGCs sufficiently different as to be classified into different GCFs under the cutoffs used, while CORASON was able to organize BGC relationships based on the single *tauD* gene (see Supplementary Figs. 3 and 10).

We hypothesized that the detoxins produced from BGCs in the unexplored clades would contain novel chemical variations related to the observed genetic variations. Fortunately, 40 of the 152 strains harbouring these BGCs were represented in our 363-strain LC-MS/MS metabolomics dataset. Molecular networking analysis of these data (Online Methods) indicated the presence of eight known detoxins, four known rimosamides and 99 putatively novel detoxin or rimosamide analogues (Fig. 3c), confirming the vast chemical diversity suggested in the BiG-SCAPE/CORASON data.

There were three detoxin BGC clades identified by BiG-SCAPE within the CORASON phylogenetic tree that captured our interest (Fig. 5). The first was named the ‘P450/enoyl clade’ because of the presence of putative cytochrome P450 and enoyl-CoA hydratase/isomerase genes in these BGCs (Fig. 5). Analysis of tandem MS data from extracts of *Streptomyces* sp. NRRL S-325, which has a BGC in this clade, and comparison to fragmentation patterns of known detoxins led to the discovery of detoxin $S_1$ (**1**; Fig. 5, Supplementary Figs. 11 and 12). This novel analogue contained a heptanamide side chain, a unique substructure among the detoxins and rimosamides that is likely installed by the condensation domain of the NRPS, potentially following processing by the predicted enoyl-CoA hydratase/isomerase and cytochrome P450s.

The second clade of interest, termed the ‘supercluster clade’, comprised BGCs with detoxin genes immediately adjacent to the known spectinomycin BGC[45] (Fig. 5). This was discovered because the spectinomycin MIBiG entry (BGC0000715) clustered with them on the CORASON tree due to it containing a *tauD* gene at its periphery (Fig. 5). Since the *tauD* gene is not known to be involved in spectinomycin biosynthesis, we hypothesized that there likely exists additional detoxin genes adjacent to this spectinomycin BGC in *S. spectabilis* NRRL 2792. We acquired this strain to determine if CORASON analysis could facilitate prediction of detoxin production based solely on the presence of a single query gene. Tandem mass spectrometry analysis of a *S. spectabilis* NRRL 2792 extract revealed production of five detoxin-like natural products (Fig. 5 and Supplementary Fig. 13), including detoxin $N_1$ (**2**), detoxin $N_2$ (**3**) and its acetoxylated analog, detoxin $N_3$ (**4**). Interestingly, ions with retention times and fragmentation patterns matching the latter two were also observed in extracts of *Streptomyces* sp. NRRL B-1347 from the supercluster clade, confirming the unique ability of CORASON to guide discovery by phylogenetically linking the limited NRRL-2792 sequence data to the detoxin supercluster clade. During

finalization of this manuscript, the genome of NRRL 2792 was published[46], and an abbreviated CORASON analysis confirmed presence of the detoxin BGC in a supercluster configuration with the spectinomycin BGC (Supplementary Fig. 14). LC-MS analysis of NRRL-2792 cultures supplemented with stable isotope-labeled amino acids corroborated structural predictions based on analysis of the closely related *Streptomyces* sp. NRRL B-1347 supercluster and tandem MS data (Fig. 5, Supplementary Figs. 15–20). All three new analogs were found to fully incorporate labeling from $^{13}C_6$-isoleucine, but $d_7$-proline was only fully incorporated into **3**. Loss of one deuteron from $d_7$-proline in **2** and **4** supported assignment of acetoxylation of the pyrrolidine ring common in reported detoxins and rimosamides[41,47]. Structural features unique to the N-series detoxins included the incorporation of an *N*-formylated tyrosine in **3** and **4** in place of the typical detoxin/rimosamide phenylalanine residue, which was supported by incorporation of ring-$d_4$-tyrosine. Compound **2** exhibited the unique incorporation of a tryptophan-derived residue at this position, made evident by retention of four deuterons when feeding indole-$d_5$-tryptophan (Supplementary Fig. 16). Although MS data was insufficient to deconvolute this substructure, compound **2** was produced by *S. spectabilis* NRRL 2792 in sufficient abundance for isolation and structure elucidation by NMR. Various 1D and 2D experiments confirmed assignments from MS data analysis and established an *N*-acetylated kynurenine as the tryptophan-derived substructure in **2** (Supplementary Fig.s 15–16, Supplementary Note 3a–h).

The third detoxin clade that we targeted comprised BGCs that were almost entirely within the genus *Amycolatopsis* (Fig. 5). This clade's BGCs also contained a unique predicted cytochrome P450 gene; hence, it was named the '*Amycolatopsis*/P450 clade' (Fig. 5). Although we did not have metabolomics data for strains with BGCs in the BiG-SCAPE-defined GCF, the CORASON visualization allowed the selection of an *Amycolatopsis* strain in our dataset with a very similar BGC (80-90% amino acid identity for the core genes) that contains a homologue of the desired P450 gene (Fig. 5, adjacent to the *Amycolatopsis*/P450 clade). Analysis of tandem MS data from an *Amycolatopsis jejuensis* NRRL B-24427 fermentation extract revealed detoxin isomers $P_1$ (**5**; Fig. 5, Supplementary Figs. 13 and 21) containing a tyrosine, $P_2$ (**6**; Fig. 5, Supplementary Figs. 13 and 22) featuring phenylalanine and a hydroxylated valine, as well as detoxin $P_3$, a closely related analog free of hydroxylation (**7**; Fig. 5, Supplementary Figs. 13 and 23). Only five of the seven novel detoxins described here appear as nodes in the Figure 3c molecular network, with the notable absence of two P-series analogs. This is because detoxin isomers $P_1$ and $P_2$ had a cosine similarity above 0.6 and were collapsed into one node, while detoxin $P_3$ was identified in fermentations following those that were a part of the original MS dataset. As before, validation of amino acid assignments observed in MS/MS fragmentation data for detoxins $P_1$–$P_3$ was achieved through several metabolic feeding experiments using stable isotopelabeled amino acids (Supplementary Figs. 24-27, 28-31, and 32-33). Detailed structural analysis for **1**–**7**, including results from feeding studies using stable isotope-labeled amino acids, deconvolution of tandem MS spectra, and full $^1H$, $^{13}C$, and 2D NMR assignments for **2**, are available in Supplementary Note 3, Supplementary Figures 15-33 and Online Methods. Previously reported detoxins and rimosamides antagonize blasticidin-S

inhibition of *Bacillus cereus*, a bioactivity that will be investigated for these analogs in follow-up studies[41,48].

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

## Discussion

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

Additionally, the ability to perform phylogenetic analyses of large sets of complete BGCs, as well as their individual genetic components, a long-standing challenge that has remained unsolved since first posed in 2008[49], will constitute a key technology to facilitate fundamental studies on the evolutionary origins of natural product chemical innovations. For example, phylogenies provide a stepping-stone to perform detailed analyses of how gene cluster architectures evolve from their constituent independent enzymes and sub-clusters. A logical next step will be the unified classification of the millions of BGCs within publicly available genome sequences, and a Pfam-like database for the assignment of biosynthetic gene cluster families to known and unknown areas of natural product chemical diversity.

## Online Methods

### Dataset

A set of 2,831 Actinobacterial genomes was downloaded from NCBI by querying for "Whole genome shotgun sequencing project" or "Complete genome" in combination with the taxonomic identifier for actinobacteria. The *Propionibacteriales*, *Micrococcales*, *Corynebacteriales* and *Bifidobacteriales* orders were excluded, as they contain large numbers of genomes without relevant natural product-producing capacity, except the *Nocardiaceae* family from the *Corynebacteriales* (see next section). To these set, 249 additional draft assemblies from the Metcalf lab were added (e.g. *Streptomyces* sp. B-1348. See BioProject

PRJNA488366). Draft genome assemblies from this BioProject were obtained by using SPAdes[50] with default options.

All files were processed with antiSMASH v4[17] (parameters: `--minimal`). The antiSMASH-annotated genome sequences are available as Online Data (antiSMASH_results_Metcalf_B, antiSMASH_results_Metcalf_J and antiSMASH_results_NCBI).

To the resulting 73,260 predicted Biosynthetic Gene Clusters (BGCs), 1,393 more were added from the Minimum Information about a Biosynthetic Gene Cluster database (MIBiG[21], release 1.3, August 2016, antiSMASH-analyzed versions from each entry) as reference data.

This final BGC set was then analyzed with BiG-SCAPE using version 31 of the Pfam database. The “hybrids” mode, which allows BGCs with mixed annotations be analyzed in their individual Class sets (e.g. a BGC annotated as *lantipeptide-t1pks* will be analyzed as both a RiPP and a PKSI) was enabled. Two results sets were created (Online Data: BiG-SCAPE Results network files): one with the default "global" mode enabled, and the other with "glocal" mode enabled (See Fig. 2).

### Actinobacteria Genomes Set

The extended set of genomes selected to be processed by antiSMASH and BiG-SCAPE was obtained by using the following query in the NCBI website on 2018-01-30 (2,891 results):

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

The CORASON and EvoMining results used the same unpublished draft genomes but a reduced set of 1,668 Actinobacterial genomes from an earlier query on the NCBI website, obtained on 2017-02-03 with the following query in the NCBI website (1,668 results):

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

## BiG-SCAPE algorithm

### Alignment Method Comparison

To compare alignment methods for domain sequences, the regular version of BiG-SCAPE was used against a custom-prepared version of the same snapshot using Muscle 3.8.1551-h6bb024c_4[51] (parameters: `-maxiters 2`) and MAFFT v7.407[52] (parameters: `--auto`); the three versions of the code are available at Online Data: Alignment Method Comparison. Muscle was parallelized using Python's `pool.map` on single-core instances for each domain sequence fasta file, while MAFFT was parallelized on each file with its `--thread` parameter. Comparison of final GCF calling (using BiG-SCAPE's `--mix` parameter) indicates high agreement between the three methods (Supplementary Figure 34), with hmmalign showing shorter runtimes as the number of BGCs in the input data increases (Supplementary Figure 35).

### Clustering algorithm optimization

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

Supplementary Figure 36 shows the dynamics and impact of the different filtering thresholds applied to the different BGC training networks, being a cutoff of 0.75 the value that maximized the number of nodes, while minimizing the impact on the structural integrity of the network. This analysis was performed using the igraph package[53] for the network analyses and ggplot2[54] for plotting.

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

**Input data—**The input of BiG-SCAPE are text files in GenBank format (.gbk extension) and the Pfam database[37] (already processed with hmmpress). While BiG-SCAPE is able to work with files not processed by antiSMASH, it relies on antiSMASH's product prediction to separate the BGCs in their correct biosynthetic class, thereby reducing computational

time. If the product annotation is unknown, missing or several different classes are mixed, the BGC will be then classified as “Other”.

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

**Distance calculation—**Pairwise distance calculation is divided between three values that measure a) the percentage of shared domain types (Jaccard Index), b) the similarity between aligned domain sequences (Domain Sequence Similarity index; domains from the same type are first matched for best similarity using the Munkres algorithm, as implemented in Scikit-Learn library[55]) and c) similarity of domain pair-types (Adjacency Index). For specific details of each index, see Supplementary Note 1.

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

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

To proceed to the next step, the LCS must be either three genes long, or contain at least one gene marked by antiSMASH as "Core Biosynthetic" (that is, genes that encode the first step in the assembly of the metabolite’s scaffold and that are used by antiSMASH as a first step in defining a biosynthetic cluster e.g. polyketide synthases or nonribosomal peptide synthetases).

In the extension stage, the selection of domains is extended for the BGC with the least number of genes up(down)stream (up until the end of the BGC or a contig break in the genome assembly). The remaining BGC domain selection (per side, i.e. both ‘left’ and ‘right’) will be subjected to expansion according to the following scoring algorithm in the Alignment Stage: for every gene in the reference BGC, a gene with the same domain organization is searched for in the remaining BGC. If such a gene is found, the score will be added a bonus (match=5) plus a penalty proportional to the distance from the current position (number of genes * gap penalty where gap=-2) and the current position will be moved to the position of the matching gene. If a gene with the same domain organization is

Europe PMC Funders Author Manuscripts

not found, the score will be decreased with a penalty (mismatch=-3). In the end, the highest-scoring extension is chosen to form the 'matching' BGC region on which the similarity will be calculated.

**GCF clustering—**Once the distance matrix is calculated for each BiG-SCAPE class (see Supplementary Table 1), Gene Cluster Family (GCF) assignment is performed for every cutoff distance selected by the user (the interactive visualization of BiG-SCAPE will show the one with the largest number) with 0.3 being the default. For every cutoff, BiG-SCAPE creates a network using all distances lower or equal than the current cutoff. The Affinity Propagation clustering algorithm[39] is applied to each subnetwork of connected components that emerge from this procedure. The similarity matrix for Affinity Propagation includes all distances between members of the subnetwork (i.e. it includes those with a distance greater than the current cutoff).

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

**Output—**BiG-SCAPE produces high-quality SVG figures for every BGC as well as text files from each of its constituent algorithms (hmmer domtable results, filtered domain results, aligned domain sequences, clustering results, and the distance network). As part of the output, BiG-SCAPE also offers an interactive HTML visualization where the user can navigate the distance network generated by the highest cutoff selected. BGCs connected and clustered into GCFs have a page on their own for closer inspection.

**CORASON family-mode—**As part of BiG-SCAPE's visual output, a CORASON-like tree is generated for every GCF page. This tree is created using the sequences of the Core Domains in the GCF. These are defined as the domain type(s) that a) appear with the highest frequency in the GCF and b) are detected in the central (or "exemplar") cluster, defined by the Affinity Propagation cluster. All copies of the Core Domains in the exemplar are concatenated, as well as those from the best matching domains of the rest of the BGCs in the GCF (aligned domain sequences are used). The tree is constructed using FastTree[43] (default options). Visual alignment is attempted using the position of the Longest Common Information from the distance calculation step (between the exemplar BGCs and each of the other clusters).

**Availability—**BiG-SCAPE written in Python and is currently compatible with both Python 2 and Python 3. It is freely available at https://git.wageningenur.nl/medema-group/BiG-SCAPE. More extensive details of the algorithm are available at the repository's wiki: https://git.wageningenur.nl/medema-group/BiG-SCAPE/wikis/home.

**Weight optimization methods—**Tuning of weights for each BiG-SCAPE class was calculated by a brute-force approach, by choosing the weight combination that maximized the correlation between BGC and Compound distances for every pair of BGCs in the same class in a manually curated Compound Group table (Supplementary Dataset). The dataset comprised all BGCs from the MIBiG database (v1.3) that had linked compound SMILES and had at least two predicted domains to filter out minimal gene cluster entries. BGC distances were calculated by moving in steps of 0.01 between the Jaccard, Domain Sequence Similarity, the original Goodman-Kruskal[56], and Adjacency indices, such that JI+DSS+GK +AI=1. The anchorboost parameter of DSS was allowed to change in the range [1,4] with steps of 0.5. For the DSS index, only the original 4 anchor domains were considered (Condensation Domain, PF00668; Beta-ketoacyl synthase N-terminal, PF00109; Beta-ketoacyl synthase C-terminal domain, PF02801, and the Terpene synthase N-terminal, PF01397). Compound distances were calculated only once, between all BGCs in the MIBiG 1.3 that had an annotated SMILES string representing the molecule. Their pairwise distance was calculated by using RDKit (Tanimoto coefficient based on Morgan fingerprinting, radius=4). The nine original curated Compound classes were used to tune the weights of 7 BiG-SCAPE classes (the Terpene BiG-SCAPE class was initially included in the Others Compound class due to a low number of points and was assigned default values of Jw = 0.2, DDSw = 0.75, AIw = 0.05).

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

**Comparison with other methods—**To compare BiG-SCAPE to the gene cluster family algorithm in Doroghazi et al. 2014[4], we reconstructed 11,618 GenBank files from data related to that study (allClusterProts.fasta file from https://www.igb.illinois.edu/labs/metcalf/gcf/search.html). We analyzed these reconstructed cluster files with antiSMASH v4, and used its output (Online Data: Comparison Doroghazi2014 reconstructed BGCs antiSMASH results) to make a run in BiG-SCAPE (Online Data: Comparison Doroghazi2014 BiG-SCAPE results).

Unlike the Doroghazi method, BiG-SCAPE follows a two-step process to infer GCFs. First, it filters the resulting network using a predefined empirical cutoff distance of 0.3, and later the GCFs are identified by the Affinity Propagation clustering algorithm. This two-step approach partitions the natural emerging components from the filtering step, increasing the resolution of the inferred GCFs. To be able to provide a fair comparison with Doroghazi et al. 2014 inferred GCFs, we used the natural emerged components after the filtering steps and compared the different clustering results and found good agreements between both methods (see Supplementary Figure 42 for details), while BiG-SCAPE took only a fraction of the runtime of the previously published tool (Supplementary Table 6).

## CORASON algorithm

CORASON inputs are a custom genomic database, a reference cluster and a query gene located within the reference cluster. The genomic database is a collection of either genomes or BGCs in GenBank format. CORASON will identify the conserved core of the reference BGC within the genomic database.

Best bidirectional hits (BBH) are pairs of genes that exist in two different sets of genes (genomes, metagenomes or BGCs) that are more similar between each other than to any other sequence in the set pair. In CORASON, this relationship was generalized in a stricter algorithm that considers all-vs-all comparisons between every set in the collection to remove paralogues and conserve only true orthologues. As a result, the conserved core is composed of gene families that are each guaranteed to be BBH across the whole collection (while they need not be contiguous).

The BGC conserved core facilitates reconstructing the BGC evolutionary history in a multilocus tree. The query gene assures that at least one element will be present in the conserved core and will also be used to visually align the BGCs variations in the graphical output.

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

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

Within a set of $N$ BGCs variations, if the set $H$ of homologous genes is defined as follows:

$$H = \{h_i | h_i \in BGC_i \ \forall \ i \ \in \{1, 2, \ldots N\}\}$$

then, $H$ belongs to the conserved core if and only if

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

**Phylogenetic reconstruction and gene cluster alignment—**For each BGC, its conserved core sequences are concatenated and then aligned using Muscle version 3.8.31[51]. The alignments are curated using Gblocks[57] with a minimum block length of 5 positions, a maximum of 10 contiguous non-conserved positions and considering only positions with a gap in less than 50% of the sequences in the final alignment. If the curation turns out to be empty, then the noncurated alignment will be used for the tree. If the alignment itself is empty, it is recommended to the reduce the score cutoff or the scope of the taxonomic diversity on the genomic database. Without the alignment, BGCs will be drawn but not sorted. Approximately-maximum-likelihood phylogenetic trees are inferred using FastTree[43] version 2.1.10 from the curated amino acid alignment.

**BGC prioritization graphical output—**CORASON produces a Scalable Vector Graphics (SVG) file containing the BGC variations sorted as stated by the phylogenetic reconstruction and aligned according to the query enzyme. The newick tree is converted to SVG applying Newick Utilities version 1.6[58] and each BGC is drawn with the Perl module SVG. As an additional feature to facilitate even more visual differentiation of BGC families within BGC clans, genes on each cluster are visually represented with a color gradient according to the sequence similarity to their homologous gene on the reference cluster. Other CORASON outputs include the Newick tree, the GenBank files of the BGC variations and the conserved core report.

CORASON was developed in Perl:5.20 and is available as free software on GitHub (https://github.com/nselem/corason) and as a downloadable image on dockerhub (https://hub.docker.com/r/nselem/corason/). A CORASON tutorial is available on-line at https://github.com/nselem/corason/wiki.

## *Streptomyces* closed genomes analysis

Sequences from 103 complete *Streptomyces* genomes were retrieved from NCBI by querying for "*Streptomyces*" and "complete genome" not “segment”. Two genomes corresponding to *C. acidiphila* and *S. arenicola* (CP001700 and CP000850) as outgroups. These genomes were analyzed by antiSMASH v4 and the resulting gene cluster (Online Data: Closed Streptomyces antiSMASH results) files were used as input for BiG-SCAPE (Online Data: Closed Streptomyces BiG-SCAPE results). The conserved core was extracted and curated with the CORASON algorithm. The tree was constructed using FastTree with default values over a matrix of 114,051 amino acids in size, from 446 conserved gene families (Online Data: StreptomycesCore).

The interactive report of BiG-SCAPE reports only 96 genomes, because genomic scaffolds that belong to the same genome are grouped by ORGANISM identifier, and the following strains have more than one assembly project associated in NCBI with the same ORGANISM identifier:

*Streptomyces clavuligerus* ATCC 27064: CM000913, CM001015

*Streptomyces cattleya* NRRL 8057 = DSM 46488”: CP003219, FQ859185

*Streptomyces lydicus*: CP007699, CP017157, CP019457

*Streptomyces venezuelae*: CP013129, LN881739

*Streptomyces albus*: CP014485, CP016825, CP010519

*Streptomyces pactum*: CP016795, CP019724

*Streptomyces pluripotens*: CP021080, CP022433

## Phylogenomic analysis

For the TauD expansions tree (Supplementary Figure 8), a *tauD* sequence from *Escherichia coli K12* was used as query to conduct a blast search against the reduced genomic database of 1917 Actinobacteria genomes (e-value .001), followed by an EvoMining analysis and a search for recruitments on MIBiG database (e-value 0.001). Recovered *tauD* orthologues were aligned with Muscle 3.8.31[51] and alignments were curated using Gblocks[57] in the same manner as described above. An unrooted approximately-maximum-likelihood tree was built using FastTree[43]. The tree was colored using Newick Utilities[58] according to BiG-SCAPE families.

The CORASON tree has as query gene *tauD* from the reference cluster of the organism *Streptomyces NRRL* B-1347 (Accession JOJM01). CORASON trees are unrooted, but this tree was posteriorly rooted with the BGC from the genome *Streptomyces* sp. NC1, because this BGC is different from all other clusters in the dimeric peptide clan, as it does not share the core but only the accessory enzyme-coding genes with other BGC clan members.

## Molecular networking methods

**Cultivation of actinomycetes for MS-based metabolomics—**All strains analyzed for metabolomics were grown on four media types: arginine/glycerol/salts, mannitol/soyflour, ISP medium 4, or glycerol/sucrose/beef extract/casamino acids as previously reported[4]. After 10 days of growth, plates were frozen, then thawed and pressed to release spent liquid media. Media was then filtered and extracted using 30 mg Supel-Select HLB SPE cartridges (Supelco) and resuspended to a concentration of approximately 2 mg/mL in 5% acetonitrile prior to LC-MS analysis.

**Acquisition and analysis of LC-MS metabolomics data—**All LC-MS/MS analyses were performed using an Agilent 1150 HPLC coupled with a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Reversed phase chromatography was performed at a 200 μL/min flow rate on a Phenomenex Kinetex C18 RP-HPLC column (150 mm x 2.1 mm i.d., 2 μm particle size, 100 Å pore size). Mobile phase A was water with 0.1% formic acid and mobile phase B was acetonitrile with 0.1% formic acid. Mass spectral data for both MS and MS/MS were acquired using a 250–3750 *m/z* scan range, a resolution of 35,000, a maximum inject time of 40 ms, and an AGC target value of $1 \times 10^6$. The top five most intense ions in each full MS spectrum were targeted for fragmentation by Higher-energy Collisional Dissociation (HCD) at 25 eV. Tandem MS data was analyzed using spectral networking as previously described[59]. Signals detected in multiple strains were determined to be the same ion if the

observed accurate masses were within 4 ppm and fragmentation cosine similarity scores were above 0.75, yielding 5,824 ions detected in two or more strains.

**LC-MS molecular networking—**Molecular networking was performed as previously reported[40,59]. Briefly, individual MS/MS scans were extracted from each mass spectrometry raw file and filtered to remove the 25% of the ions with the lowest intensity. Each MS/MS scan was further processed by taking the square root of each ion's intensity and normalizing it so that the sum of all intensities in each MS/MS scan was equal to one. Cosine similarities were calculated between all MS/MS scans, with scores ranging between 0 and 1, with a score of 1 indicating that two MS/MS scans were identical. Precursor ions were determined to be identical if they were within 0.01 m/z and their corresponding MS/MS spectra had a cosine similarity score of >0.6. A visualization of the network was constructed in Cytoscape by drawing edges between scan nodes whose cosine similarity was >0.6. The network was manually analyzed to identify ions related to known detoxins and rimosamides, which were found to cluster together as one molecular family with 99 putatively novel analogs, a subset of which were characterized herein as detoxins $S_1$, $N_1$–$N_3$, and $P_1$–$P_3$.

## Metabolic labeling of detoxins $N_1$–$N_3$, and $P_1$–$P_3$ with stable isotope-labeled amino acids

*Streptomyces spectabilis* Dietz NRRL 2792 (ATCC 27741) was obtained from the American Type Culture Collection (ATCC) and was grown on 60 mm solid agar media Petri plates containing arginine/glycerol/salts medium (1 L of DI water, 15 g agar, 1 g arginine, 12.5 g glycerol, 1 g potassium phosphate dibasic, 1 g sodium chloride, 0.5 g magnesium sulfate heptahydrate, 10 mg iron (II) sulfate hexahydrate, 1 mg copper(II) sulfate pentahydrate, 1 mg manganese(II) sulfate monohydrate, and 1 mg zinc sulfate heptahydrate). *Amycolatopsis jejuensis* NRRL B-24427 was obtained from the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA) and was grown on solid agar media Petri plates containing mannitol/soyflour medium (1 L of DI water, 15 g agar, 20 g D-mannitol, 20 g soy flour). For all metabolic labeling experiments, media was supplemented with 1 mL of a 10 mM solution of each stable isotope labeled amino acid. Stable isotope labeled amino acids used were $^{13}C_6$-isoleucine, $d_8$,$^{15}$N-phenylalanine, $d_7$-proline, 2,5,5-$d_3$-proline, phenyl-$d_4$-tyrosine, $d_8$-valine, 2-$d_1$-valine, and 3-$d_1$-valine. After five days incubation in the presence of stable isotope labeled amino acids, plates were frozen overnight at -20 °C, thawed, and pressed to release spent liquid media. Extracellular secondary metabolites were extracted using 30 mg Supel-Select HLB SPE cartridges (Supelco) and eluted with 90% acetonitrile. Samples were dried, resuspended in 5% acetonitrile, and analyzed by reversed-phase LC-MS/MS on a Q Exactive mass spectrometer as described above. The Methods used for LC-MS data acquisition on the Q Exactive were the same except for occasional parameter adjustments made to target major unnatural isotope ions for optimal fragmentation.

## Acquisition of NMR data

All NMR experiments were performed in $D_2O$. $^1$H, $^{13}$C, COSY, HSQC, HMBC, and NOESY spectra were obtained on a Bruker NEO spectrometer (600MHz for $^1$H, 150 MHz for $^{13}$C) with a QCI-F cryoprobe. The $^1$H-$^1$H TOCSY spectrum was obtained on a Bruker Avance III 500 MHz spectrometer (500 MHz for $^1$H) equipped with a DCH CryoProbe.

Chemical shifts (δ) are given in ppm and coupling constants (*J*) are reported in Hz. $^{1}H$ and $^{13}C$ chemical shifts were referenced to sodium formate ($\delta_H$ 8.44; $\delta_C$ 171.67). $^{1}H$ and $^{13}C$ NMR resonances of **2** are reported in Supplementary Note 3i.

### Metabologenomic correlations

Strains with metabolomics data were referenced against the BiG-SCAPE GCF absence/presence matrices. GCFs that had representative gene clusters in two or more strains were considered correlatable and entered into the correlations dataset. The different BiG-SCAPE modes and cutoffs produced variable numbers of correlatable GCFs and thus different numbers of ion-GCF hypotheses (Supplementary Table 2). Supplementary Figure 43 shows the full version of Figure 3d.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

We express our gratitude to the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA) for providing bacterial strains; Hyoung Sook Ann, Zachary Crispino, Yewool Kim, Natalia Ciszek, and Kyle Espejo for generating bacterial culture extracts; Ryan McClure, Matthew Robey, and Galen Miley for assistance with and contributions to metabolomic data collection methods and acquisition; and Dr. Yongbo Zhang and Dr. Yuyang Wu of the IMSERC at Northwestern University for assistance in acquiring NMR data. Some analyses were carried out using CONABIO's computing cluster, with funds from SEMARNAT. We thank Kai Blin for technical assistance with setting up the website on the secondarymetabolites.org domain.

The research reported in this publication was supported by the Netherlands Organization for Scientific Research (NWO) [VENI Grant 863.15.002 to MHM], the Graduate School for Experimental Plant Sciences (EPS grant to MHM); National Institutes of Health Genome to Natural Products Network supplementary award [U01GM110706 to MHM], CONACyT grants [CBS2017_285746 and 2017_051TAMU to FBG; postdoctoral scholarship 263661 to JCNM; PhD scholarship 204482 to NSM, (who was also supported by the Innovation Secretary of Guanajuato)], the National Cancer Institute (NCI) of the National Institutes of Health (NIH) under Award Number F32CA221327 (MWM), the National Institute of General Medical Sciences (NIGMS) under Award Number F32GM120999 (EIP), the São Paulo Research Foundation (FAPESP) under Grant Number 17/08038-8 (LTDC), the National Center for Complementary and Integrative Health (NCCIH) of the NIH under Award Number R01AT009143 (RJT, NLK) and Warwick Integrative Synthetic Biology Centre, a UK Synthetic Biology Research grant from the BBSRC and EPSRC [BB/M017982/1 to ELCdlS]. This work made use of the IMSERC at Northwestern University, which has received support from the NIH (1S10OD012016-01/1S10RR019071-01A1), the State of Illinois, and International Institute for Nanotechnology (IIN).

AF-G received funding from the European Union's Horizon 2020 research and innovation program [Blue Growth: Unlocking the potential of Seas and Oceans] under grant agreement no. [634486] (project acronym INMARE).

## Data availability

Genomes used in this study include assemblies from the sequencing project deposited in NCBI BioProject PRJNA488366, in Sequence Read Archive (SRA) runs with accession numbers SRX4638772-SRX4639021. antiSMASH, BiG-SCAPE and CORASON results for all genome assemblies, along with raw files of phylogenetic trees are available as Online Data at DOI 10.5281/zenodo.1532752. Fully annotated nucleotide sequences for the BGCs for Detoxin $S_1$, Detoxin $N_2$-$N_3$ and detoxins $P_1$–$P_3$ have been deposited in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under accession numbers BK010707, BK010852 and BK010851, respectively, and in MIBiG under accessions BGC0001840, BGC0001878 and BGC0001841, respectively. All raw mass spectrometry

Europe PMC Funders Author Manuscripts

data files for strains producing one or more of the nine compounds used for correlation analysis have been submitted to MassIVE under accession number MSV000083738. Raw mass spectrometry data files and isolated MS/MS scan files for all newly identified detoxin analogs have been uploaded to MassIVE with accession number MSV000083648, and MS/MS data for other strains is available upon request.

### Code availability

All our software is open source. An overview of both BiG-SCAPE and CORASON can be found at

https://bigscape-corason.secondarymetabolites.org/;

BiG-SCAPE project: https://git.wur.nl/medema-group/BiG-SCAPE;

CORASON project: https://github.com/nselem/corason;

## References


1. Traxler MF, Kolter R. Natural products in soil microbe interactions and evolution. Nat Prod Rep. 2015; 32(7):956–970. DOI: 10.1039/c5np00013k [PubMed: 26000872]
2. Davies J. Specialized microbial metabolites: functions and origins. J Antibiot (Tokyo). 2013; 66(7):361–364. DOI: 10.1038/ja.2013.61 [PubMed: 23756686]
3. Cimermancic P, Medema MH, Claesen J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014; 158(2):412–421. DOI: 10.1016/j.cell.2014.06.034 [PubMed: 25036635]
4. Doroghazi JR, Albright JC, Goering AW, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014; 10(11):963–968. DOI: 10.1038/nchembio.1659 [PubMed: 25262415]
5. Dejong CA, Chen GM, Li H, et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol. 2016; 12(12):1007–1014. DOI: 10.1038/nchembio.2188 [PubMed: 27694801]
6. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics. 2017; 33(20):3202–3210. DOI: 10.1093/bioinformatics/btx400 [PubMed: 28633438]
7. 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. 2017; 114(22):5601–5606. DOI: 10.1073/pnas.1614680114 [PubMed: 28461474]
8. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biol Evol. 2016; 8(6):1906–1916. DOI: 10.1093/gbe/evw125 [PubMed: 27289100]
9. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nat Chem Biol. 2015; 11(9):639–648. DOI: 10.1038/nchembio.1884 [PubMed: 26284671]
10. Katz L, Baltz RH. Natural product discovery: past, present, and future. J Ind Microbiol Biotechnol. 2016; 43(2–3):155–176. DOI: 10.1007/s10295-015-1723-5 [PubMed: 26739136]
11. Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). Nature. 2002; 417(6885):141–147. [PubMed: 12000953]
12. Schneiker S, Perlova O, Kaiser O, et al. Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol. 2007; 25(11):1281–1289. [PubMed: 17965706]

13. Bergmann S, Schumann J, Scherlach K, Lange C, Brakhage AA, Hertweck C. Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. Nat Chem Biol. 2007; 3(4):213–217. DOI: 10.1038/nchembio869 [PubMed: 17369821]
14. Medema MH, Blin K, Cimermancic P, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011; 39(W1):W339–W346. DOI: 10.1093/nar/gkr466 [PubMed: 21672958]
15. Blin K, Medema MH, Kazempour D, et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013; 41(Web Server issue):204–212. DOI: 10.1093/nar/gkt449
16. Weber T, Blin K, Duddela S, et al. antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. May; 2015 43(W1):1–7. DOI: 10.1093/nar/gkv437 [PubMed: 25505162]
17. Blin K, Wolf T, Chevrette MG, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017; 45(W1):W36–W41. DOI: 10.1093/nar/gkx319 [PubMed: 28460038]
18. Johnston CW, Skinnider MA, Wyatt MA, et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. Nat Commun. 2015; 6:8421.doi: 10.1038/ncomms9421 [PubMed: 26412281]
19. Skinnider MA, Dejong CA, Rees PN, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 2015; 43(20):9645–9662. DOI: 10.1093/nar/gkv1012 [PubMed: 26442528]
20. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 2017; 45(W1):W49–W54. DOI: 10.1093/nar/gkx320 [PubMed: 28460067]
21. Medema MH, Kottmann R, Yilmaz P, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015; 11(9):625–631. DOI: 10.1038/nchembio.1890 [PubMed: 26284661]
22. Nielsen JC, Grijseels S, Prigent S, et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in Penicillium species. Nat Microbiol. 2017; 2:17044.doi: 10.1038/nmicrobiol.2017.44 [PubMed: 28368369]
23. Tobias NJ, Wolff H, Djahanschiri B, et al. Natural product diversity associated with the nematode symbionts Photorhabdus and Xenorhabdus. Nat Microbiol. 2017; 2(12):1676–1685. DOI: 10.1038/s41564-017-0039-9 [PubMed: 28993611]
24. Grubbs KJ, Bleich RM, Santa Maria KC, et al. Greene CS. Large-Scale Bioinformatics Analysis of *Bacillus* Genomes Uncovers Conserved Roles of Natural Products in Bacterial Physiology. mSystems. 2017; 2(6):e00040–17. DOI: 10.1128/mSystems.00040-17 [PubMed: 29152584]
25. Freeman MF, Gurgui C, Helf MJ, et al. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science. 2012; 338(6105):387–390. DOI: 10.1126/science.1226121 [PubMed: 22983711]
26. Agarwal V, Blanton JM, Podell S, et al. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. Nat Chem Biol. 2017; 13(5):537–543. DOI: 10.1038/nchembio.2330 [PubMed: 28319100]
27. Owen JG, Charlop-Powers Z, Smith AG, et al. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. Proc Natl Acad Sci U S A. 2015; 112(14):4221–4226. DOI: 10.1073/pnas.1501124112 [PubMed: 25831524]
28. Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017; 2(11):1533–1542. DOI: 10.1038/s41564-017-0012-7 [PubMed: 28894102]
29. Leao T, Castelão G, Korobeynikov A, et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea. Proc Natl Acad Sci. 2017; 114(12):3198–3203. DOI: 10.1073/pnas.1618556114 [PubMed: 28265051]
30. Ziemert N, Lechner A, Wietz M, Millán-Aguiñaga N, Chavarria KL, Jensen PR. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc Natl Acad Sci U S A. 2014; 111(12):E1130–9. DOI: 10.1073/pnas.1324161111 [PubMed: 24616526]

31. Medema MH, Paalvast Y, Nguyen DD, et al. Gardner PP. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. PLoS Comput Biol. 2014; 10(9):e1003822.doi: 10.1371/journal.pcbi.1003822 [PubMed: 25188327]
32. Mohimani H, Liu W-T, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA. NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. J Nat Prod. 2014; 77(8):1902–1909. DOI: 10.1021/np500370c [PubMed: 25116163]
33. Mohimani H, Kersten RD, Liu W-T, et al. Automated Genome Mining of Ribosomal Peptide Natural Products. ACS Chem Biol. 2014; 9(7):1545–1551. DOI: 10.1021/cb500199h [PubMed: 24802639]
34. Nguyen DD, Wu CH, Moree WJ, et al. MS/MS networking guided analysis of molecule and gene cluster families. ProcNatlAcadSciUSA. 2013; 110:E2611–E2620.
35. Goering AW, McClure RA, Doroghazi JR, et al. Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. ACS Cent Sci. 2016; 2(2):99–108. [PubMed: 27163034]
36. Duncan KR, Crüsemann M, Lechner A, et al. Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. Chem Biol. 2015; 22(4):460–471. DOI: 10.1016/J.CHEMBIOL.2015.03.010 [PubMed: 25865308]
37. Punta M, Coggill P, Eberhardt R, et al. The Pfam protein families databases. Nucleic Acids Res 40 D290-D301. 2012; 30(1):1–12. DOI: 10.1093/nar/gkp985
38. Medema MHMH, Cimermancic P, Sali A, Takano E, Fischbach MAMA. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput Biol. 2014; 10(12):e1004016.doi: 10.1371/journal.pcbi.1004016 [PubMed: 25474254]
39. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007; 315(5814):972–976. DOI: 10.1126/science.1136800 [PubMed: 17218491]
40. Parkinson EI, Tryon JH, Goering AW, et al. Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. ACS Chem Biol. 2018 13(4):1029.Mar.doi: 10.1021/acschembio.7b01089 [PubMed: 29510029]
41. McClure RA, Goering AW, Ju K-S, et al. Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using Metabolite/Gene Cluster Correlations. ACS Chem Biol. 2016; 11(12):3452–3460. DOI: 10.1021/acschembio.6b00779 [PubMed: 27809474]
42. Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A. 2012; 109(26):E1743–52. DOI: 10.1073/pnas.1203689109 [PubMed: 22586093]
43. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26(7):1641–1650. [PubMed: 19377059]
44. Hausinger RP. Fe(II)/α-Ketoglutarate-Dependent Hydroxylases and Related Enzymes. Crit Rev Biochem Mol Biol. 2004; 39(1):21–68. DOI: 10.1080/10409230490440541 [PubMed: 15121720]
45. Kim K-R, Kim T-J, Suh J-W. The Gene Cluster for Spectinomycin Biosynthesis and the Aminoglycoside-Resistance Function of spcM in Streptomycesspectabilis. Curr Microbiol. 2008; 57(4):371. [PubMed: 18663525]
46. Sinha A, Phillips-Salemka S, Niraula T-A, Short KA, Niraula NP. The complete genomic sequence of Streptomyces spectabilis NRRL-2792 and identification of secondary metabolite biosynthetic gene clusters. J Ind Microbiol Biotechnol. 2019; 46(8):1217–1223. DOI: 10.1007/s10295-019-02172-8 [PubMed: 31197515]
47. Ogita T, Seto H, Otake N, Yonehara H. The Structures of Minor Congeners of the Detoxin Complex. Agric Biol Chem. 1981; 45(11):2605–2611.
48. Yonehara H, Seto H, Aizawa S, Hidaka T, Shimazu A, Otake N. The detoxin complex, selective antagonists of blasticidin S. J Antibiot (Tokyo). 1968; 21(5):369–370. DOI: 10.7164/antibiotics.21.369 [PubMed: 4973124]
49. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci U S A. 2008; 105(12):4601–4608. DOI: 10.1073/pnas.0709132105 [PubMed: 18216259]

50. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012; 19(5):455–477. DOI: 10.1089/cmb.2012.0021 [PubMed: 22506599]
51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(1362-4962; 0305-1048; 5):1792–1797. [PubMed: 15034147]
52. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–780. DOI: 10.1093/molbev/mst010 [PubMed: 23329690]
53. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Sy. 2006; 1695
54. Wickham H, Chang W. ggplot2: An implementation of the Grammar of Graphics. R Packag version 07. 2008
55. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. Oct.2011 12:2825–2830.
56. Lin K, Zhu L, Zhang DY. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics. 2006; 22(17):2081–2086. DOI: 10.1093/bioinformatics/btl366 [PubMed: 16837531]
57. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17(4):540–552. [PubMed: 10742046]
58. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010; 26(13):1669–1670. [PubMed: 20472542]
59. Henke MT, Soukup AA, Goering AW, et al. New aspercryptins, lipopeptide natural products, revealed by HDAC inhibition in Aspergillus nidulans. ACS Chem Biol. 2016; 11(8):2117–2123. [PubMed: 27310134]


Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

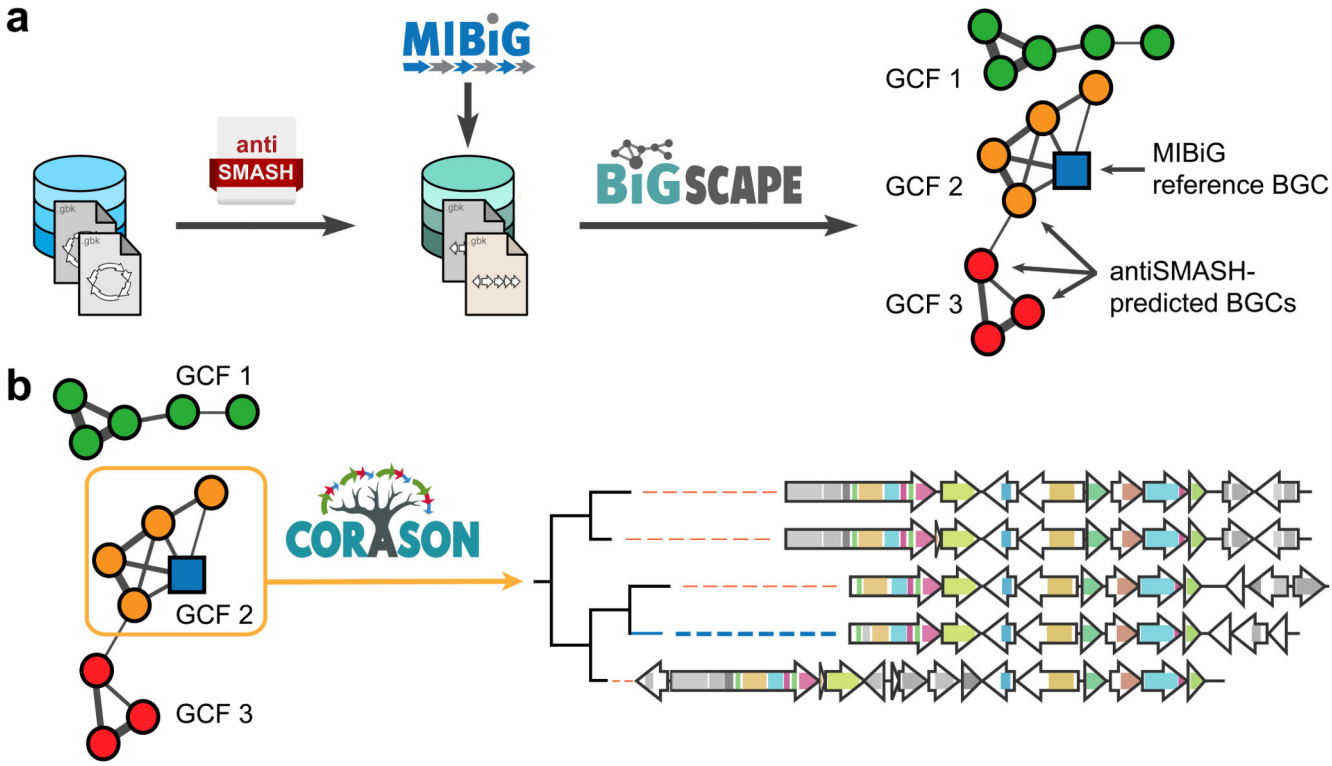


**Fig. 1. The BiG-SCAPE/CORASON workflow.**

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

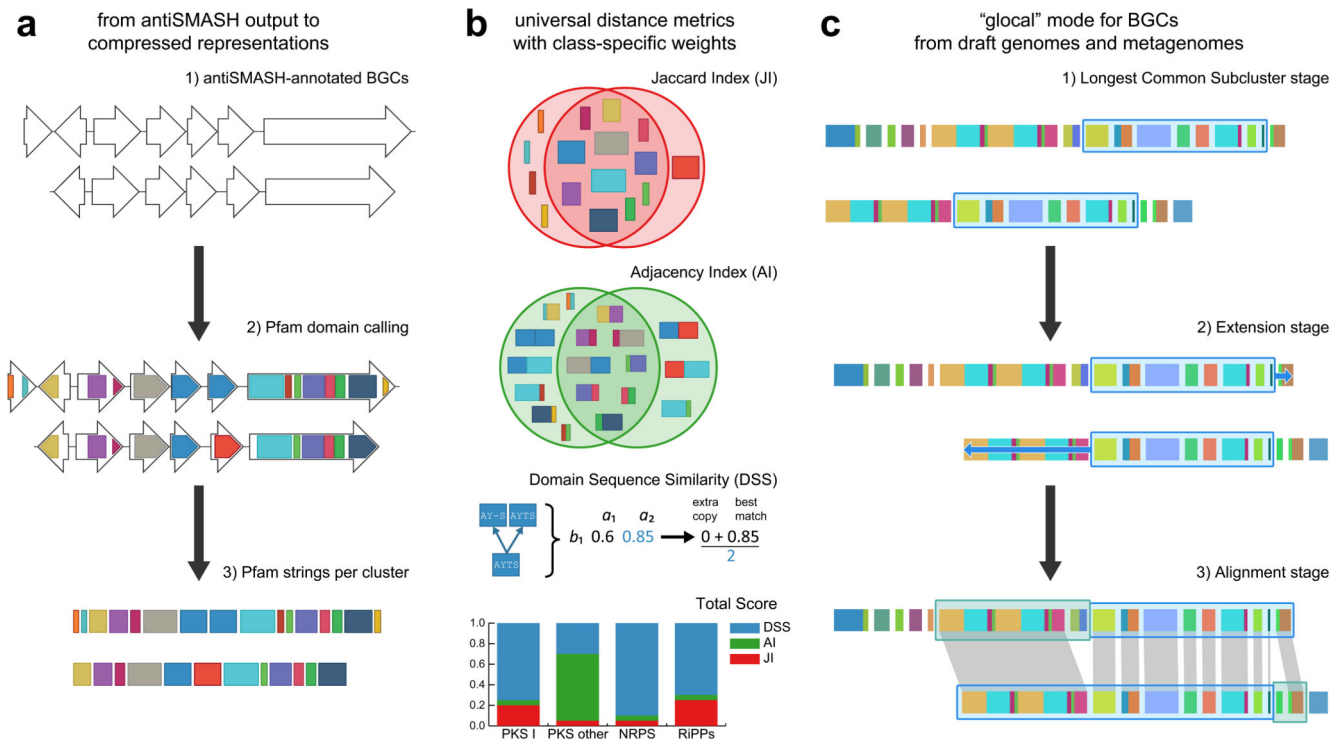


**Fig. 2. Main concepts in the BiG-SCAPE algorithm.**

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




Europe PMC Funders Author Manuscripts

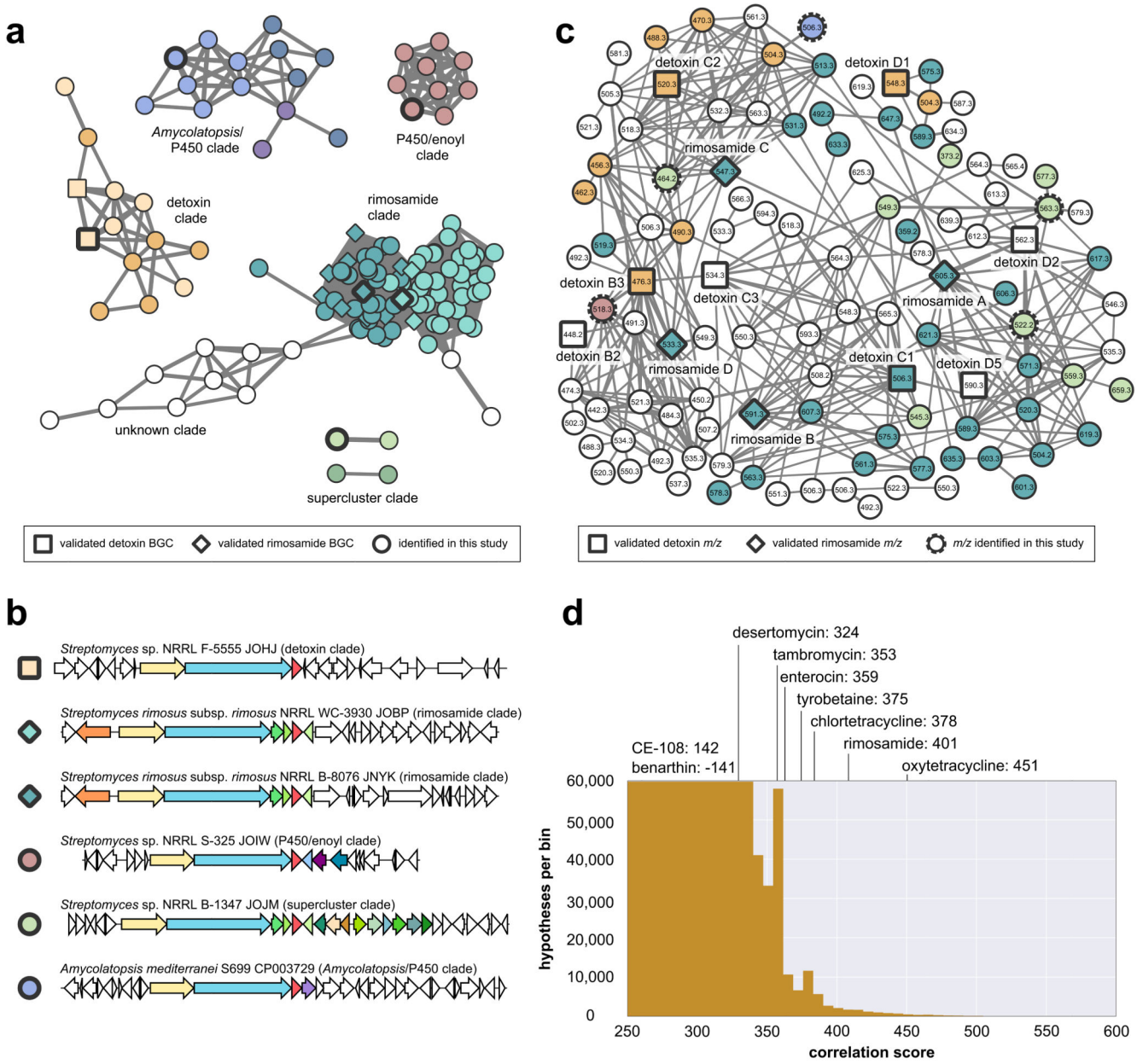


**Fig. 3. Sequence similarity and molecular networks.**

**a,** Detail of a BiG-SCAPE network containing validated detoxin and rimosamide BGCs, filtered for the presence of the taurine dioxygenase (TauD) domain. BiG-SCAPE gene cluster family classifications include the rimosamide (turquoise shades) and detoxin (orange shades) families, as well as the '*Amycolatopsis*/P450' (violet shades), 'P450/enoyl' (pink), and 'supercluster' (light green shades) families explored in this study. **b**, Validated BGCs represented by bold-outlined nodes. **c,** The detoxin and rimosamide molecular family based on tandem MS data of a 363-strain actinomycete library is colored by BiG-SCAPE family. Known detoxin (squares) and rimosamide (diamonds) nodes have solid bold outlines while putative detoxins are circular nodes and novel analogs from this study are indicated by bold,




Europe PMC Funders Author Manuscripts

dotted outlines. **d,** Histogram of all ion-GCF correlation scores resulting from the metabologenomics round run with 0.30 glocal distance cutoff. Known ion-GCF pair correlation scores are overlaid; 6 out of 9 appear in the ‘tail’ of the distribution, which would be indicative of a true connection. The low scoring for benarthin is due to the complicated fragmentation pattern of its BGCs (Supplementary Figure 3).



Europe PMC Funders Author Manuscripts

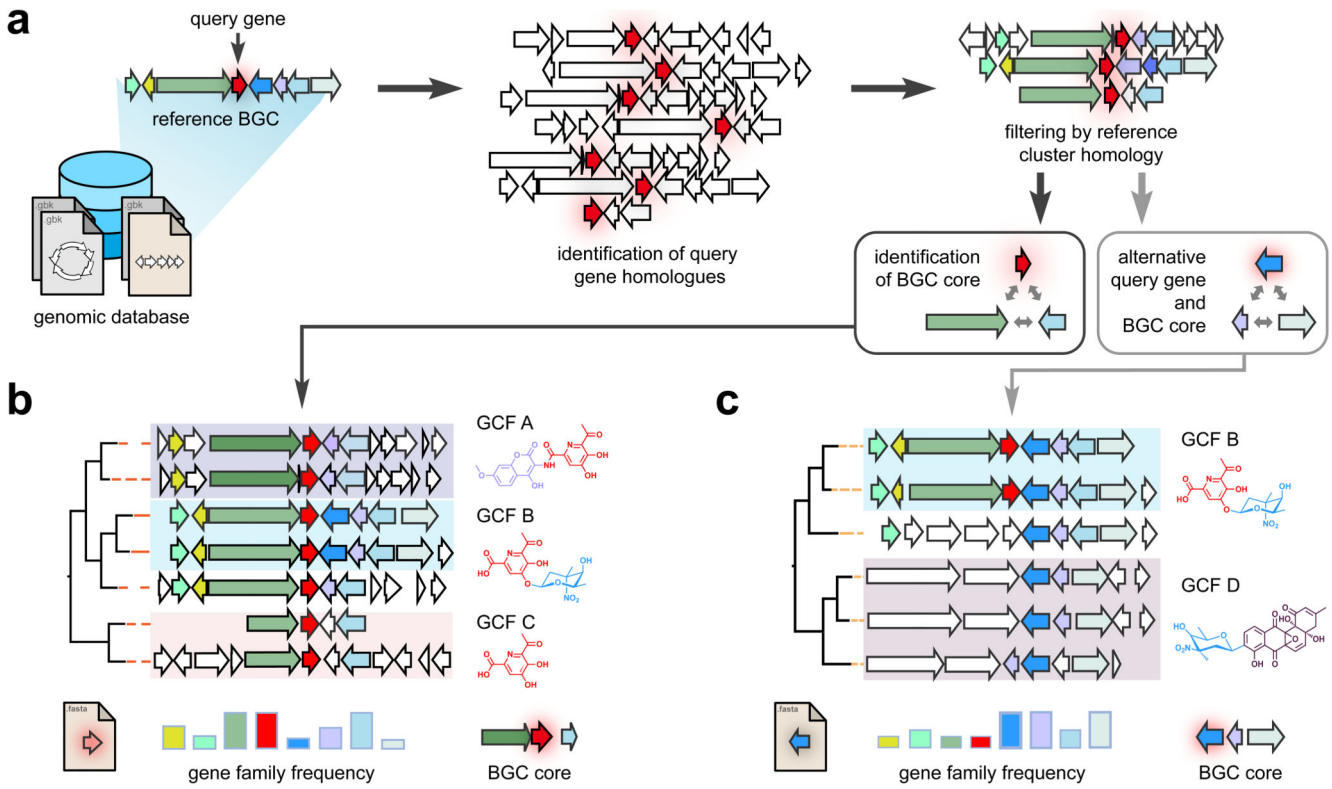


**Fig. 4. CORASON Workflow.**

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




Europe PMC Funders Author Manuscripts

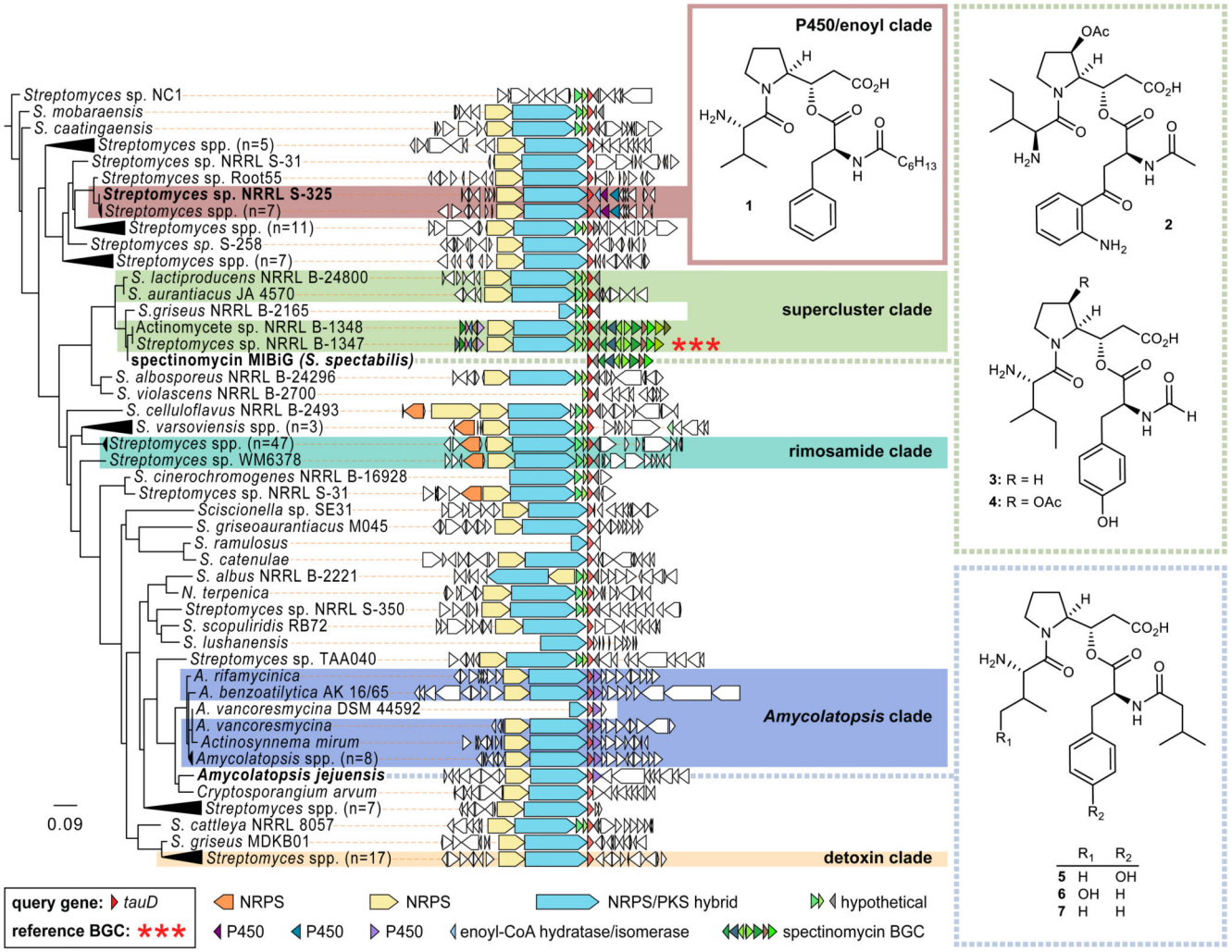


**Fig. 5. CORASON phylogeny of detoxin/rimosamide-related BGCs.**

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




Europe PMC Funders Author Manuscripts