nature chemical biology

ARTICLES

https://doi.org/10.1038/s41589-019-0400-9

A computational framework to explore large-scale biosynthetic diversity

Jorge C. Navarro-Muñoz[1,2,13], Nelly Selem-Mojica[3,13], Michael W. Mullowney[4,13], Satria A. Kautsar[1], James H. Tryon[4], Elizabeth I. Parkinson[5,11], Emmanuel L. C. De Los Santos[6], Marley Yeong[1], Pablo Cruz-Morales[3], Sahar Abubucker[7,12], Arne Roeters[1], Wouter Lokhorst[1], Antonio Fernandez-Guerra[8,9,10], 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]* and Marnix H. Medema[1]*

Genome mining has become a key technology to exploit natural product diversity. Although 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. In the present study, a streamlined computational workflow is provided, consisting of two new software tools: the 'biosynthetic gene similarity clustering and prospecting engine' (BiG-SCAPE), which facilitates fast and interactive sequence similarity network analysis of biosynthetic gene clusters and gene cluster families; and the 'core analysis of syntenic orthologues to prioritize natural product gene clusters' (CORASON), which elucidates phylogenetic relationships within and across these families. BiG-SCAPE is validated by correlating its output to metabolomic data across 363 actinobacterial strains and the discovery potential of CORASON is demonstrated by comprehensively mapping biosynthetic diversity across a range of detoxin/rimosamide-related gene cluster families, culminating in the characterization of seven detoxin analogues.

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

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

Initially, genome mining was performed on a single-genome basis: a research group or consortium would sequence the genome of a single microbial strain and attempt to identify and characterize each of its BGCs one by one. This approach has revealed much about the metabolic capacities of model natural-product-producing organisms such as *Streptomyces coelicolor*, *Sorangium cellulosum* and *Aspergillus nidulans*, and has provided clues regarding the discovery potential[10] from corresponding genera[11–13]. Computational tools for the identification of BGCs and the prediction of their products' chemical structures, such as antiSMASH[14–17] 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 (Minimum Information about a Biosynthetic Gene cluster) 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 can be extracted at once[25–28]. Such studies pave the path toward systematic investigations of the biosynthetic potential of broad taxonomic groups of organisms, as well as entire ecosystems. These large-scale analyses easily lead to the identification of thousands of BGCs with varying degrees of mutual similarity, ranging from widely distributed homologues of gene clusters for the production of well-known molecules to rare or unique gene clusters that encode unknown enzymes and pathways.

[1]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [2]Fungal Natural Products Group, Westerdijk Fungal Biodiversity Institute, Utrecht, the Netherlands. [3]Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, Mexico. [4]Department of Chemistry, Northwestern University, Evanston, IL, USA. [5]Carl R. Woese Institute for Genomic Biology and Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [6]Warwick Integrative Synthetic Biology Centre, University of Warwick, Coventry, UK. [7]Novartis Institutes for BioMedical Research, Cambridge, MA, USA. [8]Microbial Genomics and Bioinformatics, Max Planck Institute for Marine Microbiology, Bremen, Germany. [9]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [10]Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [11]Present address: Department of Chemistry, Purdue University, West Lafayette, IN, USA. [12]Present address: Sanofi, Cambridge, MA, USA. [13]These authors contributed equally: Jorge C. Navarro-Muñoz, Nelly Selem-Mojica, Michael W. Mullowney. *e-mail: n-kelleher@northwestern.edu; francisco.barona@cinvestav.mx; marnix.medema@wur.nl

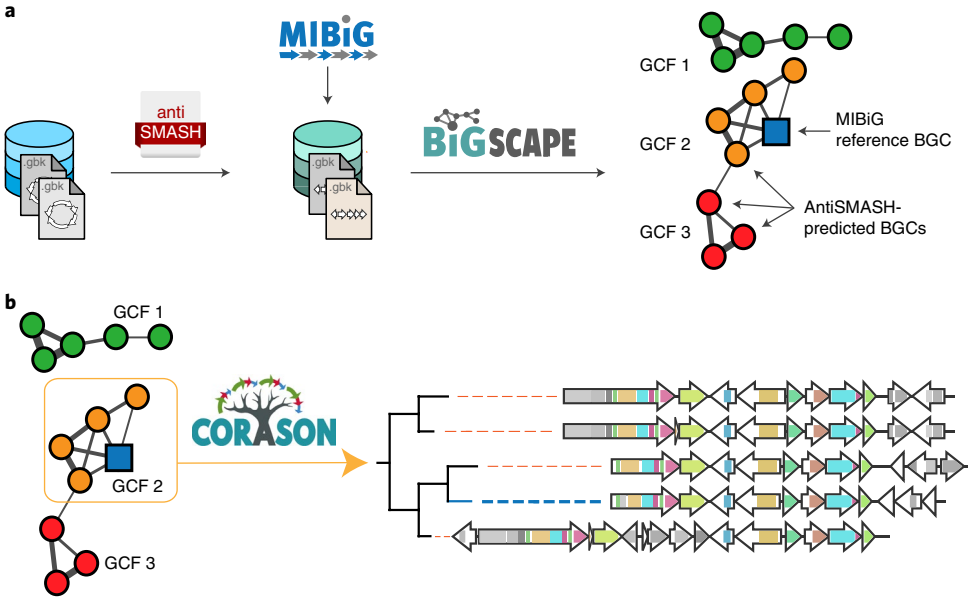

Fig. 1 | The BiG-SCAPE/CORASON workflow. a, The BiG-SCAPE approach analyzes a set of antiSMASH-detected BGCs to construct a similarity network and groups them into GCFs, together with MIBiG reference BGCs (indicated in blue). **b**, Subsequently, CORASON-based, multi-locus, phylogenetic analysis is used to illuminate evolutionary relationships of BGCs within each GCF.

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 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 molecular families 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 compute times and large-scale computing facilities when processing large datasets and lack a user-friendly implementation that interacts directly with other key resources. These shortcomings preclude adoption by the broader scientific community and impede substantial advances in natural product discovery.

In the present study, a streamlined computational workflow is provided that tightly integrates two new software tools, BiG-SCAPE and CORASON (https://bigscape-corason.secondarymetabolites.org), with the gene cluster identification and empirical biosynthetic data comparison possible through antiSMASH[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; in addition, it allows researchers to comprehensively identify all genomic contexts in which gene cassettes of interest ('subclusters' within larger BGCs) can be found. Our present study confirms that metabologenomic correlations accurately connect GCFs to mass features across metabolomic data from 363 strains. Furthermore, the power of the combined workflow is demonstrated, together with the EvoMining algorithm[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, BiG-SCAPE was built, written in Python and is freely available as open source software. BiG-SCAPE takes BGCs predicted by antiSMASH or annotated in MIBiG as inputs to automatically generate sequence similarity networks and assemble GCFs.

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

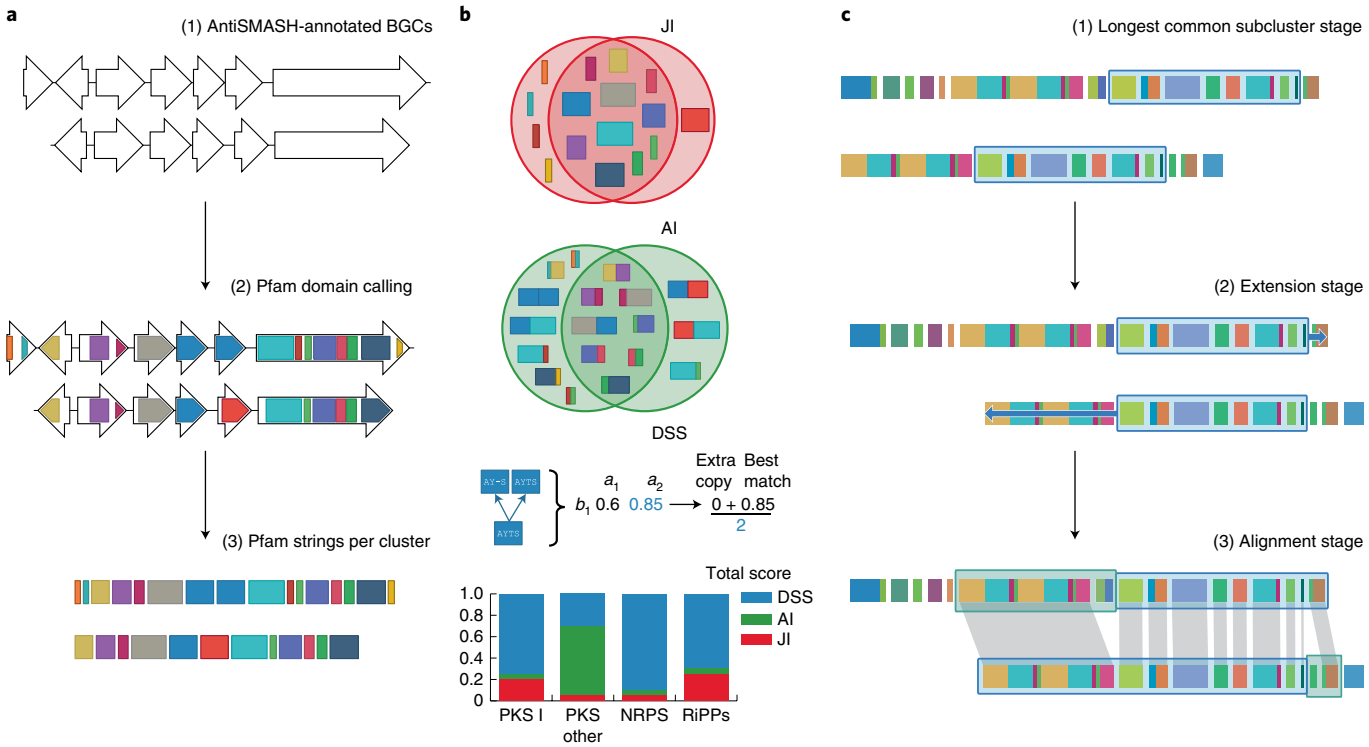

Fig. 2 | Main concepts in the BiG-SCAPE algorithm. a, Input data consist of BGC sequences directly imported from antiSMASH runs and/or MIBiG. Nucleotide sequences are translated and represented as strings of Pfam domains. **b**, The three metrics that are combined in a single distance include the JI, which measures the percentage of shared types of domains, the AI, which measures the percentage of pairs of adjacent domains, and the DSS, which is a measure of sequence identity between protein domains encoded in BGC sequences. Weights of these indices have been optimized separately for different BGC classes. For simplicity, only four classes are shown. **c**, In glocal mode, BiG-SCAPE starts with the longest common subcluster of genes between a pair of BGCs and attempts to extend the selection of genes for comparison using a match/mismatch penalty system.

both Pfam domain copy number differences and sequence identity. The combination of JI, AI and DSS indices into a new combined metric constitutes a fast and informative method to calculate distances between BGCs. BiG-SCAPE obtains very similar results in a fraction of the time compared with the previously published method[4] (see 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, whereas the amino acid sequence identity between their key biosynthetic enzymes has become less than 30–40% (ref. [3]). On the other hand, the structures of rapamycin-family polyketides exhibit major differences even when sequence identities are as high as ~80% (ref. [38]). Although there is not enough information available to construct individual metrics for each specific natural product family, specific weights of the JI, AI and DSS indices were calibrated for BGCs encoding eight different BiG-SCAPE classes (type I polyketide synthases (PKSs), other PKSs, nonribosomal peptide synthetases (NRPSs), PKS/NRPS hybrids, ribosomally synthesized and post-translationally modified peptides (RiPPs), saccharides, terpenes and others; 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 (see Methods and Supplementary Note 2). In the BiG-SCAPE output, separate networks are generated for each BiG-SCAPE class, along with an optional overall network that combines BGCs from all classes (see Supplementary Table 1).

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

BGC sequence similarity networks are then generated by applying a cutoff to the distance matrix calculated by BiG-SCAPE. Subsequently, two rounds of affinity propagation clustering[39] are performed to group BGCs into GCFs, and GCFs into 'gene cluster clans' (GCCs) (see Methods). Although tighter (lower) cutoffs are more appropriate for grouping BGCs that produce identical compounds, looser (higher) cutoffs provide a broader perspective on related families of natural products. This process of categorization facilitates calculating metabologenomic correlations[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, a chemical similarity network was constructed from all products of BGCs in MIBiG (see Methods), and this was used to derive a curated set of 376 compounds, which were manually classified into 92 groups (for example, 14-membered macrolides, benzoquinone ansamycins, quinomycin antibiotics and so on; see Supplementary Dataset) and 9

classes (for example, polyketides, NRPs, RiPPs and so on). Then BiG-SCAPE was used to group the corresponding BGCs into GCFs, and good correspondence was observed between manually curated families and those predicted by BiG-SCAPE (see Supplementary Fig. 1).

Arguably, the greatest value of BiG-SCAPE lies in the practical use of the predicted GCFs for discovery applications. Hence, the accuracy of correlations of BiG-SCAPE-predicted GCFs to MS ions was assessed from known natural products through metabologenomics[35]. First, a BiG-SCAPE analysis of 74,652 BGCs from 3,080 actinobacterial genomes (see Methods) was performed, including 1,393 reference BGCs from MIBiG[21]. BiG-SCAPE grouped these BGCs into a total of 17,718 GCFs and 801 GCCs using default parameters. Extracts from 363 actinomycete strains were analyzed using untargeted high-resolution liquid chromatography–tandem MS (LC–MS/MS)[4,35,40,41]. Exploration of gene cluster networks and molecular networks[42] highlighted the high diversity in both gene clusters and molecules, for example, 105 different BGCs were identified (at default < 0.3 distance) related to known detoxin/rimosamide gene clusters (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, using a binary scoring metric (see Supplementary Fig. 2) as described previously[4,35]. BiG-SCAPE's GCF annotations were then assessed against ion production patterns. A test dataset of nine known ion signals and their characterized gene clusters (for CE-108, benarthin, desertomycin, tambromycin, enterocin, tyrobetaine, chlortetracycline, rimosamide and oxytetracycline), which were known to be present across multiple strains in the data, were manually tracked across the four correlation rounds. Based on the metabologenomic analysis of the four rounds, the glocal mode with a 0.3 distance cutoff (Fig. 3d) was chosen as the default for BiG-SCAPE (see Supplementary Tables 2 and 3). Using these parameters, the analysis showed that at least six of these nine molecule–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 and see 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, the CORASON software was introduced, 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, approximately 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, for example, tailoring genes, to reveal various GCFs that would probably produce molecules with similar chemical modifications (Fig. 4c).

CORASON is available as 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, because GenBank outputs of antiSMASH can be used directly as inputs for the workflow (see Methods), and MIBiG reference data can be included in the analysis automatically. Although calculations on hundreds or thousands of genomes are too computationally intensive to provide them on a free public web server, the results of each BiG-SCAPE run are still made available in an interactive HTML visualization that enables efficient exploration of biosynthetic diversity across large datasets for nonprogrammers. 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 GCC. 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, an example output of an analysis with antiSMASH-predicted BGCs from 103 complete *Streptomyces* genomes (see Methods) has been provided, 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, multi-locus, whole-genome phylogeny (see Supplementary Fig. 4) was inferred from the *Streptomyces* conserved-core using CORASON, and the tree was decorated with the GCF absence/presence patterns (see Supplementary Fig. 5). As has been observed before in other genera such as *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 (1,564 GCFs (92%) occur across ≤3 genomes).

Case study: identification of new detoxin analogues. Analysis of 3,080 actinobacterial genomes revealed that detoxin and rimosamide BGCs are taxonomically widespread and architecturally diverse. Thus, the present study focused on GCFs of this class to showcase the ability of the BiG-SCAPE/CORASON workflow to analyze and map large, diverse biosynthetic datasets at high resolution[41] (see Supplementary Figs. 6 and 7). The conserved core (see 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 is present across all detoxin/rimosamide-related BGCs, but relatively unique within secondary metabolism, caught our attention. The product of the *tauD* gene belongs to the Fe(II)/2-oxoglutarate-dependent hydroxylase enzyme superfamily and is named for the commonly encoded 2-oxoglutarate-dependent taurine dioxygenase (TauD) involved in the oxygenolytic release of sulfite from the amino acid taurine[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 as being 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 paralogs

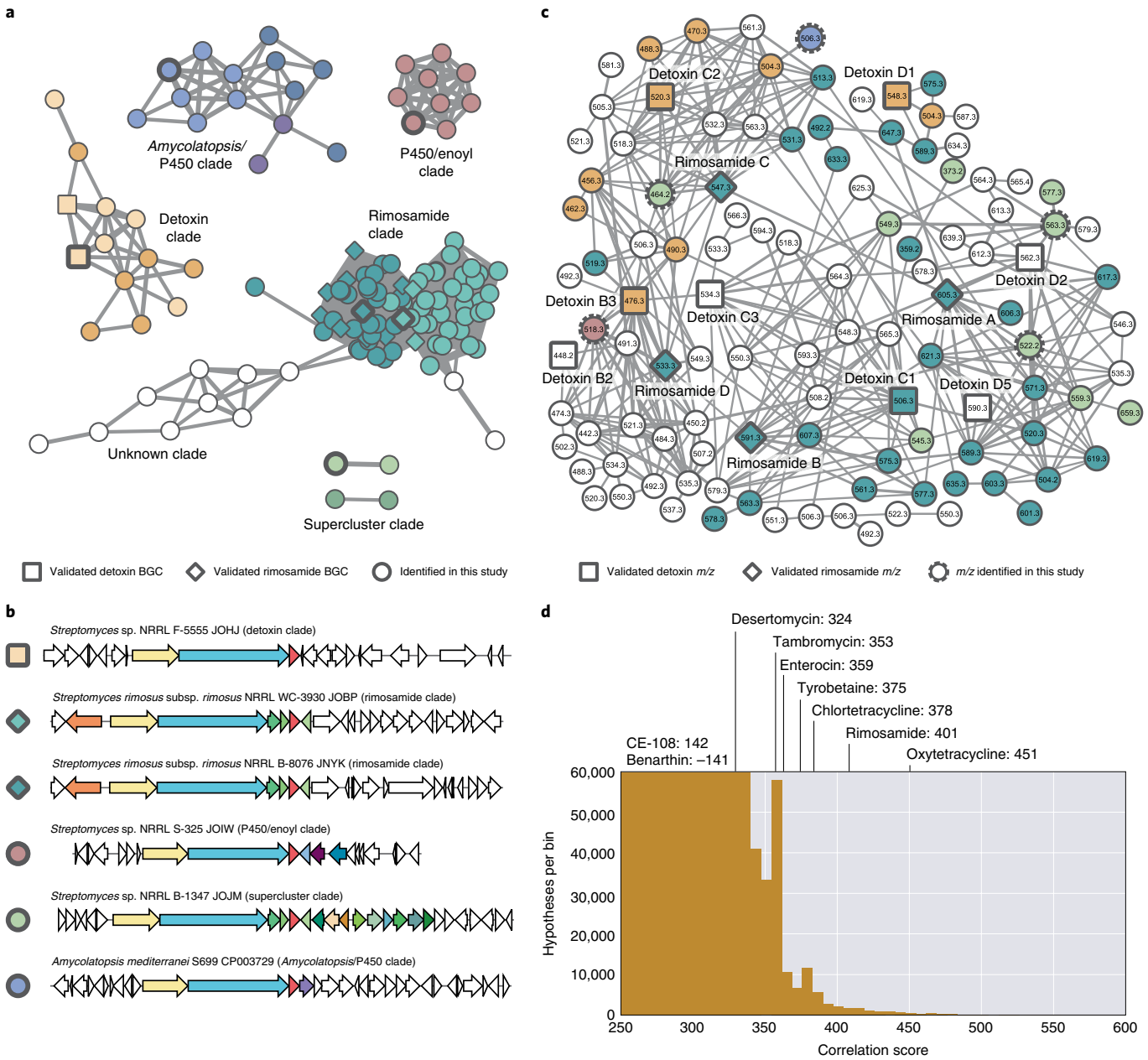

Fig. 3 | Sequence similarity and molecular networks of detoxins/rimosamides. a, Detail of a BiG-SCAPE network containing validated detoxin and rimosamide BGCs, filtered for the presence of the TauD domain. BiG-SCAPE GCF classifications include the rimosamide (turquoise shades) and detoxin (orange shades)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

across genera such as *Streptomyces*, *Rhodococcus*, *Frankia* and *Amycolatopsis* (see Supplementary Fig. 8). One expanded clade contained 15 *tauD* homologues that belonged to experimentally characterized BGCs from MIBiG v.1.3, as well as 1 within the rimosamide BGC (see Supplementary Table 4).

Next, the genomic contexts of all *tauD* expansions (comprising 1,175 BGCs) were investigated, with the goal of identifying new detoxin- and rimosamide-related BGCs. The BGCs were processed by CORASON using *tauD* as the query gene. Although ideally the detoxin/rimosamide BGC core would be defined as also containing the NRPS and NRPS/PKS hybrid genes, herein *tauD* was used as the sole member of the 'BGC core' to allow also for the identification of fragmented BGCs. Gaps in the genome sequences were observed for some organisms, including *Streptomyces humi* and *Amycolatopsis vancoresmycina* (Fig. 5). CORASON analysis revealed that the detoxin and rimosamide GCFs identified in BiG-SCAPE were part of a larger GCC related to peptide biosynthesis, which also comprised unexplored clades across the phylum Actinobacteria (Fig. 5

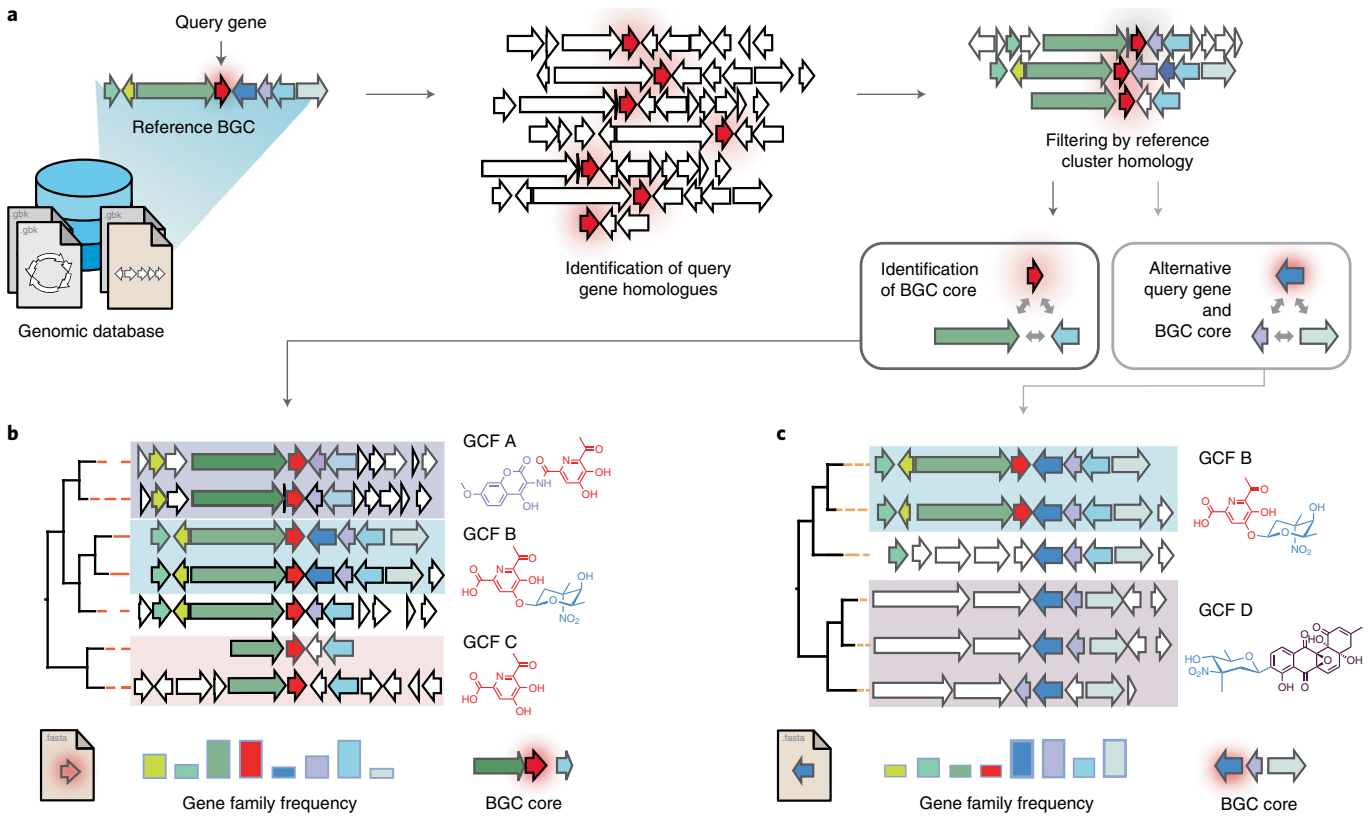

Fig. 4 | CORASON workflow. a, Given a query gene in a reference cluster and a custom genome database, CORASON (1) searches for query gene homologues, (2) creates a CVD by filtering out all genomic loci not related to the reference BGC, but keeping fragmented clusters and (3) identifies the CVD gene core based on multidirectional best hits. **b**, Then, CORASON infers a phylogenetic tree by curation and concatenation of the CVD gene core, and calculates the frequency of occurrence for each gene family from the reference BGC. The tree will reveal clades of BGCs that may correspond to GCFs from BiG-SCAPE, and may be responsible for the production of different structural analogues of a natural product family. **c**, With the same reference BGC, if a new query gene is selected from accessory enzymes rather than the current CVD core, CORASON will visualize a new phylogeny. This tree may contain clades that correspond to GCFs with diverse biosynthetic cores (of scaffold biosynthesis enzymes) that encode the same molecular modifications in different contexts.

and see 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 merging of adjacent BGCs by antiSMASH made these BGCs sufficiently different to be classified into different GCFs under the cutoffs used, whereas CORASON could organize BGC relationships based on the single *tauD* gene (see Supplementary Figs. 3 and 10).

It was hypothesized that the detoxins produced from BGCs in the unexplored clades would contain new chemical variations related to the observed genetic variations. Fortunately, 40 of the 152 strains harboring these BGCs were represented in the 363-strain, LC–MS/MS metabolomics dataset. Molecular networking analysis of these data (see Methods) indicated the presence of eight known detoxins, four known rimosamides and 99 putatively new 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 MS/MS data from extracts of *Streptomyces* sp. NRRL S-325, which has a BGC from this clade, and comparison with fragmentation patterns of known detoxins, led to the discovery of detoxin S_1 (**1**; Fig. 5 and see Supplementary Figs. 11 and 12). This contained a heptanamide side chain, a unique substructure among the detoxins and rimosamides that is probably installed by the condensation domain of the NRPS, potentially following processing by the predicted enoyl-CoA hydratase/isomerase and cytochrome P450s.

The second clade of interest, termed the 'supercluster clade', comprised BGCs with genes related to detoxin biosynthesis immediately adjacent to the known spectinomycin BGC[45] (Fig. 5). This was discovered because the spectinomycin MIBiG entry (BGC0000715) clustered with them on the CORASON tree, as it contained a *tauD* gene at its periphery (Fig. 5). Since the *tauD* gene is not known to be involved in spectinomycin biosynthesis, it was hypothesized that there were likely additional detoxin genes adjacent to this spectinomycin BGC in *S. spectabilis* NRRL-2792. This strain was acquired to determine whether CORASON analysis could facilitate prediction of detoxin production based solely on the presence of a single query gene. MS/MS analysis of 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 analogue, 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.

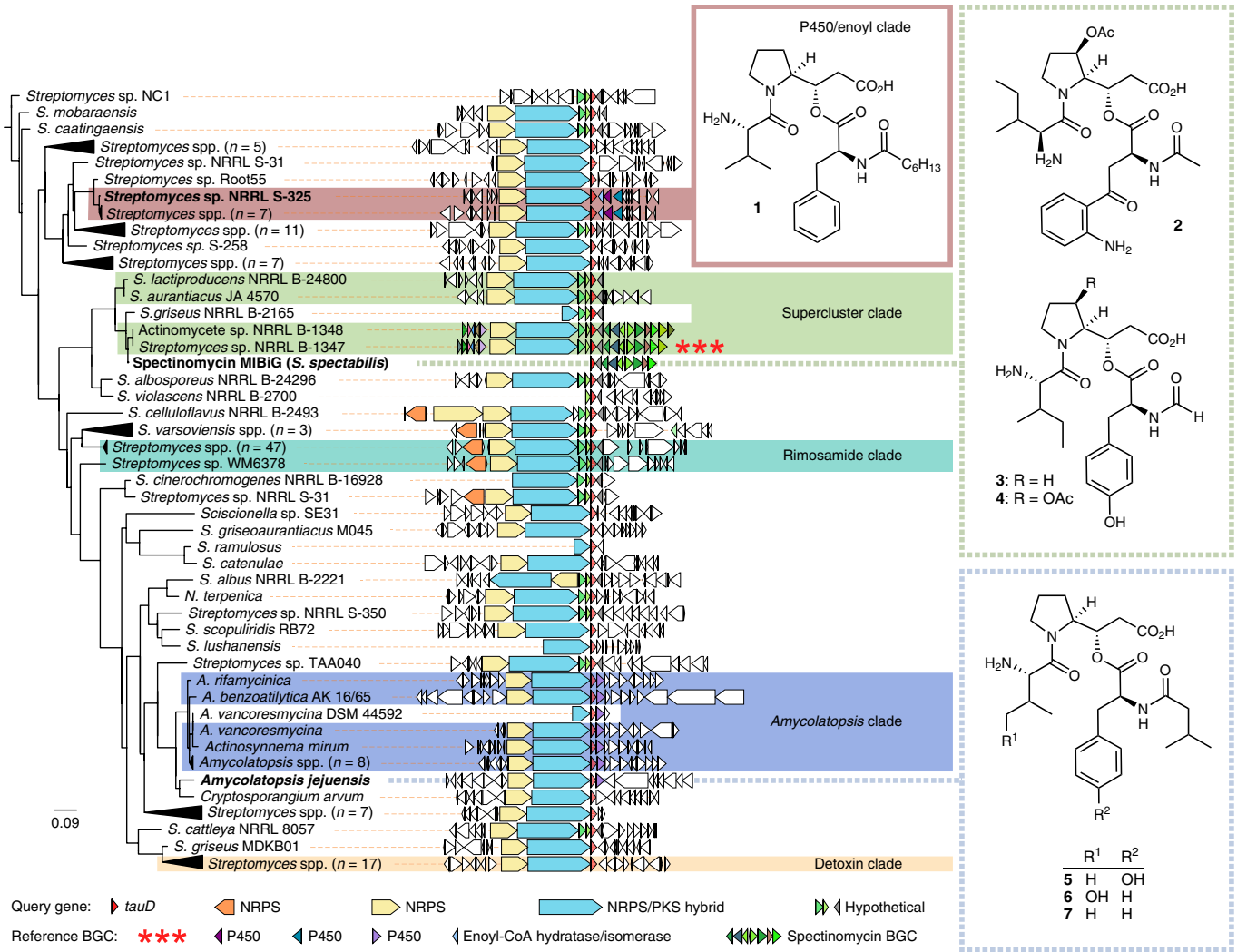

Fig. 5 | CORASON phylogeny of detoxin/rimosamide-related BGCs. CORASON phylogenetic reconstruction with *tauD* as the query gene and the *Streptomyces* sp. NRRL B-1347 BGC as query cluster, rooted with *tauD* from *Streptomyces* sp. NC1. Branches of redundant and highly divergent BGCs were compressed for readability (see the uncompressed tree in Supplementary Fig. 9). (In the Supplementary Figure, names are followed by their GenBank accession numbers when available.) Genes not found in the reference cluster are colored based on BLAST analysis. Highlighted sections on the tree correspond to BiG-SCAPE-defined families. Bolded strain/BGC names were those investigated in the present study, with dotted lines indicating BGCs and detoxins discovered just outside the BiG-SCAPE-defined families. The representative structures for each clade illustrate the correspondence between molecular and genomic variations.

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

The third detoxin clade that was 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 there were no metabolomics data for strains with BGCs in the BiG-SCAPE-defined GCF, the CORASON visualization allowed the selection of an *Amycolatopsis* strain in the present dataset with a very similar BGC (80–90% amino acid identity for the core genes) that contains a homologue of the desired P450 gene (Fig. 5, adjacent to the *Amycolatopsis*/P450 clade). Analysis of MS/MS data from

an *Amycolatopsis jejuensis* NRRL B-24427 fermentation extract revealed detoxin isomers P_1 (**5**, Fig. 5; see Supplementary Figs. 13 and 21), containing a tyrosine, P_2 (**6**, Fig. 5; see Supplementary Figs. 13 and 22), featuring phenylalanine and a hydroxylated valine, as well as detoxin P_3, a closely related analogue free of hydroxylation (**7**, Fig. 5; see Supplementary Figs. 13 and 23). Only five of the seven new detoxins described in the present study appear as nodes in the molecular network of Fig. 3c, with the notable absence of two P-series analogues. This is because detoxin isomers P_1 and P_2 had a cosine similarity >0.6 and were collapsed into one node, whereas 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 isotope-labeled amino acids (see Supplementary Figs. 24–33). Detailed structural analysis for compounds **1**–**7**, including results from feeding studies using stable isotope-labeled amino acids, deconvolution of MS/MS spectra and full ^{1}H, ^{13}C and 2D NMR assignments for compound **2**, are available in Supplementary Note 3, Supplementary Figs. 15–33 and Methods. Previously reported detoxins and rimosamides antagonize blasticidin-S inhibition of *Bacillus cereus*, a bioactivity that will be investigated for these analogues in follow-up studies[41,48].

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

Discussion

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

Furthermore, the ability to perform phylogenetic analyses of large sets of complete BGCs, as well as their individual genetic components, a long-standing challenge that has remained unsolved since first posed in 2008 (ref. [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 subclusters. A logical next step will be the unified classification of the millions of BGCs within publicly available genome sequences, and a Pfam-like database for the assignment of biosynthetic GCFs to known and unknown areas of natural product chemical diversity.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41589-019-0400-9.

Received: 3 December 2018; Accepted: 4 October 2019;
Published online: 25 November 2019

References

1. Traxler, M. F. & Kolter, R. Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.* **32**, 956–970 (2015).
2. Davies, J. Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361–364 (2013).
3. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
4. Doroghazi, J. R. et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
5. Dejong, C. A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12**, 1007–1014 (2016).
6. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
7. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl Acad. Sci. USA* **114**, 5601–5606 (2017).
8. Cruz-Morales, P. et al. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. *Genome Biol. Evol.* **8**, 1906–1916 (2016).
9. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
10. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
11. Bentley, S. D. et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147 (2002).
12. Schneiker, S. et al. Complete genome sequence of the myxobacterium *Sorangium cellulosum*. *Nat. Biotechnol.* **25**, 1281–1289 (2007).
13. Bergmann, S. et al. Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat. Chem. Biol.* **3**, 213–217 (2007).
14. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
15. Blin, K. et al. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41**, W204–W212 (2013).
16. Weber, T. et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**, W237–W243 (2015).
17. Blin, K. et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
18. Johnston, C. W. et al. An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.* **6**, 8421 (2015).
19. Skinnider, M. A. et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* **43**, 9645–9662 (2015).
20. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
21. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
22. Nielsen, J. C. et al. Global analysis of biosynthetic gene clusters reveals vast potential of secondary metabolite production in *Penicillium* species. *Nat. Microbiol.* **2**, 17044 (2017).
23. Tobias, N. J. et al. Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nat. Microbiol.* **2**, 1676–1685 (2017).
24. Grubbs, K. J. et al. Large-scale bioinformatics analysis of *Bacillus* genomes uncovers conserved roles of natural products in bacterial physiology. *mSystems* **2**, e00040–17 (2017).
25. Freeman, M. F. et al. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* **338**, 387–390 (2012).
26. Agarwal, V. et al. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat. Chem. Biol.* **13**, 537–543 (2017).
27. Owen, J. G. et al. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. *Proc. Natl Acad. Sci. USA* **112**, 4221–4226 (2015).
28. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

29. Leao, T. et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. *Proc. Natl Acad. Sci. USA* **114**, 3198–3203 (2017).
30. Ziemert, N. et al. Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc. Natl Acad. Sci. USA* **111**, E1130–E1139 (2014).
31. Medema, M. H. et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* **10**, e1003822 (2014).
32. Mohimani, H. et al. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. *J. Nat. Prod.* **77**, 1902–1909 (2014).
33. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
34. Nguyen, D. D. et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl Acad. Sci. USA* **110**, E2611–E2620 (2013).
35. Goering, A. W. 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.* **2**, 99–108 (2016).
36. Duncan, K. R. et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chem. Biol.* **22**, 460–471 (2015).
37. Punta, M. et al. The Pfam protein families databases. *Nucleic Acids Res* **40**, D290–D301 (2012).
38. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
39. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
40. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).
41. McClure, R. A. et al. Elucidating the rimosamide-detoxin natural product families and their biosynthesis using metabolite/gene cluster correlations. *ACS Chem. Biol.* **11**, 3452–3460 (2016).
42. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752 (2012).
43. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
44. Hausinger, R. P. Fe(II)/α-ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem Mol. Biol.* **39**, 21–68 (2004).
45. Kim, K.-R., Kim, T.-J. & Suh, J.-W. The gene cluster for spectinomycin biosynthesis and the aminoglycoside-resistance function of spcM in *Streptomyces spectabilis*. *Curr. Microbiol.* **57**, 371–374 (2008).
46. Sinha, A., Phillips-Salemka, S., Niraula, T.-A., Short, K. A. & Niraula, N. P. The complete genomic sequence of *Streptomyces spectabilis* NRRL-2792 and identification of secondary metabolite biosynthetic gene clusters. *J. Ind. Microbiol Biotechnol.* **46**, 1217–1223 (2019).
47. Ogita, T., Seto, H., Otake, N. & Yonehara, H. The structures of minor congeners of the detoxin complex. *Agric. Biol. Chem.* **45**, 2605–2611 (1981).
48. Yonehara, H., Seto, H., Aizawa, S., Hidaka, T. & Shimazu, A. The detoxin complex, selective antagonists of blasticidin S. *J. Antibiot. (Tokyo)* **21**, 369–370 (1968).
49. Fischbach, M. A., Walsh, C. T. & Clardy, J. The evolution of gene collectives: how natural selection drives chemical innovation. *Proc. Natl Acad. Sci. USA* **105**, 4601–4608 (2008).

Methods

Dataset. A set of 2,831 actinobacterial genomes was downloaded from the National Center for Biotechnology Information (NCBI) by querying for "Whole genome shotgun sequencing project" or "Complete genome" in combination with the taxonomic identifier for Actinobacteria. The orders *Propionibacteriales*, *Micrococcales*, *Corynebacteriales* and *Bifidobacteriales* were excluded, because they contain large numbers of genomes without relevant natural product-producing capacity, except the *Nocardiaceae* family from *Corynebacteriales* (see below). To these, 249 additional draft assemblies from the Metcalf lab were added (for example, *Streptomyces* sp. B-1348; see BioProject PRJNA488366). Draft genome assemblies from this BioProject were obtained using SPAdes[51] with default options.

All files were processed using antiSMASH v.4 (ref. [17]) (parameters: `--minimal`). The antiSMASH-annotated genome sequences are available from ref. [50] (antiSMASH_results_Metcalf_B, antiSMASH_results_Metcalf_J and antiSMASH_results_NCBI).

To the resulting 73,260 predicted BGCs, 1,393 more were added from the MIBiG[21] (release 1.3, August 2016, antiSMASH-analyzed versions from each entry) as reference data.

This final BGC set was then analyzed using BiG-SCAPE v.31 of the Pfam database. The "hybrids" mode, which allows BGCs with mixed annotations to be analyzed in their individual class sets (for example, a BGC annotated as *lanthipeptide-t1pks* will be analyzed as both an RiPP and a PKSI) was enabled. Two results sets were created (BiG-SCAPE results network files are available from ref. [50]): one with the global mode enabled and the other with glocal mode enabled (see Fig. 2).

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

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

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

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

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 (ref. [52]) (parameters: `-maxiters 2`) and MAFFT v.7.407 (ref. [53]) (parameters: `--auto`); the three versions of the code are available from ref. [50] (Alignment Method Comparison). MUSCLE was parallelized using Python's `pool.map` on single-core instances for each domain sequence fasta file, whereas MAFFT was parallelized on each file with its `--thread` parameter. Comparison of the final GCF calling (using BiG-SCAPE's `--mix` parameter) indicates high agreement across the three methods (see Supplementary Fig. 34), with hmmalign showing shorter runtimes as the number of BGCs in the input data increases (see Supplementary Fig. 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 first to identify the most suitable cutoff for clustering algorithm evaluation. The targeted attack removes the edges above a certain cutoff value while calculating, for each iteration, the number of nodes and graph density, and identifying the connected components after removal of isolated vertices (BGCs). Network statistics such as the number of vertices/edges lost for each cutoff value, as well as the size of the connected components that emerged, were calculated during the attack.

Supplementary Fig. 36 shows the dynamics and impact of the different filtering thresholds applied to the different BGC training networks, with a cutoff of 0.75 being the value that maximized the number of nodes, while minimizing the impact on the structural integrity of the network. This analysis was performed using the igraph package[54] for the network analyses and ggplot2 (ref. [55]) for plotting.

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

Input data. The input BiG-SCAPE consists of text files in GenBank format (.gbk extension) and the Pfam database[37] (already processed with hmmpress). Although BiG-SCAPE can work with files not processed by antiSMASH, it relies on antiSMASH's product prediction to separate the BGCs in their correct biosynthetic class, thereby reducing computational time. If the product annotation is unknown, missing, or several different classes are mixed, the BGC will be classified as 'Other'.

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

Distance calculation. Pairwise distance calculation is divided between three values that measure (1) the percentage of shared domain types (JI), (2) the similarity between aligned domain sequences (DSS index; domains from the same type are first matched for best similarity using Munkres' algorithm, as implemented in the Scikit-Learn library[56]) and (3) the similarity of domain pair types (AI). For specific details of each index, see Supplementary Note 1.

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

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

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

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

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

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

Output. BiG-SCAPE produces high-quality Scalable Vector Graphics (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 (1) appear with the highest frequency in the GCF and (2) are detected in the central (or exemplar) cluster, defined by the affinity propagation cluster. All copies of the Core Domains in the exemplar are concatenated, as well as those from the best matching domains of the rest of the BGCs in the GCF (aligned domain sequences are used). The tree is constructed using FastTree[43] (default parameters). Visual alignment is attempted using the position of the 'longest common information' from the distance calculation step (between the exemplar BGCs and each of the other clusters).

Availability. BiG-SCAPE is written in Python and is currently compatible with both Python v.2 and v.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, choosing the weight combination that maximized the correlation between BGC and compound distances for every pair of BGCs in the same class in a manually curated compound group table (see Supplementary Dataset). The dataset comprised all BGCs from the MIBiG database (v.1.3) that had linked compound SMILES and at least two predicted domains to filter out minimal gene cluster entries. BGC distances were calculated by moving in steps of 0.01 across the JI, DSS and original Goodman–Kruskal (GK)[57] indices and AIs, 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 four anchor domains were considered (condensation domain, PF00668; beta-ketoacyl synthase N-terminal domain, PF00109; beta-ketoacyl synthase C-terminal domain, PF02801 and the terpene synthase N-terminal domain, PF01397). Compound distances were calculated only once, between all BGCs in the MIBiG v.1.3 that had an annotated SMILES string representing the molecule. Their pairwise distance was calculated using RDKit (Tanimoto's coefficient based on Morgan's fingerprinting, radius = 4). The nine original curated compound classes were used to tune the weights of seven BiG-SCAPE classes (the terpene BiG-SCAPE class was initially included in the 'others' compound class due to a low number of points and was assigned default weights: JI = 0.2, DSS = 0.75 and AI = 0.05).

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

Comparison with other methods. To compare BiG-SCAPE with the GCF algorithm in Doroghazi et al.[4], 11,618 GenBank files were reconstructed from data related to that study (allClusterProts.fasta file from https://www.igb.illinois.edu/labs/metcalf/gcf/search.html). These reconstructed cluster files were analyzed using antiSMASH v.4, and its output (ref. [50]; Comparison Doroghazi2014 reconstructed BGCs antiSMASH results) to make a run in BiG-SCAPE (ref. [50]; Comparison Doroghazi2014 BiG-SCAPE results).

Unlike Doroghazi's method, BiG-SCAPE follows a two-step process to infer GCFs. First, it filters the resulting network using a predefined empirical cutoff distance of 0.3, and later the GCFs are identified by the affinity propagation clustering algorithm. This two-step approach partitions the natural emerging components from the filtering step, increasing the resolution of the inferred GCFs. To provide a fair comparison with GCFs inferred by Doroghazi et al.[4], the natural emerged components were used after the filtering steps and the different clustering results were compared; good agreements were found between both methods (see Supplementary Fig. 42 for details), although BiG-SCAPE took only a fraction of the runtime of the previously published tool (see 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 are pairs of genes that exist in two different sets of genes (genomes, metagenomes or BGCs) that are more similar with each other than with any other sequence in the set pair. In CORASON, this relationship was generalized in a stricter algorithm that considers all-versus-all comparisons between every set in the collection to remove paralogues and conserve only true orthologues. As a result, the conserved core is composed of gene families that are each guaranteed to be a best bidirectional hit across the whole collection (although they need not be contiguous).

The BGC-conserved core facilitates reconstruction of the BGC evolutionary history in a multi-locus tree. The query gene assures that at least one element will be present in the conserved core and will also be used to visually align the BGC variations in the graphic output.

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

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

Formally, for H defined as:

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

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

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

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

Phylogenetic reconstruction and gene cluster alignment. For each BGC, its conserved core sequences are concatenated and then aligned using MUSCLE v.3.8.31 (ref. [52]). The alignments are curated using Gblocks[58] with a minimum block length of five positions, a maximum of ten contiguous nonconserved positions and considering only positions with a gap in less than 50% of the sequences in the final alignment. If the curation turns out to be empty, then the noncurated alignment will be used for the tree. If the alignment itself is empty, it is recommended to reduce the score cutoff or the scope of the taxonomic diversity on the genomic database. Without the alignment, BGCs will be drawn but not sorted. Approximately maximum-likelihood phylogenetic trees are inferred using FastTree[43] v.2.1.10 from the curated amino acid alignment.

BGC prioritization graphic output. CORASON produces an SVG file containing the BGC variations sorted as stated by the phylogenetic reconstruction and aligned according to the query enzyme. The Newick tree is converted to SVG by applying Newick Utilities v.1.6 (ref. [59]) 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 online at https://github.com/nselem/corason/wiki.

***Streptomyces* closed genome analysis.** Sequences from 103 complete *Streptomyces* genomes were retrieved from the NCBI by querying for "*Streptomyces*" and "complete genome" not "segment". Two genomes corresponding to *C. acidiphila* and *S. arenicola* (CP001700 and CP000850) were used as outgroups. These genomes were analyzed by antiSMASH v.4 and the resulting gene cluster (ref. [50]; Closed_Streptomyces_antiSMASH_results) files were used as input for

BiG-SCAPE (ref. [50]; Closed_Streptomyces_BiG-SCAPE_results). The conserved core was extracted and curated using the CORASON algorithm. The tree was constructed using FastTree with default values over a matrix of 114,051 amino acids in size, from 446 conserved gene families (ref. [50]; 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 (see Supplementary Fig. 8), *a tauD* sequence from *Escherichia coli K12* was used as query to conduct a blast search against the reduced genomic database of 1,917 Actinobacteria genomes (*e*-value 0.001), followed by an EvoMining analysis and a search for recruitments on MIBiG database (*e*-value 0.001). Recovered *tauD* orthologues were aligned with MUSCLE v.3.8.31 (ref. [52]) and alignments were curated using Gblocks[58] as described above. An unrooted approximately maximum-likelihood tree was built using FastTree[43]. The tree was colored using Newick Utilities[59] according to BiG-SCAPE families.

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

Molecular networking methods. *Cultivation of actinomycetes for MS-based metabolomics.* All strains analyzed for metabolomics were grown on four media types: arginine/glycerol/salts, mannitol/soy flour, ISP medium 4 or glycerol/sucrose/beef extract/casamino acids as previously reported[4]. After 10 d of growth, plates were frozen, then thawed and pressed to release spent liquid media. Media were then filtered and extracted using 30 mg Supel-Select HLB SPE cartridges (Supelco) and resuspended to a concentration of approximately 2 mg ml^{-1} in 5% acetonitrile before LC–MS analysis.

Acquisition and analysis of LC–MS metabolomics data. All LC–MS/MS analyses were performed using an Agilent 1150 HPLC coupled with a Q-Exactive mass spectrometer (Thermo Fisher Scientific). Reversed-phase chromatography was performed at a 200 µl min^{-1} flow rate on a Phenomenex Kinetex C18 RP-HPLC column (150 × 2.1 mm^2 inner diameter, 2-µm particle size, 100 Å pore size (1 Å = 0.1 nm). Mobile phase A was water with 0.1% formic acid and mobile phase B was acetonitrile with 0.1% formic acid. Mass spectral data for both MS and MS/MS were acquired using a 250–3750 *m/z* scan range, a resolution of 35,000, a maximum inject time of 40 ms and an AGC target value of 1×10^6. The top five most intense ions in each full MS spectrum were targeted for fragmentation by higher-energy collisional dissociation at 25 eV. MS/MS data were analyzed using spectral networking as previously described[60]. 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 >0.75, yielding 5,824 ions detected in two or more strains.

LC–MS molecular networking. Molecular networking was performed as previously reported[40,60]. Briefly, individual MS/MS scans were extracted from each MS raw file and filtered to remove the 25% of ions with the lowest intensity. Each MS/MS scan was further processed by taking the square root of each ion's intensity and normalizing it so that the sum of all intensities in each MS/MS scan was equal to 1. Cosine similarities were calculated between all MS/MS scans, with scores ranging between 0 and 1, with a score of 1 indicating that two MS/MS scans were identical. Precursor ions were determined to be identical if they were within 0.01 *m/z* and their corresponding MS/MS spectra had a cosine similarity score >0.6. A visualization of the network was constructed in Cytoscape by drawing edges between scan nodes with a cosine similarity >0.6. The network was manually analyzed to identify ions related to known detoxins and rimosamides, which were found to cluster together as one molecular family, with 99 putatively new analogues, a subset of which were characterized herein as detoxins 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 medium Petri plates containing arginine/glycerol/salts medium (1 l of distilled water, 15 g of agar, 1 g of arginine, 12.5 g of glycerol, 1 g of potassium phosphate dibasic, 1 g of sodium chloride, 0.5 g of magnesium sulfate heptahydrate, 10 mg of iron(II) sulfate hexahydrate, 1 mg of copper(II) sulfate pentahydrate, 1 mg of manganese(II) sulfate monohydrate and 1 mg of zinc sulfate heptahydrate).

Amycolatopsis jejuensis NRRL B-24427 was obtained from the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA) and was grown on solid agar medium Petri plates containing mannitol/soy flour medium (1 l of distilled water, 15 g of agar, 20 g of D-mannitol and 20 g of soy flour). For all metabolic labeling experiments, the medium 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 5 d of incubation in the presence of stable isotope-labeled amino acids, plates were frozen overnight at −20 °C, thawed and pressed to release spent liquid medium. Extracellular secondary metabolites were extracted using 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 ^2H_2O. ^{1}H, ^{13}C, correlation spectroscopy, heteronuclear single quantum coherence, heteronuclear multiple bond correlation and nuclear Overhauser enhancement spectroscopy spectra were obtained on a Bruker NEO spectrometer (600 MHz 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 (δ_H 8.44; δ_C 171.67). ^{1}H and ^{13}C NMR resonances of compound **2** are reported in Supplementary Note 3i.

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

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Genomes used in this study include assemblies from the sequencing project deposited in NCBI BioProject PRJNA488366, in Sequence Read Archive runs with accession numbers SRX4638772 to SRX4639021. AntiSMASH, BiG-SCAPE and CORASON results for all genome assemblies, along with raw files of phylogenetic trees, are available from ref. [50]. Fully annotated nucleotide sequences for the BGCs for detoxin S_1, detoxins 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 accession numbers BGC0001840, BGC0001878 and BGC0001841, respectively. All raw MS data files for strains producing one or more of the nine compounds used for correlation analysis have been submitted to MassIVE under accession number MSV000083738. Raw MS data files and isolated MS/MS scan files for all newly identified detoxin analogues have been uploaded to MassIVE with accession number MSV000083648, and MS/MS data for other strains are available upon request.

Code availability

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 at https://git.wur.nl/medema-group/BiG-SCAPE and CORASON project at https://github.com/nselem/corason.

References

50. Navarro-Muñoz J. C., Selem-Mojica N., Mullowney M. et al. Zenodo https://doi.org/10.5281/zenodo.1532752 (2018)..
51. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
53. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
54. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
55. Wickham, H. et al. ggplot2: an implementation of the grammar of graphics. R package version 7 http//CRANR-projectorg/package=ggplot2 (2008).
56. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn Res.* **12**, 2825–2830 (2011).

57. Lin, K., Zhu, L. & Zhang, D. Y. An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* **22**, 2081–2086 (2006).
58. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
59. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
60. Henke, M. T. et al. New aspercryptins, lipopeptide natural products, revealed by HDAC inhibition in *Aspergillus nidulans*. *ACS Chem. Biol.* **11**, 2117–2123 (2016).

Acknowledgements

We thank the following: the ARS of the USDA for providing bacterial strains; H. Sook Ann, Z. Crispino, Y. Kim, N. Ciszek and K. Espejo for generating bacterial culture extracts; R. McClure, M. Robey and G. Miley for assistance with and contributions to metabolomic data collection methods and acquisition; and Dr. Y. Zhang and Dr. Y. Wu of the Integrated Molecular Structure Education and Research Center (IMSERC) at Northwestern University for assistance in acquiring NMR data. Some analyses were carried out using CONABIO's computing cluster, with funds from the Secretariat of Environment and Natural Resources. We thank K. Blin for technical assistance with setting up the website on the secondarymetabolites.org domain. The research reported in this publication was supported by the Netherlands Organization for Scientific Research (grant no. 863.15.002 to M.H.M.), the Graduate School for Experimental Plant Sciences (grant to M.H.M.); National Institutes of Health (NIH) Genome to Natural Products Network supplementary award (no. U01GM110706 to M.H.M.), CONACyT grants (grant nos. CBS2017_285746 and 2017_051TAMU to F.B.-G.; postdoctoral scholarship 263661 to J.C.N.M.; PhD scholarship 204482 to N.S.M. (who was also supported by the Innovation Secretary of Guanajuato)), the National Cancer Institute of the NIH (award no. F32CA221327 to M.W.M.), the National Institute of General Medical Sciences (award no. F32GM120999 to E.I.P.), the São Paulo Research Foundation (FAPESP, grant no. 17/08038-8 to L.T.D.C.), the National Center for Complementary and Integrative Health of the NIH (award no. R01AT009143 to R.J.T. and N.L.K.) and Warwick Integrative Synthetic Biology Centre, a UK Synthetic Biology Research grant from the Biotechnology and Biological Sciences Research Council and Engineering and Physical Sciences Research Council (grant no. BB/M017982/1 to E.L.C.D.L.S). This work made use of the IMSERC at Northwestern University, which has received support from the NIH (grant nos. 1S10OD012016-01/1S10RR019071-01A1), the State of Illinois and the International Institute for Nanotechnology. A.F.-G. received funding from the European Union's Horizon 2020 research and innovation program (Blue Growth: Unlocking the Potential of Seas and Oceans; grant agreement no. 634486).

Author contributions

R.J.T., W.W.M., N.L.K., F.B.-G. and M.H.M. originally conceived of the research and coordinated the work. J.C.N.M. designed and developed BiG-SCAPE, with the help of S.A.K., E.L.C.D.L.S., M.Y., S.A., A.R., W.L., A.F.-G. and M.H.M. S.A.K. designed the output visualizations with the help of J.C.N.M. and E.L.C.D.L.S. N.S.M. designed and developed CORASON, with the help of P.C.M. and F.B.-G. M.W.M., J.H.T., E.I.P., L.T.D.C., A.W.G., R.J.T., W.W.M. and N.L.K. designed and performed the experimental research. J.C.N.M., N.S.M., M.W.M., J.H.T., F.B.-G. and M.H.M. wrote the first draft of the manuscript and all authors participated in editing the manuscript.

Competing interests

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

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-019-0400-9.

Correspondence and requests for materials should be addressed to N.L.K., F.B.-G. or M.H.M.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Marnix Medema

Last updated by author(s): Aug 28, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Thermo Xcalibur version 4.0.27.10, MestReNova 10.0.2-15465
Data analysis	antiSMASH v4.0, BiG-SCAPE commit 05518eed9dcf9b73da6ace10096323300388f665, biopython 1.70, fasttree 2.1.10, hmmer 3.1b2, Pfam version 31, networkx 2.1, using numpy 1.14.0, scikit-learn 0.19.1, scipy 1.0.0, CORASON commit 3a288a5fd4b3e1d47f9d95abf7e317acb8533f2c, Muscle 3.8.31, MAFFT v7.407, Gblocks 0.91b, Newick Utilities 1.6, xmeasures v4.0.2, igraph (R) 1.1.0, ggplot2 v3.1.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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 S1, Detoxin N2-N3 and detoxins P1–P3 have been deposited in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under accession numbers BK010707, BK010851 and BK010852, respectively, and in MIBiG under accessions BGC0001840, BGC0001878 and BGC0001841, respectively. All raw mass spectrometry 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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size was precalculated; rather, all publicly available genomic and metabolomic data was used for the data analysis, thus maximizing the amount of input data.
Data exclusions	No data was excluded.
Replication	Analytical chemistry experiments (e.g., mass spectrometry) were replicated multiple times as indicated in the Supplementary Information.
Randomization	Randomization was not relevant for our study, as we did not perform group comparisons.
Blinding	No blinding was performed in our study, as no group comparisons were performed and there was no specific reason to expect bias, as MS/MS and NMR are unbiased techniques..

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging