iPRESTO: automated discovery of biosynthetic sub-clusters linked to specific natural product substructures

Joris J.R. Louwen[1], Satria A. Kautsar[1], Sven van der Burg[2], Marnix H. Medema[1*], Justin J.J. van der Hooft[1,3*]

1. Bioinformatics Group, Wageningen University, Wageningen, the Netherlands

2. Netherlands eScience Center, Amsterdam, the Netherlands

3. Department of Biochemistry, University of Johannesburg, Johannesburg, South Africa

* Corresponding authors

E-mail: marnix.medema@wur.nl, justin.vanderhooft@wur.nl

Abstract

Microbial specialised metabolism is full of valuable natural products that are applied clinically, agriculturally, and industrially. The genes that encode their biosynthesis are often physically clustered on the genome in biosynthetic gene clusters (BGCs). Many BGCs consist of multiple groups of co-evolving genes called sub-clusters that are responsible for the biosynthesis of a specific chemical moiety in a natural product. Sub-clusters therefore provide an important link between the structures of a natural product and its BGC, which can be leveraged for predicting natural product structures from sequence, as well as for linking chemical structures and metabolomics-derived mass features to BGCs.

While some initial computational methodologies have been devised for sub-cluster detection, current approaches are not scalable, have only been run on small and outdated datasets, or produce an impractically large number of possible sub-clusters to mine through.

Here, we constructed a scalable method for unsupervised sub-cluster detection, called iPRESTO, based on topic modelling and statistical analysis of co-occurrence patterns of enzyme-coding protein families. iPRESTO was used to mine sub-clusters across 150,000 prokaryotic BGCs from antiSMASH-DB. After annotating a fraction of the resulting sub-cluster families, we could predict a substructure for 16% of the antiSMASH-DB BGCs. Additionally, our method was able to confirm 83% of the experimentally characterised sub-clusters in MIBiG reference BGCs. Based on iPRESTO-detected sub-clusters, we could correctly identify the BGCs for xenorhabdin and salbostatin biosynthesis (which had not yet been annotated in BGC databases), as well as propose a candidate BGC for akashin biosynthesis. Additionally, we show for a collection of 145 actinobacteria how substructures can aid in linking BGCs to molecules by correlating iPRESTO-detected sub-clusters to MS/MS-derived Mass2Motifs substructure patterns.

This work paves the way for deeper functional and structural annotation of microbial BGCs by improved linking of orphan molecules to their cognate gene clusters, thus facilitating accelerated natural product discovery.

Author summary

In this work, we introduce iPRESTO, a tool for scalable unsupervised sub-cluster detection in biosynthetic gene clusters. This detection is important because these biosynthetic hotspots encode many products useful for humanity, such as antibiotics, antitumor agents, or herbicides. Recent technological developments have made identification of biosynthetic loci in genomes straightforward. Yet, methods to connect these inferred biosynthetic genes to the final chemical structures of their cognate metabolites are largely lacking. Being able to reliably predict parts of the final product would constitute a real step forward in natural product genome mining. Therefore, we focussed on constructing a tool to systematically detect and annotate small regions called sub-clusters, which code for the biosynthesis of substructures in the final product, across all genomically inferred biosynthetic diversity. iPRESTO makes it possible to query unknown biosynthetic regions and infer which substructures are present in their metabolic products. This will facilitate more effective prioritization of chemical novelty, as well as linking activities from bioassays and microbiome-associated phenotypes to the metabolites responsible for them.

Introduction

A considerable part of bacterial metabolism is dedicated to the biosynthesis of specialised metabolites. These natural products (NPs) have many uses as pharmaceuticals, crop protection agents, and ingredients for foods and cosmetics [1, 2]. NPs consist of a spectrum of different chemical classes, which are often highly complex in structure [3]. Intriguingly, the genes necessary for the biosynthesis of NPs cluster together physically in biosynthetic gene clusters (BGCs) [4]. The search and discovery of new BGCs accelerates identification of new NPs, which is especially important in the field of antibiotics, as antibiotic-resistant bacteria are becoming increasingly prevalent [5].

Due to the growing availability of genomic data, genome mining approaches have become more and more useful for NP discovery. Currently, multiple algorithms exist that mine bacterial genomes for putative BGCs, such as antiSMASH, ClusterFinder and PRISM [6-8]. These methods have provided a better understanding of BGC diversity and the evolutionary mechanisms that govern BGC diversity.

Many classes of BGCs display a modular architecture [4]. As such, a BGC can be divided into multiple modules or sub-clusters, where each sub-cluster is a group of co-evolving genes responsible for the biosynthesis of a specific chemical moiety in the NP [4, 9, 10]. Sub-clusters therefore provide a direct link between the substructures of an NP and its BGC. This makes information about sub-clusters and the substructures they synthesise highly valuable for genome-based structure prediction, which would be a great asset for tools like antiSMASH. Apart from enhancing structural predictions for existing BGC classes, sub-cluster knowledge would facilitate predicting novel (partial) structures of currently unclassified BGCs, such as the thousands of unclassified BGCs with yet unknown products in the antiSMASH-DB [11].

Additionally, BGC modularity poses a great opportunity to connect metabolomics experiments to sub-cluster data. Chemical moieties identified from fragments in mass spectrometry (MS) data could be linked to sub-clusters responsible for their synthesis, as part of MS-guided genome mining strategies [10, 12, 13]. Recent advances in substructure modelling [14] may aid such co-occurrence-based metabologenomic approaches [15] by automating the identification of substructures from MS/MS data.

Recently, Del Carratore et al. [10] introduced an initial method for the detection of sub-clusters in BGCs. By constructing Clusters of Orthologous Groups (COGs) and by using a statistical approach to group co-occurring COGs in sub-clusters, they were able to detect several experimentally characterised sub-clusters, as well as to discover novel ones. However, COG construction is not very scalable due to the all-vs-all BLAST calculation required. As a result, their analysis was performed on a relatively small dataset that is by now almost a decade old, and the chosen approach is hard to scale up to the massive amounts of genomic data that have become available in recent years. Additionally, the proposed statistical approach greatly overestimates the numbers of sub-clusters. This is due to the presence of redundant BGCs, which leads to artificial sub-clusters spanning entire BGCs, and caused by the inherently nested structure of the sub-clusters, where smaller, less specific sub-clusters are contained in larger, more specific sub-clusters.

Here, we propose an improved scalable method for unsupervised sub-cluster detection which we called the integrated Prediction and Rigorous Exploration of biosynthetic Sub-clusters Tool (iPRESTO). iPRESTO is scalable to large datasets and takes phylogenetic bias into account by filtering the input in a more advanced way. To detect sub-clusters, iPRESTO uses a statistical approach (PRESTO-STAT) as well as a topic modelling algorithm (PRESTO-TOP). As a data source, we used the antiSMASH-DB, which is one of the largest collections of BGCs that currently exists, and which has been scrutinized for underlying genome assembly quality [11]; it contains over 150,000 BGCs from almost 25,000 bacterial species selected to reduce taxonomical bias. These numbers represent a considerable improvement in comparison with the previous method as it contains over ten times as many BGCs, while being less redundant. After applying iPRESTO on this large collection of BGCs, we were able to annotate 45 sub-cluster motifs based on occurrences in known BGCs from the MIBiG reference BGC database [16]. Using these annotated sub-cluster motifs, we zoomed in on relevant sub-clusters, and showed direct usefulness of our method by correctly predicting the BGCs for xenorhabdin and salbostatin biosynthesis (which have been published but were missing from BGC databases) and identifying a candidate BGC for akashin biosynthesis. Finally, as a starting point for the automated connection of BGCs to their NPs, we were able to systematically link sub-clusters to substructures by using a metabologenomic correlation method in a paired-genome-metabolome dataset of 145 actinobacteria.

Results & Discussion

Overview of iPRESTO

iPRESTO prepares each BGC for sub-cluster detection by tokenising each gene in a BGC as a combination of Pfam domains (**Fig 1** and S1 Fig). If a pair of proteins share the same Pfam domains, this provides an

effective indication of (at least distant) sequence similarity, while Pfam detection is highly scalable. As Pfams are quite broad sequence models (which would be a major disadvantage compared to using COGs), we increased the resolution by splitting the 112 most abundant biosynthetic Pfams into a number of subPfams, akin to the implementation in BiG-SLICE [17]. Each subPfams constitutes a narrower domain model that covers a subset of a Pfam's sequence space. We only considered biosynthetic domains (see Methods) to limit the search space and focus solely on finding biosynthetic sub-clusters. With a graph-based filtering step, redundant BGCs are removed, after which iPRESTO detects sub-clusters using PRESTO-STAT and PRESTO-TOP. PRESTO-STAT is based on the previously published statistical method, which we expanded by partly removing nested sub-clusters, collapsing similar sub-clusters into families, and joining similar families into clans.

Fig 1. Outline of the iPRESTO workflow for the detection of sub-clusters. All genes in BGCs are converted into strings of Pfam domains, after which redundant BGCs are filtered out based on an Adjacency Index of domains. Sub-clusters are detected using two methods: PRESTO-TOP (TOP) and PRESTO-STAT (STAT). BGCs from the MIBiG database are used to annotate putative sub-clusters with sub-structures. These annotations are used to predict sub-structures in unknown BGCs.

To enrich the discovery of sub-clusters with a method that does not produce nested sub-clusters, we introduce PRESTO-TOP as a novel approach for sub-cluster detection. PRESTO-TOP is built on Latent Dirichlet Allocation (LDA), which is used to model topics in text documents. LDA has already been used successfully in genome and metabolome data analysis before [14, 18]. In the case of PRESTO-TOP, a text document is a BGC, a word is a gene represented as a domain combination, and a topic can be thought of as a sub-cluster motif. This highlights the use of PRESTO-TOP for sub-cluster detection, as we assume that a BGC is a combination of multiple different sub-clusters, which consist of co-evolving genes that co-occur in multiple BGCs. Another benefit of PRESTO-TOP is that a topic or sub-cluster motif will usually consist of a set of core genes that encode the enzymes to synthesise the base of a substructure, while various combinations of additional modifying genes can be found in PRESTO-STAT-detected (nested) sub-clusters. In this way, the two iPRESTO methods can jointly capture substructure diversity, by identifying the sub-cluster cores as well as their variants.

The resulting sub-clusters of both methods can be annotated with substructures and subsequently be used to predict sub-structures in BGCs. iPRESTO is readily usable for anyone who wants to detect sub-clusters in their own datasets, both by creating new sub-cluster models and by querying BGCs to the collection of sub-clusters we detected in this study. iPRESTO can handle large amounts of BGCs: tokenising and reducing redundancy in the 150,000 BGCs in the antiSMASH-DB dataset took around 48 hours each using 32 CPU cores on an Intel Xeon CPU E5-2670 v3. Detecting sub-clusters with PRESTO-STAT and PRESTO-TOP completed in 24 and 8 hours, respectively. iPRESTO can query around 20 BGCs per minute to the sub-clusters detected in this study including the tokenisation steps. iPRESTO also contains a visualisation module to visualise the results of querying a BGC to PRESTO-STAT or PRESTO-TOP output (see S2 Fig for an example of querying the rifamycin BGC).

PRESTO-STAT improves comprehensibility of existing statistical method

We applied iPRESTO to the antiSMASH-DB v2 dataset, which contained, after pre-processing, 60,028 BGCs with 10,539 domain combinations (Table A in S1 Text). Using the PRESTO-STAT method, we found 108,085 sub-clusters in the dataset. Over 80% of the statistical sub-clusters contain fewer than ten genes, and 17% of the sub-clusters occur in more than 10 BGCs (S3 Fig). When comparing PRESTO-STAT with the previous version of the method by Del Carratore et al. [10], we observed that PRESTO-STAT produces on average roughly two sub-clusters per BGC, while the previous method resulted in roughly fourteen sub-clusters per BGC. This indicates that we end up with fewer nested sub-cluster structures, which is most likely due to our extended redundancy filtering that removed almost half of the dataset (Table A in S1 Text). Even so, nested structures are still very apparent in our results (S2 Fig). For example, thousands of BGCs have more than 30 sub-clusters, many of which overlap with one another (S4A Fig). Not only do the nested structures inflate the results, but they also have the additional disadvantage that their presence makes it harder to connect BGCs with similar yet distinct sub-clusters.

To facilitate the sub-cluster analysis, we connected related sub-clusters by clustering the statistical sub-clusters into 10,000 sub-cluster families (SCFs) and the SCFs into 2,000 sub-cluster clans (SCCs). We used K-means clustering and represented the statistical sub-clusters as a presence/absence matrix of the tokenised genes. Although some SCCs grouped seemingly unrelated sub-clusters together that share only one gene (based on having the same Pfam domain content), most SCCs (81%) provided groups of related sub-clusters, sharing at least three genes.

Apart from the nested structures, the statistical method produces many sub-clusters of which only a fraction probably provides meaningful information. This is illustrated by the fact that the PRESTO-STAT results can be very noisy: in a group of BGCs sharing multiple sub-clusters, all combinations of these shared sub-clusters could form new sub-clusters, which happens frequently (S2 Fig). Additionally, it is rather difficult to query a BGC using the statistical sub-clusters while allowing inexact matching, as this would quickly become very time consuming.

PRESTO-TOP identifies characterised and novel sub-clusters

The drawbacks of PRESTO-STAT present a clear reason as to why we chose to also develop PRESTO-TOP, which can find multiple sub-clusters in a BGC and is able to capture sub-cluster diversity within sub-cluster motifs. Furthermore, LDA, upon which PRESTO-TOP is built, allows for a scalable way to build and query sub-cluster motifs.

We used PRESTO-TOP to train and query a model on the antiSMASH-DB dataset with 1,000 sub-cluster motifs. In the Methods section, we provide information on (hyper)parameters used and reasoning for the chosen settings. Over 80% of the BGCs in the dataset contained at least one sub-cluster motif (S4B Fig). To assess the quality of the sub-cluster motifs, we visualised all sub-clusters individually, where each sub-cluster is a group of genes matching against a sub-cluster motif (Fig 2A). For a sub-cluster to be interesting, we would expect its size to be between 2-12 genes, as experimentally characterised sub-clusters fall in this range [19]. Upon checking our results, most sub-clusters that were present across a considerable number of BGCs were within this expected size range (Fig 2A), while some sub-clusters were uninformative as they encompass (nearly) entire BGCs (Fig 2B). To validate the sub-cluster motifs, we assessed whether we could detect a set of 109 experimentally verified sub-clusters, which are stored in the SubClusterBlast module within the antiSMASH framework. The sub-cluster motifs from PRESTO-TOP matched to 91 (83%) validated sub-clusters, where the methoxymalonate and AHBA sub-clusters of macbecin are shown as examples (**Fig** C). Additionally, PRESTO-STAT was able to detect 78 of the validated sub-clusters, of which 75 overlap with the sub-cluster motifs (S5 Fig). In general, we see that PRESTO-TOP generates a more restricted amount of sub-cluster data, which might contain less meaningful sub-clusters compared to PRESTO-STAT in absolute numbers but has a considerably higher ratio of valid sub-cluster information.

Fig 2. BGC length versus sub-cluster length. (a) Scatterplot of the length of each BGC (number of non-empty genes) from the antiSMASH-DB dataset versus the length of a match to a topic or sub-cluster motif, representing a sub-cluster. The colour of each dot indicates how many times a BGC with a certain length contains a sub-cluster with a certain length. (b) BGC for sipanmycin where the identified sub-cluster encompasses the entire BGC, demonstrating an uninformative result. (c) BGC for macbecin where the two characterised sub-clusters for AHBA (red) and methoxymalonyl (blue) are highlighted in the structure of macbecin [20]. Sub-clusters from (b) and (c) are linked to their corresponding location in (a).

Our results provide clear examples of sub-cluster motifs that capture sub-cluster variety, by containing a set of core genes responsible for synthesising the base of a substructure, and a set of modifying genes that may not be present in all sub-clusters. For example, a motif like the sugar-related sub-cluster motif 680 is present in 134 MIBiG BGCs that represent different biosynthetic classes, such as different types of polyketide synthases and nonribosomal peptide synthetases. This motif codes for the biosynthesis of different (di)deoxy-sugars that are sometimes modified with amino or methyl-amino groups. However, for some sub-cluster motifs, the biosynthetic context had an impact on shaping the motif. The sugar-related sub-cluster motif 207, for example, contains several indolocarbazole biosynthesis genes as some MIBiG BGCs matching to this motif encode the production of indolocarbazoles, and some of the indolocarbazole-related genes ended up in this motif as weak features.

Exploring the sub-cluster motifs

Among the 90 identified characterised sub-clusters from the antiSMASH SubClusterBlast module, we could readily annotate 23 sub-cluster motifs covering around 4,000 of the PRESTO-TOP-detected sub-clusters. To extend on the sub-cluster knowledge stored in the SubClusterBlast module, we annotated another 22 PRESTO-TOP-detected sub-cluster motifs for which sub-cluster instances were found inside MIBiG BGCs. Together, these 45 annotations constitute 24 different types of substructures at different levels of detail and allow us to explore the discovered sub-clusters more deeply (Fig 3 and S1 File). In the non-redundant antiSMASH-DB dataset, around 9,500 (16%) putative BGCs contain at least one of these annotated sub-cluster motifs. Through iPRESTO, we now gained relevant knowledge about these putative BGCs that we can use to predict part of the structures of the products they encode.

Fig 3. Sub-cluster motif annotations. The pie chart visualises the annotations for the 45 sub-cluster motifs divided into general substructure groups, where an example substructure is shown for several groups. Additionally, examples of eight of the substructures are shown in the structures of apoptolidin, platencin, fluvirucin b2 and pyralomicin 1a, where the colours of the substructures correspond to the sub-cluster motif annotations in the pie chart. For these four metabolites, their respective BGCs are shown where the sub-cluster motifs are highlighted in the same colour as the substructures they encode.

On average, an annotated sub-cluster motif occurs in 239 non-redundant BGCs, ranging from 19 BGCs for sub-cluster motif 190, to 873 BGCs for sub-cluster motif 220, which encode the biosynthesis of caprazol and dihydroxybenzoic acid moieties, respectively (S6 Fig). Some of the annotated sub-cluster motifs are mainly present in one BGC class, while others occur in diverse BGC classes (S6 and S7 Figs). An example of the latter is sub-cluster motif 773, which occurs in 153 BGCs mostly encoding nonribosomal peptide synthetases and type I polyketide synthases. This sub-cluster motif encodes the production of a 3-amino-2-methylpropionyl starter unit that appears in the known gene cluster BGC0001597 (fluvirucin b2) (Fig 3). Interestingly, the motif also occurs in some BGCs of the class "Other", meaning they cannot be classified by antiSMASH, like two BGCs from *Amycolatopsis alba* DSM 44262 (NZ_KB913032.1.cluster021; AMYAL_RS0129245 - AMYAL_RS0129610) and *Bradyrhizobium sp. Ec3.3* (NZ_AXAS01000001.cluster006; YUU_RS0100020 - YUU_RS49645). This does not only provide interesting leads for these BGCs with previously unknown structural predictions, but it also adds to their validity. In total, 6.5% of the 10,000 "Other" class BGCs in the antiSMASH-DB contain one of the annotated sub-cluster motifs.

iPRESTO can identify BGCs of orphan metabolites through sub-cluster presence

Information about the sub-clusters present in a BGC is not only useful to predict the product of a BGC, but it could also be used as a tool to identify BGCs for 'orphan' known metabolites. To demonstrate this, we searched NPAtlas [21] with substructures that are encoded by our annotated sub-cluster motifs and looked for metabolites without a MIBiG BGC that are found in one of the strains in the antiSMASH-DB dataset. We first searched for metabolites that contain the dithiolopyrrolone substructure for which the biosynthesis is encoded by sub-cluster motif 517, annotated as such based on the MIBiG BGCs encoding thiomarinol, holomycin and thiolutin [22-24]. In doing so, we found xenorhabdins 1-6, produced by many *Xenorhabdus* strains that are also present in the antiSMASH-DB [25]. By searching for BGCs in those strains that contain a match to the dithiolopyrrolone sub-cluster motif, we found 12 *Xenorhabdus* strains that contain such a BGC (Fig 4). In one of those strains, *X. doucetiae*, the BGC for xenorhabdin biosynthesis has recently been described, corroborating that we accurately identified BGCs for xenorhabdin biosynthesis based on iPRESTO-detected sub-clusters [26]. Next, we searched NPAtlas for metabolites with the valienol moiety present in validamycin and pyralomicins, which is encoded by sub-cluster motif 940 [27, 28]. As a result, we found salbostatin, which is produced by *Streptomyces albus* ATCC 21838 in our dataset [29]. By investigating BGCs in that strain, we identified a BGC that contains sub-cluster motif 940 and should therefore be responsible for salbostatin biosynthesis (Fig 4). Indeed, it turned out that this BGC has already been described in 2008 to encode the production of salbostatin [30], but it has been lacking from the MIBiG database [16]. This valienol sub-cluster motif encoding C7-cyclitol-like substructures is an interesting example of a sub-cluster motif that can be found in different biosynthetic contexts, *i.e.*, PKS-NRPS-like pyralomicins and different kinds of saccharides like validomycin and salbostatin. This analysis highlights that iPRESTO allows identifying correct links between BGCs and molecules that are published but were yet missing in public BGC databases (and which can thus be added to these resources).

Fig 4. Connecting non-MIBiG BGCs to their metabolic products through iPRESTO-detected sub-clusters. (a) Phylogenetic tree made with CORASON of 12 *Xenorhabdus* BGCs and 3 MIBiG BGCs, that contain an iPRESTO-predicted sub-cluster for dithiolopyrrolone biosynthesis [31]. The A-domain containing gene of NZ_FO704550.1.cluster001 was used as query for CORASON. Structures of thiomarinol (1), thiolutin (2) and holomycin (3) are linked to their MIBiG BGCs. Xenorhabdins (4-9) are encoded by *X. doucetiae* str. FRM16 as indicated by the asterisk, while we infer based on sub-cluster presence that the other *Xenorhabdus* BGCs are also responsible for xenorhabdin biosynthesis. (b) Phylogenetic tree made with CORASON NZ_CP010519.1.cluster004 from *S. albus* ATCC 21838 and 4 MIBiG BGCs, that contain an iPRESTO-predicted sub-cluster for C7 cyclitol biosynthesis. The predicted 2-epi-5-epi-valiolone synthase from NZ_CP10519.1.cluster004 was used as query for CORASON. Structures of validomycin A (10) and pyralomycin 1A (11) are linked to their MIBiG BGCs. Salbostatin (12) is encoded by *S. albus* ATCC 21838 as indicated by the hash symbol.

By searching in NPAtlas for chlorinated indoles, we found the orphan metabolites akashin A-C produced by the diazaquinomycins producer *Streptomyces* sp. F001 [32]. The BGC of akashins has not been described before in literature. As this strain was not present in the antiSMASH-DB, we ran antiSMASH 6 on the genome of this strain and used iPRESTO to infer sub-clusters in the predicted BGCs. As akashins have chlorinated-indole moieties and are glycosylated, we sought for such sub-cluster motifs in the BGCs of *S.* sp. F001. Interestingly, we identified the genomic region in QZWF01000007.1.region003 (StrepF001_25985 - StrepF001_26130) directly upstream of the diazaquinomycin BGC, based on the presence of sub-cluster motifs 194, 607 and 680 that were annotated as methylaminosugar, halogenated aromatic ring, and (amino)deoxysugar, respectively (Fig 5). The formation of the indigo-derived backbone of akashins could potentially be formed by the two p450 enzymes, akin to CYP102G4, a recently described p450 enzyme from *S. cattleya* [33]. This p450 enzyme can catalyse the reaction from indole to 3-hydroxyindole after which spontaneous oxidation forms indigo. CYP102G4 was even shown to accept chloro-indole as substrate, in the case that chlorination occurs before indole formation in akashin biosynthesis. This shows that iPRESTO can aid in generating meaningful hypotheses about the biosynthesis of orphan metabolites.

Fig 5. Putative BGC for akashin A biosynthesis. The antiSMASH-predicted BGC QZWF01000007.1.region003 is shown (StrepF001_26130-StrepF001_26145), which is hypothetically responsible for akashin A biosynthesis in *S.* sp. F001. Genes are coloured by their iPRESTO-predicted sub-clusters or predicted function based on Pfam domains.

Correlation analysis in substructure-based integrative omics mining

To automatically link unknown molecules to BGCs at a larger scale, correlating substructures predicted from metabolomics data to sub-clusters from genome data would potentially be of great added value [12, 13]. To test such an approach, we used a previously defined correlation score which assumes that a BGC is needed to synthesise a product, but that a BGC may be cryptic and not synthesise anything [15]. Ernst et al. [34] used the MS2LDA tool to discover substructure mass patterns, called Mass2Motifs, from metabolomics data of 145 *Salinispora* and *Streptomyces* species for all of which (except one) genomic data and BGC predictions are also available (the 'Streptomyces/Salinispora dataset') [14]. To identify sub-clusters in these, we used iPRESTO to query all Streptomyces/Salinispora BGCs on the sub-cluster motifs and sub-cluster clans (SCCs) of the antiSMASH-DB dataset. For each of the 107,590 pairs of Mass2Motifs and sub-cluster motifs, we used the correlation score from Doroghazi et al. [15] to calculate how frequently they co-occur across the Streptomyces/Salinispora strains, while we did the same for the 122,404 pairs of Mass2Motifs and SCCs (S8 Fig). To prioritise interesting substructure-sub-cluster pairs, we performed permutation tests for all pairs to assess the likelihood of a high scoring pair arising by chance. This was especially needed as the Streptomyces/Salinispora dataset includes highly related strains, in which many BGCs and compounds are shared. Abundant sub-clusters and substructures therefore get high correlation scores by default. Permutation testing resulted in 3,230 and 1,939 'significant' pairs of Mass2Motifs and sub-cluster motifs or SCCs, respectively (S8 Fig). As an example of how such an approach connects substructure information inferred from genome and metabolome mining, we identified 5 high correlation scores with low p-values between two staurosporine-related mass2motifs and both sub-cluster motifs and SCCs constituting the amino-sugar moiety of staurosporine (Fig 6). Since currently only a fraction of the Mass2Motifs, sub-cluster motifs and SCCs are annotated, our analysis serves as an illustration of how such an approach could help to link metabolome and genome data in the future.

This correlation method generally results in a lot of noise, as sub-clusters and substructures that occur in a shared subset of strains will all correlate to each other. Such co-correlating structures make the identification of the actual correlating pair therefore difficult, especially with limited annotations. Identifying clusters of co-correlating pairs could therefore provide a way to make the interpretation of this analysis easier. Additionally, the correlation analysis is not perfect in our case, as multiple different sub-clusters are often responsible for synthesising the same kind of substructure. For example, we identified multiple sub-cluster motifs that can encode methylated aminosugars, while only one mass2motif is annotated as a methylated aminosugar. In future approaches, such mismatches between genome and metabolome could be overcome by finding ways to group sub-cluster motifs together that encode similar structures before running such metabologenomic correlation analyses. Combining such solutions with the integration of more diverse species, new annotations, and improved correlation scoring methods like the one developed in Hjörleifsson Eldjárn et al. [35] would improve such analyses drastically. Furthermore, we expect that combining co-occurrence based scores (such as standardized Metcalf) with feature-based scores, such as NPClassScore [36], and the here developed iPRESTO, will further help to prioritize plausible BGC-MS/MS spectral links [12, 13]. Indeed, we expect that tools like iPRESTO could in the future be built into frameworks like NPLinker [35]. As our current contribution represents a first step in linking substructure- and sub-cluster models with rather limited (annotated) information, we expect that analyses like these will have great impact in the future to facilitate metabologenomics experiments that use integrative omics mining.

Fig 6. Metabologenomic correlation scores between sub-clusters and mass2motifs. Stacked histogram of the correlation scores across the Streptomyces/Salinispora strains between the mass2motifs paired with either the SCCs or sub-cluster motifs with a p-value below 0.1. Highlighted with their scores are the pairs mass2motif_108 with SSC_452, SSC_1010, sub-cluster_motif_207 and sub-cluster_motif_680, and the pair mass2motif_8 with SSC_452. The aforementioned sub-cluster motifs (blue) and SCCs (brown) are responsible for sugar synthesis in staurosporine, while both mass2motifs (red) are staurosporine related.

Conclusion and future perspectives

This study introduces the iPRESTO concept and makes it available as a command line tool. We plan to include iPRESTO in one of the future releases of antiSMASH, so the collection of sub-clusters we generated in this study can be used to detect and visualize them in antiSMASH-predicted BGCs. We anticipate that this will enhance the current scope of sub-cluster detection, as antiSMASH's current sub-cluster predictor SubClusterBlast offers a limited amount of sub-cluster data, whereas our sub-cluster set will allow making more connections between predicted BGCs and MIBiG reference BGCs. This will accelerate NP discovery by linking structural information from genome and metabolome data.

Due to the above discussed limitations of PRESTO-STAT, we plan to use PRESTO-TOP as the main method for sub-cluster detection in the antiSMASH implementation, as it also captures sub-cluster variety in the sub-cluster motifs and yet can be used easily to query BGCs for sub-cluster motifs. PRESTO-STAT could still be used to identify the sub-cluster boundaries better, by for example linking groups of related PRESTO-STAT sub-clusters to 'parent' PRESTO-TOP sub-cluster motifs, and by using the PRESTO-STAT modules to more specifically identify the sub-cluster variant found in a given BGC. The drawback of the statistical method that it produces highly nested and variable sub-clusters could as such be used as a strength. A way to further improve PRESTO-TOP would be to apply PRESTO-TOP in a semi-supervised manner, which constitutes a major potential benefit of this approach. Before training an LDA model, certain motifs could be seeded beforehand, which allows accurate sub-cluster motifs to be reused in new analyses, analogous to the metabolomics database MotifDB, in which annotated Mass2Motifs are stored in MotifSets [37]. Such semi-supervised approaches would allow for noise to be eliminated from sub-cluster motifs and sub-cluster motifs to be finetuned. Another way to reduce noise and to identify the more robust sub-cluster motifs would be to train multiple PRESTO-TOP models on the same dataset. Sub-cluster motifs that are found in every PRESTO-TOP model would constitute conclusive sub-cluster motifs, whereas sub-cluster motifs that are identified in most cases would still be considered reasonably accurate. In this manner, noisy sub-cluster motifs that arise through chance would be filtered out, as they would only occur in one or a few of the many LDA models. Noisy genes in accurate sub-cluster motifs could be filtered out by taking intersects of multiple similar sub-cluster motifs. As another option, each BGC could be represented multiple times in training to increase the observations of less frequently occurring sub-clusters. This could lead to better

363 estimation of the sub-cluster motif distributions over the data and cause less erroneous mixed sub-cluster
364 motifs. We have attempted this for a small subset and noticed that the overlap with SubClusterBlast
365 increased slightly, making this an interesting avenue to continue PRESTO-TOP sub-cluster algorithmic
366 developments.

367 Using iPRESTO, in our current study we were able to characterise 45 different sub-cluster motifs present
368 in diverse BGC classes. The remaining 955 sub-cluster motifs remain largely unexplored, of which many
369 are likely to encode useful substructures. We expect that, in the future, more annotations will increase the
370 value of our results even more, which will be aided by the inclusion of updated (expanded) versions of the
371 MIBiG database. Using one of the characterised sub-cluster motifs, we showed a direct practical application
372 of our method by hypothesising a putative BGC for akashin A production. Additionally, we provided the
373 initial step for linking sub-clusters to substructures in a systematic way, which in the future could facilitate
374 the automated connection of BGCs to their NPs.

375 Methods

376 Data and Code availability

377 iPRESTO is available as a command-line tool at https://git.wageningenur.nl/bioinformatics/iPRESTO/. The
378 annotated sub-cluster motifs and other relevant data can be found at
379 https://doi.org/10.5281/zenodo.6953657. The following sections describe the most important steps of this
380 project, while the Supplementary methods in S1 Text provide more detailed explanations.

381 Data selection

382 The antiSMASH-DB dataset consisted of three data sources: the MIBiG database, the
383 Streptomyces/Salinispora dataset and the antiSMASH-DB. Version 1.4 of the MIBiG database was used
384 which contains 1,819 BGCs (https://dl.secondarymetabolites.org/mibig/mibig_gbk_1.4.tar.gz). The
385 Streptomyces/Salinispora dataset consists of 5,927 BGCs that originate from the 146 *Streptomyces* and
386 *Salinispora* strains investigated by Crüsemann et al. [38]. antiSMASH 3.0 was used for the detection of
387 BGCs in the Streptomyces/Salinispora dataset. The antiSMASH-DB version 2 is comprised of 152,122 BGCs
388 detected with antiSMASH 4.0, where we included BGCs from draft genomes (Table A in S1 Text;
389 https://dl.secondarymetabolites.org/database/2.0/asdb_20180828_all_results.tar.xz). BGCs were
390 discarded if they were flagged by antiSMASH as lying on a contig-edge, as these BGCs are probably
391 incomplete (fragmented) and less accurate. Additionally, BGC class information was included in the
392 analysis, by using the assigned antiSMASH biosynthetic classes.

393 Data pre-processing

394 BGCs were tokenised by converting each gene into a string of (sub)Pfam domains. To detect (sub)Pfams,
395 the HMMER3 tool hmmscan was used with a custom profile hidden Markov model (pHMM) database
396 consisting of Pfam database version 32.0, where 112 Pfams were replaced by corresponding subPfams [39,
397 40]. These 112 Pfams were selected as they are the most abundant biosynthetic Pfams in the antiSMASH-
398 DB (S2 File). To create subPfams, the multiple sequence alignment of a Pfam is split into clades, after
399 which a new pHMM is built for each clade, each of which constitutes a subPfam (S1A Fig and
400 https://github.com/satriaphd/build_subpfam).

401 Redundant BGCs were removed from the analysis using a similarity network of BGCs, where BGCs were
402 connected based on an Adjacency Index of domains higher than 0.95 or if BGCs were fully contained within
403 one another. From each maximal clique in the network, only the BGC with the most domains was chosen
404 to remain in the analysis (Table A in S1 Text and S9 Fig) [41]. After redundancy filtering, all non-
405 biosynthetic domains were removed from all BGCs. To select biosynthetic domains, EC-associated Pfams
406 were collected with ECDomainMiner, from which Pfams were selected if they occurred in pre-calculated
407 BGCs [42]. After manual curation, this resulted in a list of 1,839 biosynthetic Pfams (S3 File). Additionally,
408 Pfams that occurred less than three times in the dataset were removed as well as BGCs that contained less
409 than two non-empty genes (S4 File).

PRESTO-STAT

The statistical method for sub-cluster detection was re-implemented in Python based on Del Carratore et al. [10] with some alterations, resulting in PRESTO-STAT. Instead of representing genes as COGs as in the previous method, we represent each gene as a combination of its domains. First, all possible adjacency and co-localisation interactions between each pair of genes are counted. To assess whether an observed interaction between two genes occurs more than by random chance, one needs to distribute such a pair of genes randomly through the dataset and calculate the probability of the observed interaction. To reduce the computational burden of a permutation-based approach, for each pair of genes one gene is kept fixed while the other is being randomly distributed throughout the data. For an adjacency interaction this gives a hypergeometric equation describing all available positions of one gene while the other is fixed (Table B1 in S1 Text). This follows from the fact that there are three options for the position of gene B while keeping gene A fixed: not adjacent to gene A (B_1), adjacent to gene A (B_2), or adjacent to gene A on both sides (B_3). N_1, N_2 and N_3 represent all available positions in these three categories, while N_{tot} represents all positions and B_{tot} all occurrences of gene B. For a co-localisation interaction the same applies, except for the fact that gene B can be co-localised with n_{max} genes A, where n_{max} is the number of genes A co-localised with gene B (Table B2 in S1 Text). When n_{max} is large this becomes computationally hard, which is why we replaced duplicate genes with an empty gene (a dash) and placed one copy of the duplicate gene at the end of the cluster separated by an empty gene. This simplifies the equation as only two types of co-localisations need to be counted: co-localisation and no co-localisation (Table B3 in S1 Text). A p-value can be calculated by summing all probabilities in the hypergeometric distribution that correspond to several interactions higher or equal to the observed number of interactions. Or, to make it easier, by subtracting the sum of all possible interactions smaller than the observed interaction from one (Table B4 in S1 Text).

Calculating an interaction between each pair of genes results in two p-values, one coming from gene A and one coming from gene B. Only the largest p-value for both the co-localisation, and the adjacency interactions is considered, to be conservative. To control false discovery rate under dependency we used the Benjamini–Yekutieli method on both the co-localisation and adjacency p-values [43].

To group interacting pairs of genes into sub-clusters, undirected graphs are constructed, where each gene is a node. An edge is made between two genes if they have an adjacency or co-localisation p-value below a threshold of 0.1. All maximal cliques are selected as sub-clusters, while changing the threshold iteratively to all the p-values in the dataset smaller than the original threshold of 0.1. To reduce false positives, we removed putative sub-clusters if they contained fewer than three genes and if they only occurred in one BGC. Next, we grouped similar sub-clusters together using K-means clustering into sub-cluster families and sub-cluster clans and removed redundant sub-clusters (Supplementary methods in S1 Text) [44, 45].

PRESTO-TOP

PRESTO-TOP uses Latent Dirichlet Allocation (LDA) latent sub-cluster composition in BGCs [46]. LDA assumes a bag-of-words representation, where each BGC is depicted as a frequency vector of its domain combinations, not taking gene order into account. We used the multicore LDA implementation from Gensim, that makes use of online variational Bayes [47, 48]. In this implementation, an LDA model is trained by updating it with mini-batches from the data, which has low time and memory complexity. We chose the chunk size of each mini-batch to be 5% of the data with a minimum chunk size of 2,000, which is loosely based on testing different chunk sizes by Hoffman et al. [48]. We considered that using 500 iterations to train a model was enough after assessing that the log-likelihood converged sufficiently (S10 Fig). For the sake of computational resources, we did limited hyperparameter optimisation for the number of sub-cluster motifs (topics) N, α, and β. To test the performance of the different models, we considered the coherence score as measured with the u_mass method [49] and the overlap with validated sub-clusters from SubClusterBlast (Supplementary methods in S1 Text). Based on the coherence score of the different models, choosing 250 sub-cluster motifs seemed optimal (S11A Fig). However, upon manual inspection of some of the motifs, it turned out that many motifs are hard to annotate with a single substructure due to the presence of many noisy features. This is corroborated by the fact that choosing 250 sub-cluster motifs does not produce the highest overlap with SubClusterBlast (S11B Fig). Instead, the model with 1000 sub-

cluster motifs produced the highest overlap with SubClusterBlast while having a similar coherence score to the model with 250 motifs, which is why we chose 1000 sub-cluster motifs. We chose the default setting of a symmetric 1/N for hyperparameters α and β, as we could not find better SubClusterBlast overlap when setting α and β to symmetric, asymmetric, auto, or 1.

Each sub-cluster motif in an LDA model consists of a probability vector of domain combinations, representing the contribution of each domain combination to a sub-cluster motif. To filter out noise, we sorted this vector from high to low probability, summed the probabilities and included all domain combinations until 0.95 was reached. When a group of genes from a BGC match to a sub-cluster motif, each gene is assigned a feature probability describing how well it fits in the sub-cluster motif, for which we set a cut-off of 0.3. For a sub-cluster to be considered it needs to consist of more than one gene, for which we set a cut-off of 1.1 on the summed feature probabilities. Additionally, we calculated an overlap score for each match, which we computed by summing the domain combination probabilities from the sub-cluster motif present in the match [50]. We set a threshold of 0.15 on the overlap score, as this was the highest threshold that did not remove manually validated SubClusterBlast sub-clusters from the analysis.

Acknowledgements

We thank Dr Dick de Ridder and Dr Simon Rogers for useful comments and discussions.

Supporting information

S1 Text. Supplementary information for *iPRESTO: automated discovery of biosynthetic sub-clusters linked to specific natural product substructures.*

S1 Fig. Schematic depiction of BGC tokenisation. (A) subPfams are constructed for the 112 most frequent Pfam domains in the antiSMASH-DB by dividing the multiple sequence alignment of a Pfam into clades and converting each clade into a new pHMM. (B) The BGCs predicted by antiSMASH are tokenised by detecting (sub)Pfams in each gene, where non-biosynthetic Pfams are removed. After tokenising the BGCs, sub-cluster can be detected with the statistical method (Stat), where the tokenised genes are represented in their original order, or by LDA, which assumes a bag of words model where original gene order is not considered.

S2 Fig. Result of querying rifamycin (BGC0000373) to the PRESTO-TOP and PRESTO-STAT sub-clusters generated in this project. Only around 25% of the PRESTO-STAT sub-clusters are shown. Each gene is depicted as a token, where all (sub)Pfam domains are coloured. The visualisation of the BGC, the PRESTO-TOP and PRESTO-STAT output are separated by a dashed line, respectively. All PRESTO-STAT sub-clusters clearly exhibit a nested structure, where all combinations of genes in an actual sub-cluster are detected as individual sub-clusters. The PRESTO-STAT sub-clusters shown here are also examples of noisy sub-clusters comprised of combinations of genes from different actual sub-clusters, like detected PRESTO-STAT sub-clusters that are combinations of genes responsible for the biosynthesis of AHBA (green), sugars (blue) and the polyketide scaffold (purple).

S3 Fig. Information about the PRESTO-STAT sub-clusters. (A) The distribution of the number of genes per PRESTO-STAT sub-cluster in the antiSMASH-DB dataset. (B) The distribution of the log10 transformed PRESTO-STAT sub-cluster occurrences in the antiSMASH-DB dataset.

S4 Fig. Number of PRESTO-STAT and PRESTO-TOP sub-clusters per BGC. (A) Distribution of the log10 transformed number of PRESTO-STAT sub-clusters per BGC in the non-redundant antiSMASH-DB dataset, where the bin with the seemingly negative value represents BGCs without any PRESTO-STAT sub-cluster. (B) The number of topics or sub-cluster motifs per BGC in the non-redundant antiSMASH-DB dataset, not counting sub-clusters of length one as these are almost definitely noise (see Methods). (C) All BGCs with at least one annotated sub-cluster motif grouped by how many annotated sub-cluster motifs they have. In total there are 9,425 putative BGCs with at least one annotated sub-cluster motif, and 350 MIBiG BGCs.

S5 Fig. PRESTO-STAT and PRESTO-TOP overlap with validated sub-clusters from SubClusterBlast. Overlap between detected SubClusterBlast sub-clusters and output of both sub-cluster detection methods applied on the antiSMASH-DB dataset according to different overlap cut-offs. The overlap expresses the fraction of genes

from the original SubClusterBlast sub-cluster that is found in the iPRESTO-detected sub-cluster. We considered an overlap of 0.6 sufficient for having detected a sub-cluster (see Supplementary methods in S1 Text).

S6 Fig. Degrees (occurrences) of the annotated sub-cluster motifs within the antiSMASH-DB dataset (non-redundant).

S7 Fig. BGC class distribution across sub-cluster motifs. Relative abundance of antiSMASH classes when querying the non-redundant antiSMASH-DB dataset on the 45 annotated sub-cluster motifs. Matches of length 1 are ignored and hybrid class BGCs are counted for all classes they contain. RIPPs classes are grouped together.

S8 Fig. Correlation scores between Mass2Motifs and sub-clusters. (A) Correlation scores between Mass2Motifs and SCCs. (B) Correlation scores between Mass2Motifs and sub-cluster motifs. In both panels the significant pairs are highlighted.

S9 Fig. Graphical representation of graph-based filtering for the small dataset: MIBiG-and Streptomyces/Salinispora BGCs. Each node represents a BGC and an edge represents an adjacency index (AI) of 0.95 or higher. In blue are the BGCs chosen as representatives, while BGCs that are filtered out are shown in black. We show the small dataset here as it was difficult to visualize this process for the antiSMASH-DB dataset.

S10 Fig. LDA model convergence. Convergence of the log-likelihood of an LDA model with 1,000 topics/sub-cluster motifs trained on the non-redundant 60,028 BGCs from the antiSMASH-DB dataset, which also contains the Streptomyces/Salinispora dataset and the MIBiG database, using 2,000 iterations of chunk size 3,000. Log-likelihood based on 28 held out BGCs.

S11 Fig. Coherence scores and overlap with SubClusterBlast sub-clusters for different LDA models. (A) Coherence scores of different LDA models trained using PRESTO-TOP on the non-redundant antiSMASH-DB dataset with different number of topics. (B) Number of validated SubClusterBlast sub-clusters found with different LDA models trained using PRESTO-TOP on the non-redundant antiSMASH-DB dataset with different number of topics.

S1 File. Excel sheet containing the current information about the 45 annotated sub-cluster motifs.

S2 File. The 112 domains for which we created subPfams.

S3 File. The biosynthetic domains we considered in this study.

S4 File. All used domain-combinations present in the antiSMASH-DB dataset after filtering.

References

1. Dayan FE, Cantrell CL, Duke SO. Natural products in crop protection. Bioorganic & medicinal chemistry. 2009;17(12):4022-34.
2. Li JWH, Vederas JC. Drug Discovery and Natural Products: End of an Era or an Endless Frontier? Science. 2009;325(5937):161. doi: 10.1126/science.1168243.
3. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci U S A. 2017;114(22):5601-6. Epub 2017/05/04. doi: 10.1073/pnas.1614680114. PubMed PMID: 28461474; PubMed Central PMCID: PMCPMC5465889.
4. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput Biol. 2014;10(12):e1004016. Epub 2014/12/05. doi: 10.1371/journal.pcbi.1004016. PubMed PMID: 25474254; PubMed Central PMCID: PMCPMC4256081.
5. Chevrette MG, Currie CR. Emerging evolutionary paradigms in antibiotic discovery. J Ind Microbiol Biotechnol. 2018. Epub 2018/10/01. doi: 10.1007/s10295-018-2085-6. PubMed PMID: 30269177.
6. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014;158(2):412-21. Epub 2014/07/19. doi: 10.1016/j.cell.2014.06.034. PubMed PMID: 25036635; PubMed Central PMCID: PMCPMC4123684.
7. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 2021;49(W1):W29-W35. doi: 10.1093/nar/gkab335.
8. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nature Communications. 2020;11(1):6058. doi: 10.1038/s41467-020-19986-1.

9. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proceedings of the National Academy of Sciences. 2008;105(12):4601. doi: 10.1073/pnas.0709132105.
10. Del Carratore F, Zych K, Cummings M, Takano E, Medema MH, Breitling R. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. Communications Biology. 2019;2(1). doi: 10.1038/s42003-019-0333-6.
11. Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Research. 2020;49(D1):D639-D43. doi: 10.1093/nar/gkaa978.
12. Louwen JJ, Van Der Hooft JJ. Comprehensive large-scale integrative analysis of omics data to accelerate specialized metabolite discovery. Msystems. 2021;6(4):e00726-21.
13. van der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH. Linking genomics and metabolomics to chart specialized metabolic diversity. Chemical Society Reviews. 2020;49(11):3297-314. doi: 10.1039/D0CS00162G.
14. van der Hooft JJ, Wandy J, Barrett MP, Burgess KE, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. Proc Natl Acad Sci U S A. 2016;113(48):13738-43. Epub 2016/11/20. doi: 10.1073/pnas.1608041113. PubMed PMID: 27856765; PubMed Central PMCID: PMCPMC5137707.
15. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol. 2014;10(11):963-8. doi: 10.1038/nchembio.1659. PubMed PMID: 25262415; PubMed Central PMCID: PMCPMC4201863.
16. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research. 2019;48(D1):D454-D8. doi: 10.1093/nar/gkz882.
17. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience. 2021;10(1). doi: 10.1093/gigascience/giaa154.
18. Chen X, Hu X, Shen X, Rosen G, editors. Probabilistic topic modeling for genomic data interpretation. 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2010: IEEE.
19. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, Takano E, et al. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013;41(Web Server issue):W204-12. Epub 2013/06/06. doi: 10.1093/nar/gkt449. PubMed PMID: 23737449; PubMed Central PMCID: PMCPMC3692088.
20. Zhang M-Q, Gaisser S, Nur-E-Alam M, Sheehan LS, Vousden WA, Gaitatzis N, et al. Optimizing Natural Products by Biosynthetic Engineering: Discovery of Nonquinone Hsp90 Inhibitors. Journal of Medicinal Chemistry. 2008;51(18):5494-7. doi: 10.1021/jm8006068.
21. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Central Science. 2019;5(11):1824-33. doi: 10.1021/acscentsci.9b00806.
22. Li B, Walsh CT. Identification of the gene cluster for the dithiolopyrrolone antibiotic holomycin in *Streptomyces clavuligerus*. Proceedings of the National Academy of Sciences. 2010;107(46):19731-5. doi: doi:10.1073/pnas.1014140107.
23. Fukuda D, Haines AS, Song Z, Murphy AC, Hothersall J, Stephens ER, et al. A Natural Plasmid Uniquely Encodes Two Biosynthetic Pathways Creating a Potent Anti-MRSA Antibiotic. PLOS ONE. 2011;6(3):e18031. doi: 10.1371/journal.pone.0018031.
24. Huang S, Him Tong M, Qin Z, Deng Z, Deng H, Yu Y. Identification and characterization of the biosynthetic gene cluster of thiolutin, a tumor angiogenesis inhibitor, in Saccharothrix algeriensis NRRL B-24137. Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents). 2015;15(3):277-84.
25. McInerney BV, Gregson RP, Lacey MJ, Akhurst RJ, Lyons GR, Rhodes SH, et al. Biologically Active Metabolites from Xenorhabdus Spp., Part 1. Dithiolopyrrolone Derivatives with Antibiotic Activity. Journal of Natural Products. 1991;54(3):774-84. doi: 10.1021/np50075a005.
26. Bode E, Brachmann AO, Kegler C, Simsek R, Dauth C, Zhou Q, et al. Simple "On-Demand" Production of Bioactive Natural Products. ChemBioChem. 2015;16(7):1115-9. doi: https://doi.org/10.1002/cbic.201500094.
27. Bai L, Li L, Xu H, Minagawa K, Yu Y, Zhang Y, et al. Functional analysis of the validamycin biosynthetic gene cluster and engineered production of validoxylamine A. Chemistry & biology. 2006;13(4):387-97.
28. Flatt PM, Wu X, Perry S, Mahmud T. Genetic Insights into Pyralomicin Biosynthesis in Nonomuraea spiralis IMC A-0156. Journal of Natural Products. 2013;76(5):939-46. doi: 10.1021/np400159a.
29. Vértesy L, Fehlhaber H-W, Schulz A. The Trehalase Inhibitor Salbostatin, a Novel Metabolite from Streptomyces albus, ATCC21838. Angewandte Chemie International Edition in English. 1994;33(18):1844-6. doi: https://doi.org/10.1002/anie.199418441.
30. Choi WS, Wu X, Choeng Y-H, Mahmud T, Jeong BC, Lee SH, et al. Genetic organization of the putative salbostatin biosynthetic gene cluster including the 2-epi-5-epi-valiolone synthase gene in Streptomyces albus ATCC 21838. Applied Microbiology and Biotechnology. 2008;80(4):637-45. doi: 10.1007/s00253-008-1591-2.
31. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, et al. A computational framework to explore large-scale biosynthetic diversity. Nature Chemical Biology. 2020;16(1):60-8. doi: 10.1038/s41589-019-0400-9.
32. Braesel J, Clark CM, Kunstman KJ, Green SJ, Maienschein-Cline M, Murphy BT, et al. Genome Sequence of Marine-Derived *Streptomyces* sp. Strain F001, a Producer of Akashin A and Diazaquinomycins. Microbiology Resource Announcements. 2019;8(19):e00165-19. doi: doi:10.1128/MRA.00165-19.
33. Kim J, Lee P-g, Jung E-o, Kim B-G. In vitro characterization of CYP102G4 from Streptomyces cattleya: A self-sufficient P450 naturally producing indigo. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2018;1866(1):60-7. doi: https://doi.org/10.1016/j.bbapap.2017.08.002.

34. Ernst M, Kang KB, Caraballo-Rodríguez AM, Nothias L-F, Wandy J, Chen C, et al. MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites. 2019;9(7):144. PubMed PMID: doi:10.3390/metabo9070144.
35. Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLOS Computational Biology. 2021;17(5):e1008920. doi: 10.1371/journal.pcbi.1008920.
36. Louwen JJ, Medema MH, van der Hooft JJ. Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. 2022. doi: https://doi.org/10.21203/rs.3.rs-1391827/v2.
37. Rogers S, Ong CW, Wandy J, Ernst M, Ridder L, van der Hooft JJJ. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. Faraday Discussions. 2019;218(0):284-302. doi: 10.1039/C8FD00235E.
38. Crüsemann M, O'Neill EC, Larson CB, Melnik AV, Floros DJ, da Silva RR, et al. Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. J Nat Prod. 2017;80(3):588-97. doi: 10.1021/acs.jnatprod.6b00722. PubMed PMID: 28335604; PubMed Central PMCID: PMCPMC5367486.
39. Bateman A, Smart A, Luciani A, Salazar GA, Mistry J, Richardson LJ, et al. The Pfam protein families database in 2019. Nucleic Acids Research. 2018;47(D1):D427-D32. doi: 10.1093/nar/gky995.
40. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic acids research. 2013;41(12):e121-e.
41. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Commun ACM. 1973;16(9):575-7. doi: 10.1145/362342.362367.
42. Alborzi SZ, Devignes M-D, Ritchie DW. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics. 2017;18(1):107. doi: 10.1186/s12859-017-1519-x.
43. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The annals of statistics. 2001;29(4):1165-88.
44. Arthur D, Vassilvitskii S, editors. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; 2007: Society for Industrial and Applied Mathematics.
45. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
46. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.
47. Rehurek R, Sojka P, editors. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010: Citeseer.
48. Hoffman M, Bach FR, Blei DM, editors. Online learning for latent dirichlet allocation. advances in neural information processing systems; 2010.
49. Röder M, Both A, Hinneburg A, editors. Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining; 2015.
50. van der Hooft JJJ, Wandy J, Young F, Padmanabhan S, Gerasimidis K, Burgess KEV, et al. Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. Anal Chem. 2017;89(14):7569-77. Epub 2017/06/18. doi: 10.1021/acs.analchem.7b01391. PubMed PMID: 28621528; PubMed Central PMCID: PMCPMC5524435.

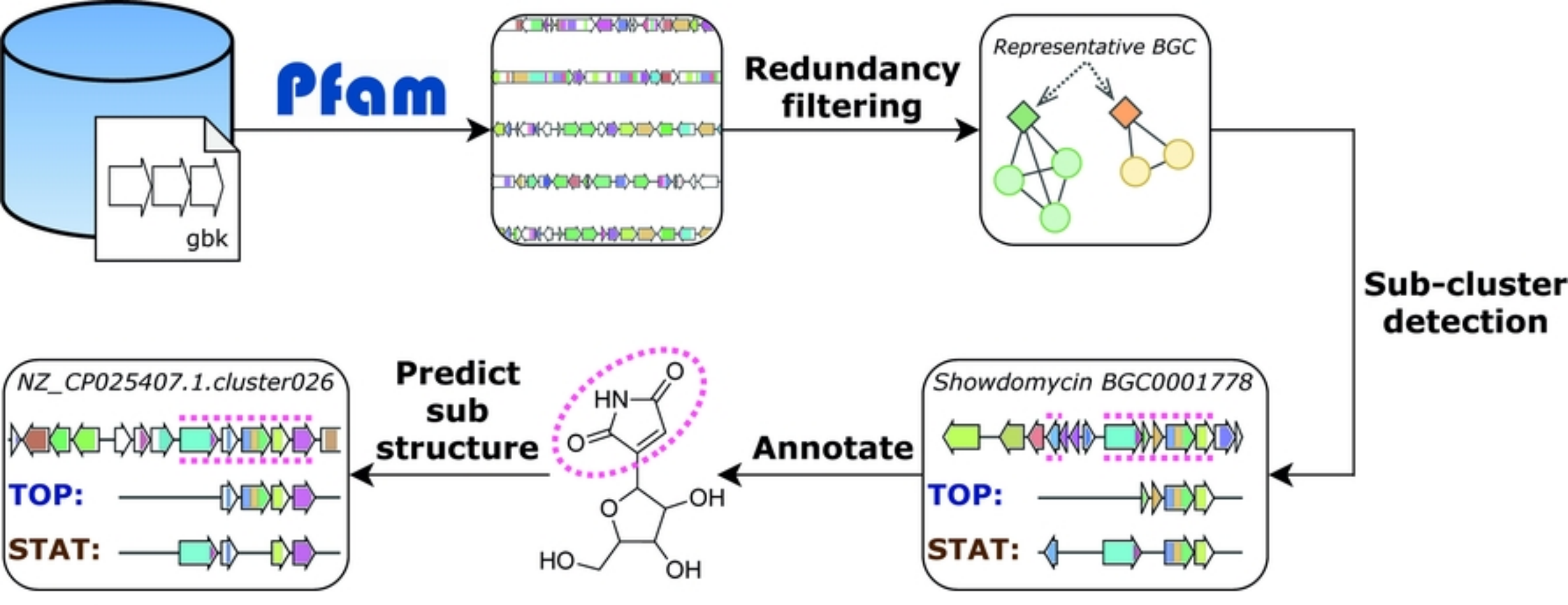

gbk
Pfam
Redundancy filtering
Representative BGC
Sub-cluster detection
Showdomycin BGC0001778
TOP:
STAT:
Annotate
HN
O
O
OH
O
HO
OH
Predict sub structure
NZ_CP025407.1.cluster026
TOP:
STAT:

Fig 1

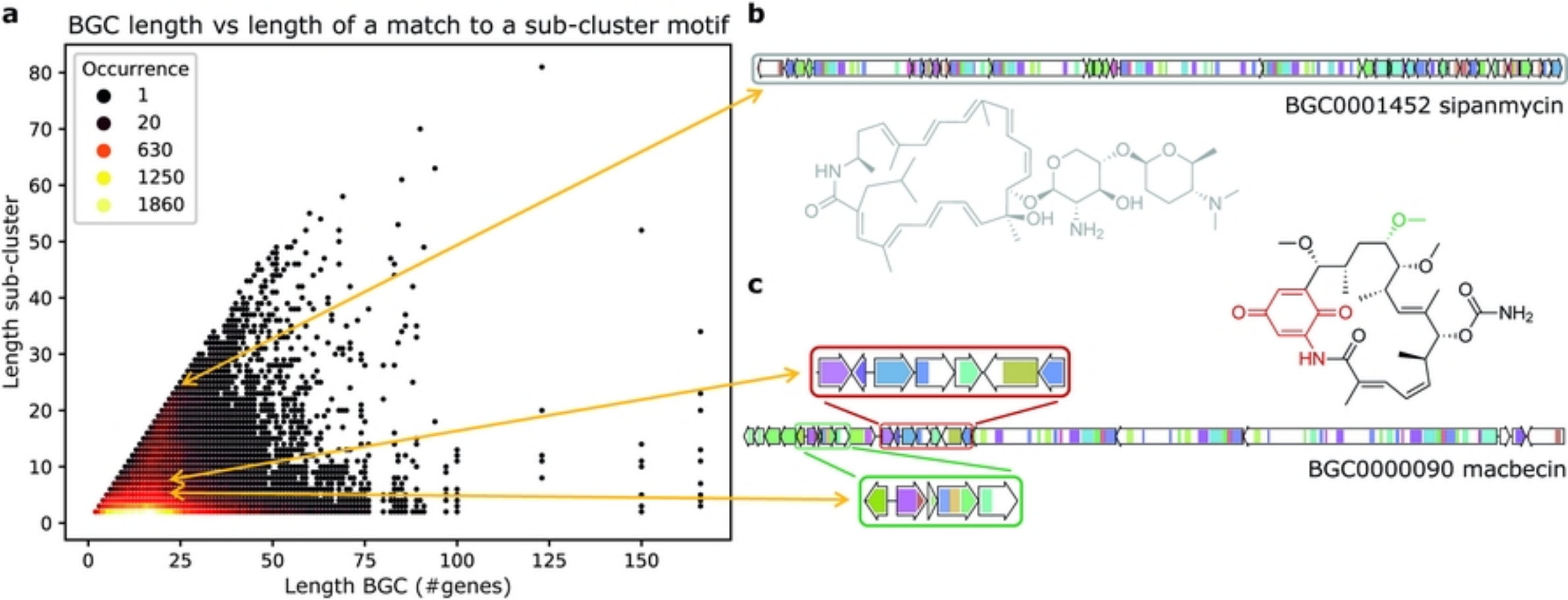

a
BGC length vs length of a match to a sub-cluster motif
Occurrence
1
20
630
1250
1860
Length sub-cluster
Length BGC (#genes)
0
10
20
30
40
50
60
70
80
0
25
50
75
100
125
150
b
BGC0001452 sipanmycin
c
BGC0000090 macbecin

Fig 2

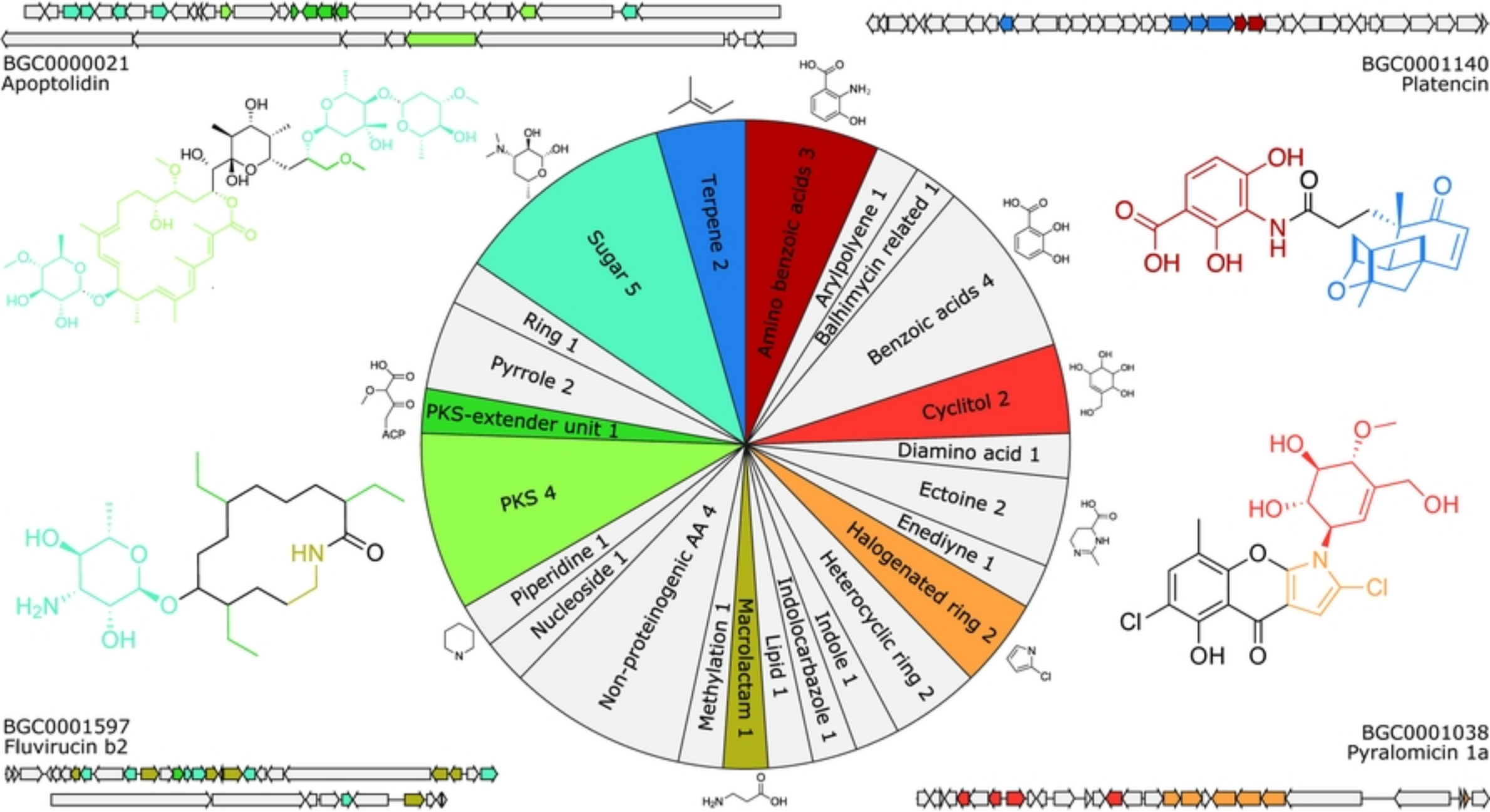

BGC0000021
Apoptolidin
BGC0001140
Platencin
Amino benzoic acids 3
Arylpolyene 1
Balhimycin related 1
Benzoic acids 4
Cyclitol 2
Diamino acid 1
Ectoine 2
Enediyne 1
Halogenated ring 2
Heterocyclic ring 2
Indole 1
Indolocarbazole 1
Lipid 1
Macrolactam 1
Methylation 1
Non-proteinogenic AA 4
Nucleoside 1
Piperidine 1
PKS 4
PKS-extender unit 1
Pyrrole 2
Ring 1
Sugar 5
Terpene 2
BGC0001597
Fluvirucin b2
BGC0001038
Pyralomicin 1a

Fig 3

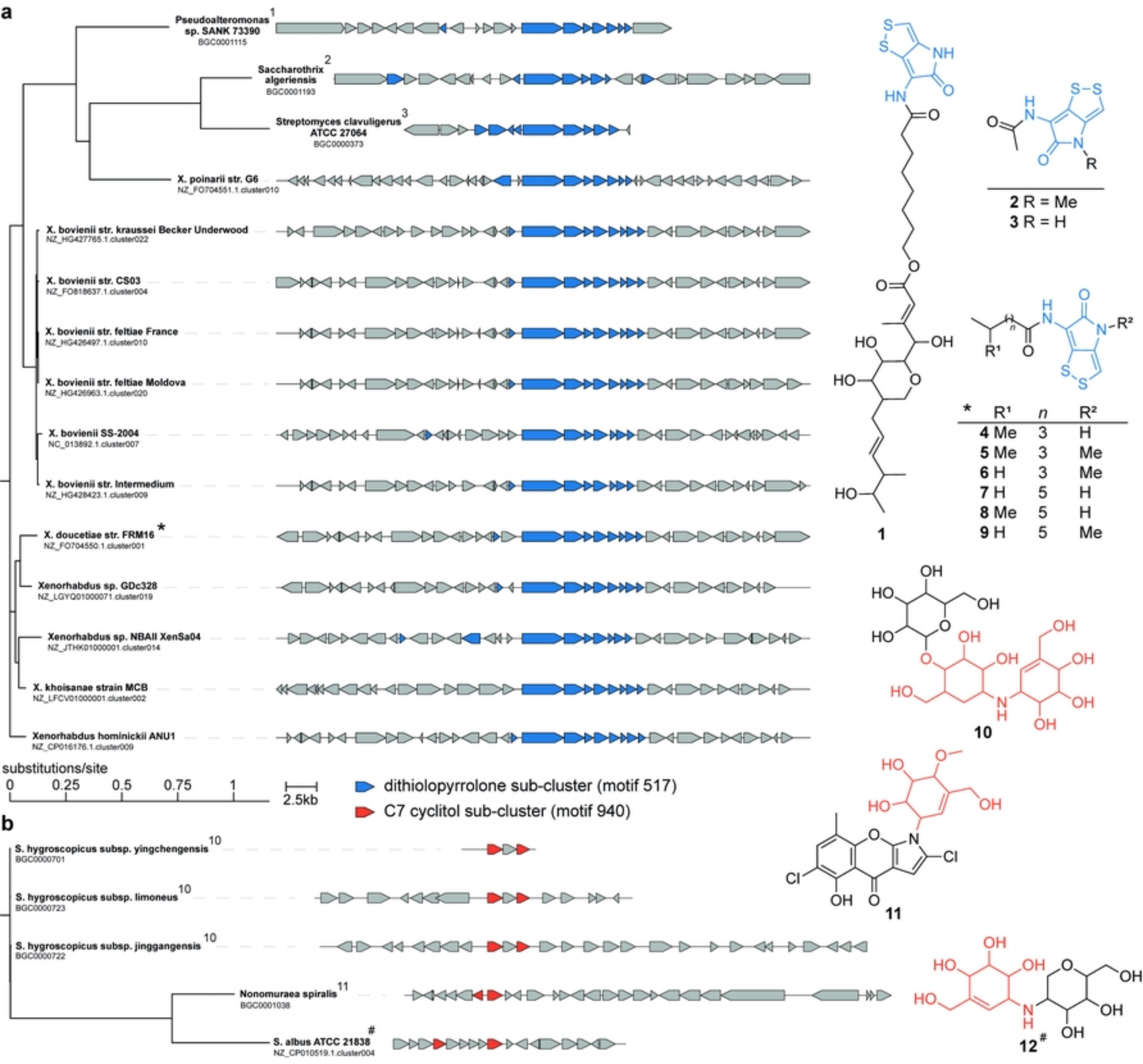
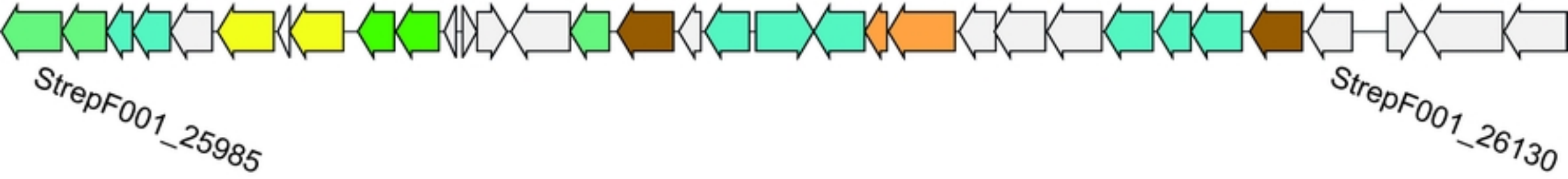

a
Pseudoalteromonas sp. SANK 73390
BGC0001115
Saccharothrix algeriensis
BGC0001193
Streptomyces clavuligerus ATCC 27064
BGC0000373
X. poinarii str. G6
NZ_FO704551.1.cluster010
X. bovienii str. kraussei Becker Underwood
NZ_HG427765.1.cluster022
X. bovienii str. CS03
NZ_FO818637.1.cluster004
X. bovienii str. feltiae France
NZ_HG426497.1.cluster010
X. bovienii str. feltiae Moldova
NZ_HG426963.1.cluster020
X. bovienii SS-2004
NC_013892.1.cluster007
X. bovienii str. Intermedium
NZ_HG428423.1.cluster009
X. doucetiae str. FRM16
NZ_FO704550.1.cluster001
Xenorhabdus sp. GDc328
NZ_LGYQ01000071.cluster019
Xenorhabdus sp. NBAII XenSa04
NZ_JTHK01000001.cluster014
X. khoisanae strain MCB
NZ_LFCV01000001.cluster002
Xenorhabdus hominickii ANU1
NZ_CP016176.1.cluster009
substitutions/site
0 0.25 0.5 0.75 1
2.5kb
dithiolopyrrolone sub-cluster (motif 517)
C7 cyclitol sub-cluster (motif 940)
b
S. hygroscopicus subsp. yingchengensis
BGC0000701
S. hygroscopicus subsp. limoneus
BGC0000723
S. hygroscopicus subsp. jinggangensis
BGC0000722
Nonomuraea spiralis
BGC0001038
S. albus ATCC 21838
NZ_CP010519.1.cluster004
1
2 R = Me
3 R = H
* R¹ n R²
4 Me 3 H
5 Me 3 Me
6 H 3 Me
7 H 5 H
8 Me 5 H
9 H 5 Me
10
11
12#

Fig 4

QZWF01000007.1.region003

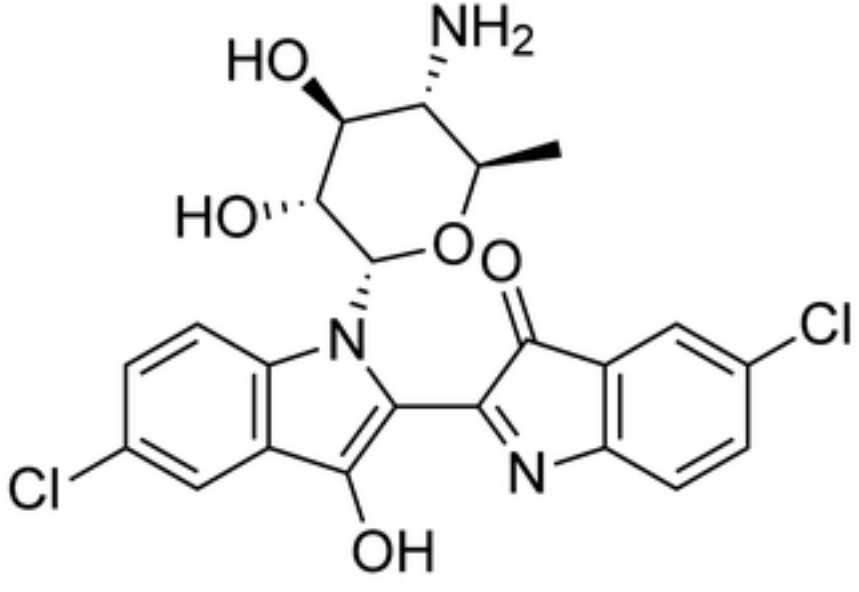


iPRESTO predictions

- deoxy-aminosugar (motif 194 & 680)
- halogenated ring (motif 607)

Other features

- sugar-related
- transport
- p450
- oxidoreductase



akashin a

Fig 5

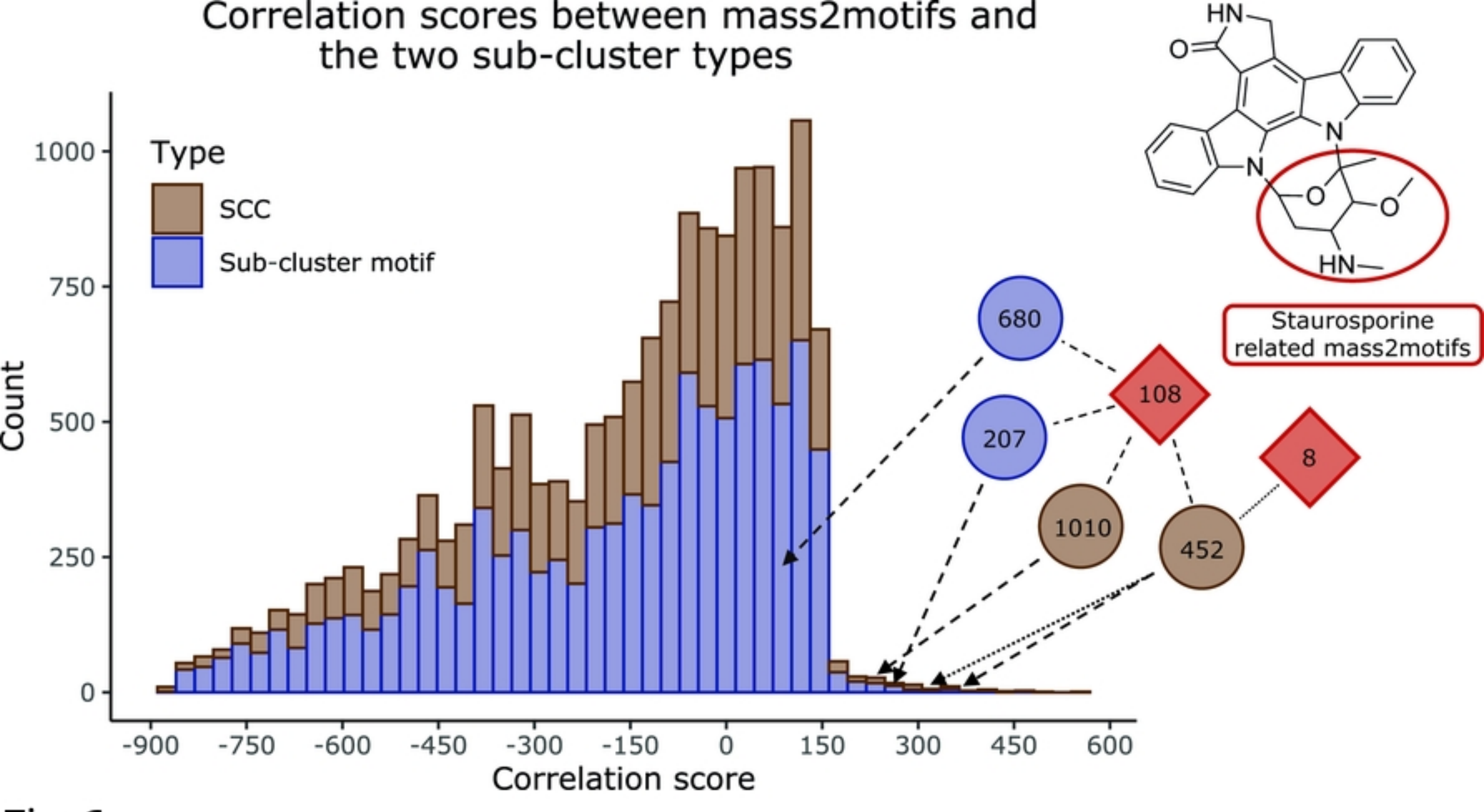

Correlation scores between mass2motifs and the two sub-cluster types
Type
SCC
Sub-cluster motif
Count
1000
750
500
250
0
-900
-750
-600
-450
-300
-150
0
150
300
450
600
Correlation score
680
207
108
1010
452
8
Staurosporine related mass2motifs
HN
O
N
N
O
O
HN

Fig 6