Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS

Santiago Sanchez[1], Joel D. Rogers[2], Alexander B. Rogers[1], Maaly Nassar[3,4], Johanna McEntyre[3], Martin Welch[2], Florian Hollfelder[2], Robert D. Finn[1]

[1]Microbiome Informatics Team, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[2]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

[3]Literature Services, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[4]SciBite - an Elsevier Company, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1DR, UK

Corresponding author email: rdf@ebi.ac.uk

Abstract

Natural products (bio)synthesised by microbes are an important component of the pharmacopeia with a vast array of biomedical applications, in addition to their key role in many ecological interactions. One approach for the discovery of these metabolites is the identification of biosynthetic gene clusters (BGCs), genomic units which encode the molecular machinery required for producing the natural product. Genome mining has revolutionised the discovery of BGCs, yet metagenomic assemblies represent a largely untapped source of natural products. The imbalanced distribution of BGC classes in existing databases restricts the generalisation of detection patterns and limits the ability of mining methods to recognise a broader spectrum of BGCs. This problem is further intensified in metagenomic datasets, where BGC genes may be split across multiple contigs. This work presents SanntiS, a new machine learning-based approach for identifying BGCs. SanntiS achieved high precision and recall in both genomic and metagenomic datasets, effectively capturing a broad range of BGCs. Application of SanntiS to

metagenomic assemblies found in MGnify led to a resource containing 1.1 million BGC predictions with associated contextual data from diverse biomes. Additionally, experimental validation of a previously undescribed BGC, detected solely by SanntiS, further demonstrates the potential of this approach in uncovering novel bioactive compounds. The study illustrates the significance of metagenomic datasets in comprehensively understanding the diversity and distribution of BGCs in microbial communities.

Introduction

The microbial world harbours an enormous reservoir of natural products (also termed secondary metabolites), many of which are widely studied and exploited as therapeutic molecules in a vast range of medical applications (Sassone-Corsi et al. 2016; Spradlin et al. 2019; Newman and Cragg 2020), especially in the area of antimicrobials. Many natural products have industrial applications, including food preservation (Gaber El-Saber Batiha, Diaa E. Hussein, Abdelazeem M. Algammal, Toyosi T. George, Philippe Jeandet, Ali Esmail Al-Snafi, Achyut Tiwari, Jorge Pamplona Pagnossa, Clara Mariana Lima, Nanasaheb D. Thorat, Muhammad Zahoor, Mohamed El-Esawi, Abhijit Dey, Saad Alghamdi, Helal F. Hetta, Natália Cruz-Martins 2021), animal growth promoters (Kierończyk et al. 2020), and phytostimulants (Grover et al. 2021). Studies on the role of natural products in their native environments suggest they play a major part in many ecological processes due to their wide range of functions such as microbial defence systems, signalling molecules for biofilm formation, or long-distance mediation between bacteria and their surrounding neighbours (Tyc et al. 2017).

The era of genomics has revolutionised approaches for detecting biosynthetic gene clusters (BGCs), which are sets of colocated genes that encode the molecular machinery necessary for producing natural products. Classically, molecular biology techniques such as gene knockout have been used to identify BGCs, but these methods are time- and resource-intensive. Today, comparative genomics has emerged as a powerful tool for identification and systematic evaluation of BGCs, greatly increasing the discovery rate by reducing both time and cost. Despite these advancements, the discovery of BGCs remains limited by our ability to cultivate and isolate microorganisms for genome sequencing, a feature often exacerbated when organisms are found living in complex environments.

Metagenomics provides a cultivation-independent approach for understanding the taxonomic and functional compositions of microbial communities, and has provided new insights into their complex dynamics and their relationship to their environment. The past five years have

witnessed an increasing availability of deeply sequenced short-read metagenomic datasets, which has enabled the assembly of contigs capable of spanning BGCs. This, coupled with emerging long-read metagenomic sequence datasets, provides a rich source for identifying BGCs and understanding their heterogeneity and distribution in different environments. The study of natural products and the BGCs encoding them at the microbial-community level have underlined their potential as a source of molecular novelty (Paoli et al. 2022; Nayfach et al. 2021), the importance of chemical interactions in shaping these communities (Ibrahim, Raajaraam, and Raman 2021), and highlighted the vast potential applications (Sayed et al. 2020).

BGCs are commonly classified based on the chemical structure of the natural product and/or the characteristics of the biosynthetic pathway, with broad classes such as Alkaloids, Non-Ribosomal Peptides (NRPs), Polyketides, Ribosomally synthesised and Post-translationally modified Peptides (RiPPs), Saccharides or Terpenes widely accepted (Terlouw et al. 2023), with each class being subdivided based on specific characteristics. Computational approaches to detecting BGCs have often focused on detecting specific classes and/or subclasses of BGC (Skinnider et al. 2017; Tietz et al. 2017; Wolf et al. 2016), while others have aimed to predict BGCs from all classes (e.g. Cimermancic et al. 2014; Blin et al. 2021; Hannigan et al. 2019; Carroll et al. 2021). Despite these efforts, a key challenge persists: the generalisation of detection patterns across a diverse array of BGCs. This challenge is confounded by the significant imbalance of experimentally validated BGC classes in established databases, such as MIBiG (Kautsar et al. 2020; Supplementary Figure 1b). This constraint limits the ability to identify a wide spectrum of BGCs. Fragmented datasets, typically produced from metagenomic assembly, further exacerbate this limitation, as BGC genes may be distributed across multiple contigs. Consequently, current methods have difficulties in accurately detecting diverse BGCs in these complex datasets, highlighting the need for a more robust and adaptable approach. In this work, we address the challenges associated with current BGC prediction through the development of a novel machine learning-based approach that is suitable for both genomic and metagenomic datasets, and capable of detecting a broad range of BGCs. We demonstrate the potential of this method through extensive benchmarking and experimental validation of a natural product exclusively detected by our method. Applying SanntiS to 24,408 metagenomic assemblies from a wide array of biomes, we provide new insights into the diversity and distribution of BGCs found in various environments.

Results and Discussion

A machine learning approach for BGC detection

Several widely used methods are currently available for prediction of BGCs in genomic datasets, including ClusterFinder, antiSMASH, deepBGC, and GECCO (Cimermancic et al. 2014; Blin et al. 2021; Hannigan et al. 2019; Carroll et al. 2021). Our benchmarking exercises (see section below) show that the ability of these methods to detect certain classes of BGCs is limited or impacted by high false positive rates. Furthermore, the machine learning-based methods ClusterFinder and deepBGC have high recall, but at the expense of precision. To address these limitations, we developed a BGC detection method based on deep learning, SanntiS (**S**econdary metabolite gene cluster **a**nnotations using **n**eural **n**etworks **t**rained on **I**nterPro **s**ignatures; Supplementary Figure 1a). SanntiS was trained using a positive set of 1,858 non-redundant experimentally validated BGCs, collected from MiBIG DB v2.0 (Kautsar et al. 2020) and a negative set of 5,574 randomly selected nucleotide regions from the species representative genomes of GTDB (release 86; Parks et al. 2022).

At the core of SanntiS is the detection model, an artificial neural network (ANN) with a one-dimensional convolutional layer, plus a bidirectional long short-term memory (BiLSTM). The model was developed using linearized sequences of protein annotations based on a subset of InterPro (see Methods) as input. The use of InterPro results in a 13% increase in the number of annotated proteins in BGCs (Supplementary Figure 1b) when compared with using Pfam (Mistry et al. 2021) alone, which is the common functional annotation method used by other methods. The detection model was further optimised to address the challenge of predicting BGC classes that have a limited representation in the training set. To achieve this, we employed a duration robust loss function (RLF, orange in Supplementary Figure 1c; Akiyama et al. 2020). Unlike traditional methods, RLF mitigates the issue of class imbalance, which can arise from the disparities in BGC counts by class and the variation in the duration of detection events - in this context, the disparities in length across different BGC classes. It accomplishes this by adjusting the binary-cross entropy loss function based on the complexity of detecting a specific event. This results in an increase in the overall performance of the model (Supplementary Figure 1c), from an area under the precision-recall curve (AUCPR) of 0.60 to 0.64, when compared with the use of the standard binary cross-entropy (BCE, blue in Supplementary Figure 1c) loss function. Finally, a post-processing random forest classifier (RF) sorts the regions with a high probability of being BGCs (default probability of 0.85) as BGC or non-BGC.

High F1 and F2 scores in genomic and metagenomic datasets

We compared the predictive power of our BGC-detection method, SanntiS, with three other methods with similar goals (antiSMASH, deepBGC, and GECCO) using the metrics precision, recall, F1, and F2 scores. Our assessment of the predictions focused on three key aspects. Firstly, the use of cluster and average metrics. This approach was specifically chosen to mitigate the limitations of point-based metrics in representing range events - a term referring to events that span a sequence of consecutive time points or locations. This is pertinent in situations where the datasets are class unbalanced and the length distributions of BGCs are uneven. In the cluster approach the overlap of a predicted region with the ground truth BGC is considered as a true positive (TP) and a predicted region not overlapping any ground truth BGC regions is a false positive (FP; Supplementary Figure 2a). In the average-based approach, precision is calculated for each predicted BGC by dividing the number of TP coding sequences (CDS) by the total CDSs in the prediction, followed by averaging these values across all predicted regions. Similarly, recall is computed for each ground truth BGC by dividing the number of TP CDSs by the CDSs in the BGC, and then averaging the results across all predicted regions (Supplementary Figure 2b). The second key aspect of our assessment consisted of statistics for individual BGC classes to prevent biases in the results towards the most prevalent classes (see Methods); and, finally, evaluation on both isolate genomes and metagenomics datasets.

To construct our benchmark, we used four distinct datasets: (i) the widely utilised, manually annotated "9 genomes" for BGC and non-BGC regions; (ii) the "15 genomes" which includes the "9 genomes" from (i) and the "6 genomes" datasets (genomes with manually annotated BGC regions; Hannigan et al. 2019); (iii) a synthetic isolate genome, constructed by combining manually annotated BGCs from version 3.0 of the MiBIG database (Terlouw et al. 2023) with randomly selected proteins from the representative genomes of GTDB (Parks et al. 2022). This approach ensured that we did not introduce any previously unidentified BGCs into the benchmark; and (iv) a synthetic metagenomic assembly, derived from fragmenting the synthetic isolate genome, used in (iii), with lengths following the distribution of metagenomic assemblies from diverse environments (see Methods and Supplementary Figure 3). We maintained the integrity of our evaluation by confirming that the BGCs labelled as positive in the benchmark dataset were not already included in the training set.

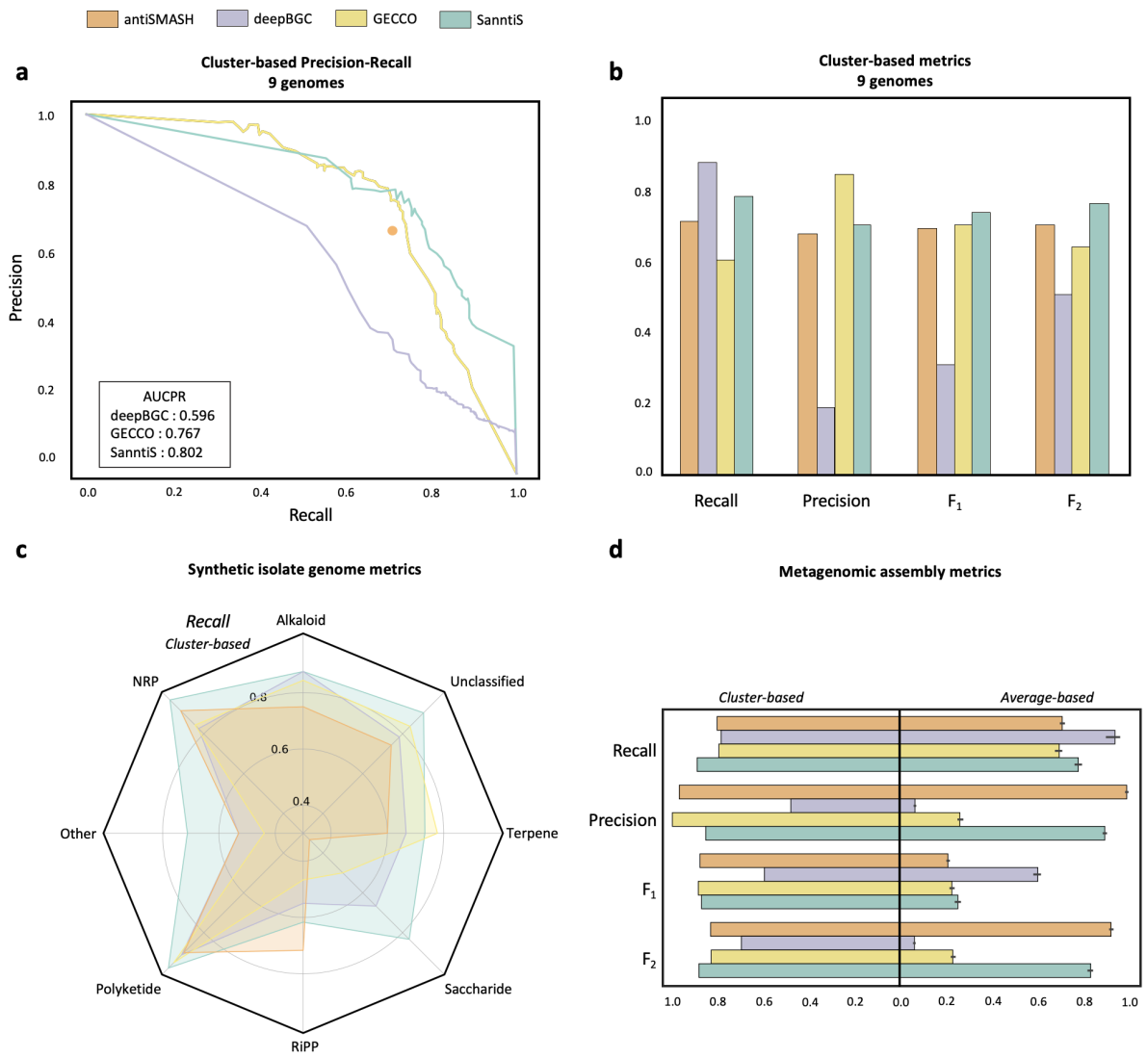


Figure 1: SanntiS performance on genomic and metagenomic datasets with comparison to antiSMASH, deepBGC, and GECCO. **(a)** Cluster-based precision-recall curve of each method at different probability thresholds in 9 genomes. The area under the precision-recall curve (AUCPR) is shown for those methods with the option of changing probability thresholds. **(b)** Cluster-based recall, precision, F1 and F2 scores of each method using default parameters in 9 genomes. **(c)** Cluster-based recall for each BGC class in a synthetic isolate genome. The following correspond to the number of BGCs in each class: Alkaloid, *n*=32; NRP, *n*=168; Other, *n*=66; Polyketide, *n*=183; RiPP, *n*=60; Saccharide, *n*=6; Terpene, *n*=45; Unclassified, *n*=74. **(d)** Cluster-based and average-based recall, precision, F1 and F2 scores and standard deviation bars of each method in the 3 replicates of synthetic metagenomic assembly.

SanntiS achieved the highest AUCPR among all the methods evaluated, but note that this metric could not be calculated for antiSMASH due to its rules-based approach. Overall, the results of the benchmark show that SanntiS outperforms other methods when high recall is desired, as shown in the bottom right of Figure 1a. This high recall rate, alongside high precision, is reflected in SanntiS F2 score (best on all test datasets), and F1 score. On the 9 genomes test set, the recall of SanntiS is the second-highest, which is only exceeded by

deepBGC. However, in terms of combined precision-recall, SanntiS demonstrates superior performance, maintaining high precision unlike deepBGC, which sacrifices precision for high recall. As a result of the precision and recall, SanntiS has the highest F1 (0.76) and F2 (0.77) scores, when using default parameters (Figure 1b). A similar level of performance is also exhibited when SanntiS is applied to the synthetic isolate genome (Supplementary Figure 4a).

The overall high recall of our method is a direct result of the ability to detect BGCs consistently across different classes of BGC. The synthetic isolate genome revealed that the macro-average recall of SanntiS (0.83) exceeds that of the other methods (Figure 1c). Specifically, the predictive power in the *Saccharides* and *Others* classes is improved compared to the other methods. A similar result is also achieved in the 15 genomes test set (Supplementary Figure 1b), but in this case the SanntiS macro-average recall is 0.75, surpassed only by deepBGC (0.84).

On highly fragmented sequences from metagenomics assembly, SanntiS outperformed other methods with the highest F1 and F2 scores in cluster (0.82 and 0.79) and average (0.38 and 0.29) based metrics (Figure 4d). SanntiS and antiSMASH precision notably outperform the other methods when applied to metagenomics datasets.

SanntiS detects BGCs undetected by other methods

Encouraged by the benchmarking results, we applied SanntiS to isolate genomes to understand the diversity of the predictions compared to the same methods evaluated in the aforementioned benchmarking. To do so we predicted BGCs in the species representative genomes of GTDB (release 86). The results of our analysis indicate that SanntiS produces results highly consistent with other methods as indicated by the upset plot (Figure 2a), which shows a high level of overlap between SanntiS results and those produced by other methods. However, we were also able to identify a significant number of unique observations using our method, with a total of 29,656 BGCs uniquely identified by SanntiS.

To gain further insights into the novel 29,656 BGC predictions produced by SanntiS (insert panel in Figure 2a), we calculated the Jaccard distance (JD) between the BGCs predicted by SanntiS and the BGCs found in MiBIG (version 2.0). As BGCs can undergo rearrangement, the JD was based on the Pfam domain composition, with smaller JD values indicating a closer similarity between the prediction and MiBIG. This approach indicated that the more dissimilar predictions came predominantly from the BGC classes Polyketide and Other, while more similar predictions were typically found for RiPP and Other. Overall, the median JD was 0.66, which was within the expected range, as more similar BGCs (with smaller JD scores) are more

likely to have been detected by the other methods, and hence not represent a unique SanntiS prediction.

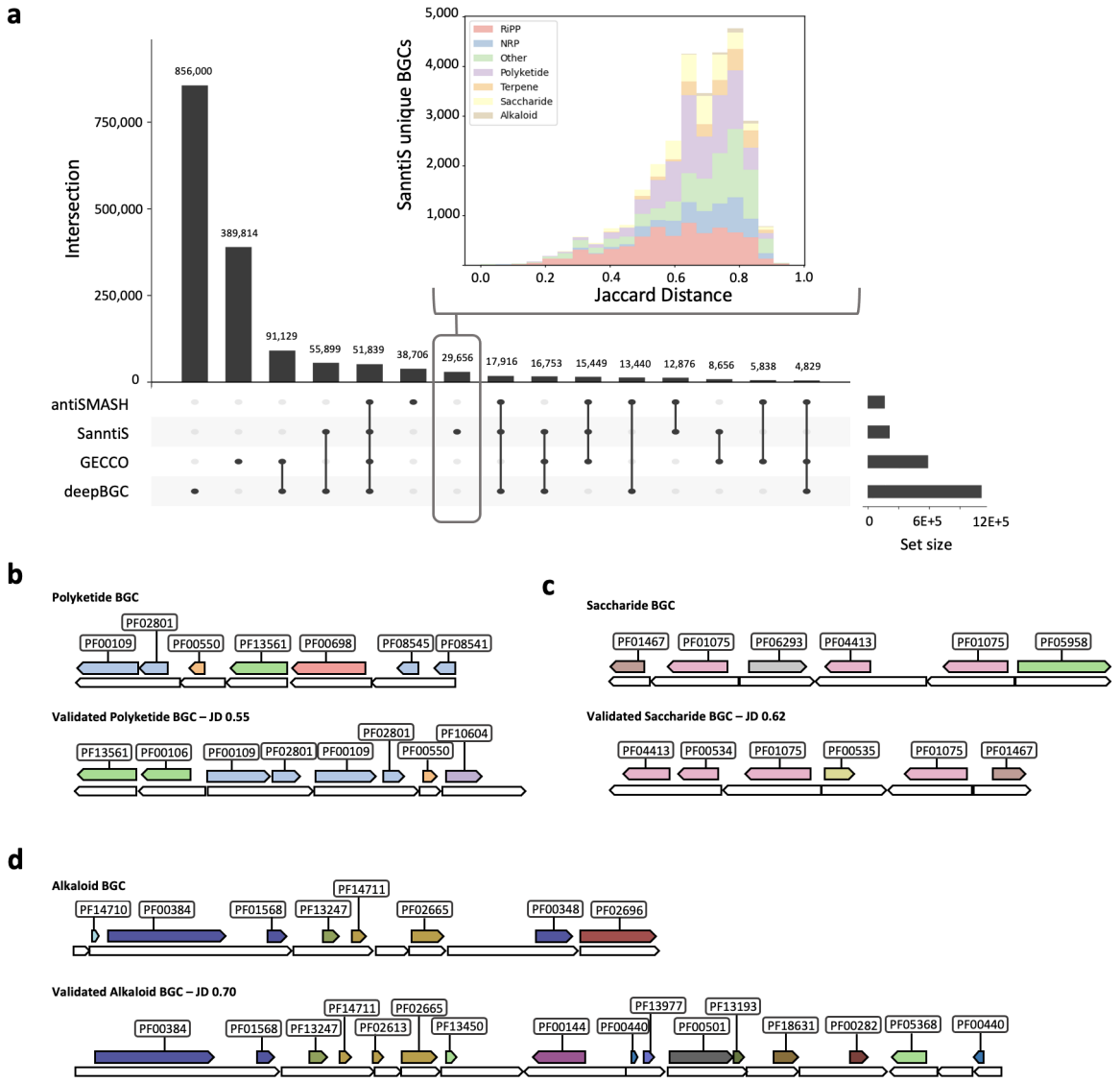


Figure 2: Putative BGCs detected only by SanntiS. **(a)** Upset plot showing the intersection of BGC regions predicted by each method. Histogram of Jaccard Distance from BGC regions predicted only by SanntiS to the nearest BGC in MiBIG DB. Polyketide, 25.1%; RiPP, 24.0%; Other, 22.0%; NRP, 11.2%; Saccharide, 10.3%; Terpene, 6.4%; Alkaloid, 1.0%. **(b, c, and d)** Gene architecture of BGCs from the GTDB representatives found only by SanntiS (top) and their nearest MiBIG BGC by Jaccard distance (bottom). White/empty elements indicate proteins. Coloured elements indicate Pfam domains (coloured by Pfam clan). Labelled boxes indicate the Pfam domain accession. **(b)** Putative polyketide compared to BGC0000228. **(c)** Putative saccharide compared to BGC0000779. **(d)** Putative alkaloid compared to BGC0001137.

While the JD scores provide an overall impression of the predictions provided by SanntiS compared to known BGCs, we next investigated how the structure of the BGC compared to the nearest known BGC for different classes of BGCs. There were multiple examples across

different JD scores and classes where the predicted BGC and a validated BGC shared many similar features (e.g. Figure 2b, c, d). In our detailed analysis of uniquely detected RiPPs — driven by the relative ease to experimentally validate them compared to other BGC classes — we discovered a particularly interesting putative BGC within the *Brevibacillus laterosporus* genome. This BGC contained an Open Reading Frame (ORF) annotated as a laterosporulin peptide (PF17861; Figure 3a), a domain known to be associated with antimicrobial activity.

As there are no BGCs with this domain in the MiBIG database (and hence our training set), we searched the literature for experimentally validated BGCs with similar characteristics and found two additional examples, which encode for the RiPPs LS (Singh et al. 2012) and LS10 (Baindara et al. 2016), that had a comparable gene architecture and domain composition. Comparison of the BGC organisation revealed an additional ORF in the 3' end of the BGC that was not covered by the SanntiS prediction. Furthermore, multiple sequence alignment of the predicted laterosporulin (designated LS-E) and the experimentally validated laterosporulins (Figure 3c) indicated that LS-E is more closely related to LS10 than to LS, but the sequences shared many invariant residues in both the leader sequence and the mature peptide (Figure 3c). This high degree of similarity was also found when comparing the determined structure of LS10 (PDB accession 4OZK) and the structural model of the mature LS-E (obtained using AlphaFold 2.0; Jumper et al. 2021), with a root-mean-square deviation of <0.6Å. The N terminus of mature LS (positions 2 to 7 in the alignment in Figure 3c) has been proposed as a key feature of the sequence for forming a possible trimer (Singh et al. 2015). This region undergoes a structural transformation from coil (when monomer) to a β-sheet (when trimer), and is stabilised with the β-strand 1 (Figure 3b). Interestingly, LS-E has two insertions (corresponding to positions 4 and 5 in the alignment shown in Figure 3c) in this region that are also present in LS10. LS-E also contains a Gln→Glu substitution at position 6. Another key feature associated with trimer formation are salt bridge interactions between Arg45 and Glu47. In LS-E Arg45 is substituted with Lys, a conservative substitution. The substitutions at these two positions do not appear to invoke major structural changes in the position of the side chains (Figure 3b, structure alignment, side chains).

In order to validate this SanntiS prediction further, we carried out antimicrobial susceptibility testing of the predicted mature LS-E. To this end, we incubated the chemically synthesised peptide, at a range of concentrations, with *Staphylococcus aureus* ATCC 25923 (a known target of LS and LS10, with the strain used for these tests being MTCC 1430), and assessed the impact on bacterial cell growth via optical density (absorbance) measurements (Figures 3d and 3e). We found that the peptide clearly affected antimicrobial activity (IC_{50} = 55 ± 13 μM), even though this activity was somewhat lower than that demonstrated for LS or LS10 against the

same bacterial species. This suggests that LS-E may have evolved different target specificity or reduced potency. Nevertheless, our findings demonstrate that SanntiS can detect a novel RiPP BGC, validating the extrapolation from sequence to function.

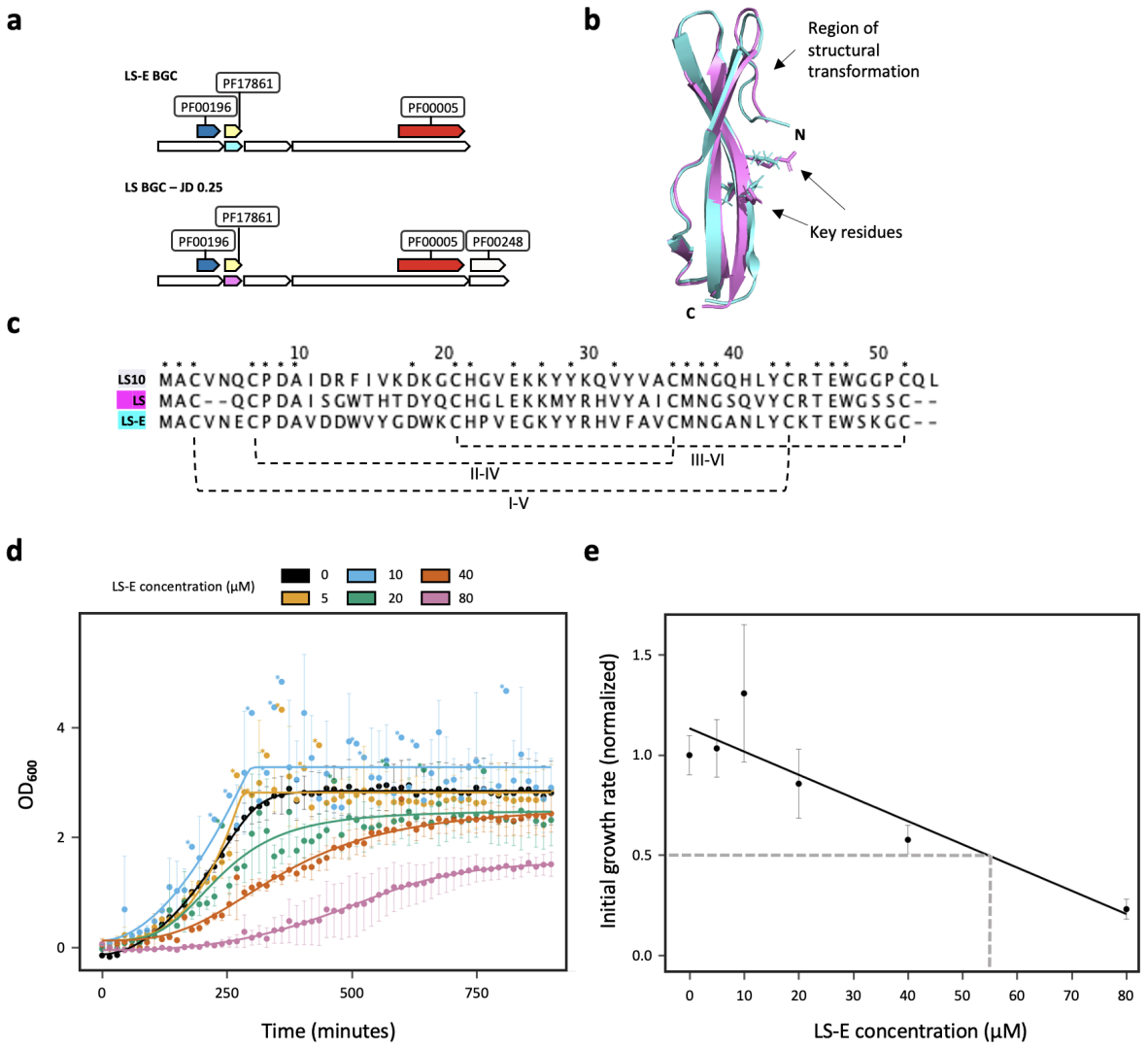


Figure 3: Laterosporulin discovery and experimental validation. **(a)** Gene architecture of the BGC encoding LS-E compared with the BGC of experimentally validated laterosporulin, LS. Protein domains are coloured according to their Pfam clan. **(b)** Structure alignment of mature LS (PDB accession 4OZK; magenta) and mature LS-E (predicted structure; cyan). Side chains of key residues in positions 43-45 are shown. **(c)** Multiple protein sequence alignment of LS-E, LS and LS10. An asterisk (*) indicates positions which have fully conserved residues in this alignment. Disulfide bonds of LS are shown with dashed lines. **(d)** Growth curves based on OD_{600} measurements of *S. aureus* cultures incubated with varying concentrations of LS-E. The data are overlaid with curves based on five-parameter Hill equations. Data points are the means of three replicates, with standard deviation indicated by error bars. Data points with very large standard deviations (greater than 1.25 OD_{600} units) are denoted with an asterisk. **(e)** Initial growth rates *vs* LS-E concentration. Lag-phase extension, cell killing and growth inhibition are collapsed into a factor representing the average gradient between the zero time point and the sigmoidal curve's

inflection point. All factors are normalised to the negative control (untreated culture). LS-E's IC_{50} (dashed line) was estimated using linear regression.

Expanding BGC diversity with metagenomic assemblies from MGnify

To date, the majority of experimentally validated BGCs have been confined to a limited number of taxonomic clades (27 classes in MiBIG 2.0). However, the taxonomic distribution of predicted BGCs on genomes found in GTDB (Supplementary Table 8) suggests that the actual distribution of BGCs may be different. Considering the strong performance of SanntiS on simulated metagenomic datasets and the broader sampling of microbes enabled by culture-independent approaches, we further explored the BGC repertoire encoded in metagenomes. We applied SanntiS to 24,408 public metagenomic assemblies from the MGnify resource, which originated from a diverse range of ecologically and geographically distributed samples.

This analysis identified 1,145,552 putative BGC regions containing BGCs from all classes (Supplementary Figure 5). This dataset is now publicly available in MGnify (Supplementary Figure 6). To assess the similarity of these BGCs with those found in other large-scale works, we compared our results, denoted as MGnify-SanntiS, with a study that identified approximately 1.2 million BGCs in 188,622 isolate genomes and 20,584 metagenome-assembled genomes (MAGs) (Kautsar et al. 2021). To our knowledge, this latter study represents the largest BGC-mining study to date, and we refer to this dataset as BiG-SLiCE. To understand the overlap in BGC predictions between the two datasets, we combined and compared the two sets of predictions to identify similar BGC predictions (see methods). This analysis revealed 376,922 unique sequences in the BiG-SLiCE dataset, 570,603 in MGnify-SanntiS, and only 2,661 in both datasets.

Our previous benchmarking results, using both SanntiS and antiSMASH (the detection method used in BiG-SLiCE), suggested an expected overlap of approximately 47% if both methods were applied to identical sequences (see Supplementary Figure 7). The smaller overlap identified can likely be attributed to the distinct sequence sets used in each study. Further, the comparative approach reduced the number of BiG-SLiCE BGCs by 68%, indicating a significant redundancy in this dataset and reflecting the bias in isolate genome sequencing. A reduction of 50% was observed for the MGnify-SanntiS dataset, underscoring the diversity of sequences within MGnify.

Certain BGC classes, such as NRP and Polyketides, were detected less frequently in metagenomic assemblies compared to the BiG-SLiCE dataset, likely due to challenges in effectively assembling these long sequences with repetitive domains. As a result, the

representation of these classes in metagenomes was lower than other classes, highlighting a limitation in metagenomic assembly rather than the detection method itself. Nevertheless, the comparison of MGnify-SanntiS to BiG-SliCE BGC predictions led to a significant 1.5-fold increase in the number of non-redundant putative BGCs. This increase was most pronounced in the RiPP, Other, and Saccharide BGC classes (Figure 4a), supporting the notion that the MGnify-SanntiS dataset harbours hitherto unknown BGC diversity.

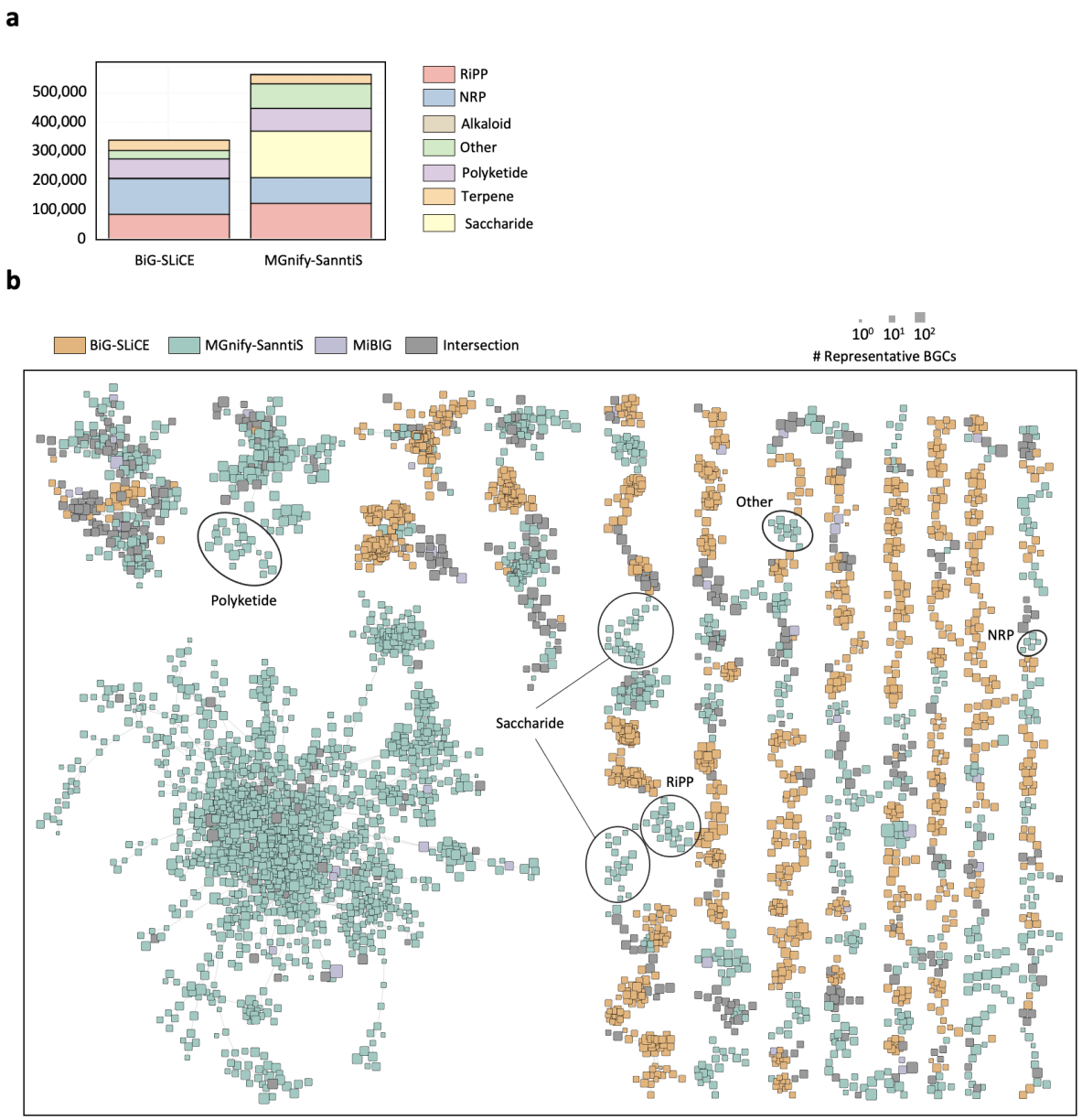


Figure 4: Metagenomic assemblies harbour BGC novelty. Comparison of BiG-SLiCE, MGnify-SanntiS, and MiIBIG BGC datasets. **(a)** Class distribution of deduplicated sequences by dataset. **(b)** Similarity network of BGC domain composition cluster representatives. Size of nodes reflects the number of representative BGCs in a cluster. Colours represent dataset composition of the cluster. Orange contains only BGCs from BiG-SLiCE dataset; green contains only BGCs from MGnify-SanntiS dataset; purple contains at least one MiBIG representative; grey, cluster contains BGCs from both BiG-SLiCE and MGnify-SanntiS datasets.

To assess the novelty of BGCs in the MGnify-SanntiS dataset, we conducted a clustering analysis of BGCs from both the MGnify-SanntiS and BiG-SLiCE datasets. This analysis was based on functional domain composition, independent of BGC organisation. The analysis identified 24,794 unique BGC clusters: 10,135 exclusively within the BiG-SLiCE dataset, 9,732 solely in the MGnify-SanntiS dataset, and 4,927 consisting of representatives from both datasets. A key finding in the meta-cluster analysis is a significant number of Saccharide BGCs (Figure 4b, bottom left), mostly consisting of exclusive SanntiS predictions.

Many of these meta-clusters are compositionally very different from any other BGC in the genomic space and from any BGC included in the MiBIG DB (Figure 4b). These entirely uncharacterized BGC regions hold potential as sources of novel therapeutic molecules, but require experimental validation to confirm their functionality.

Biome range of BGCs in metagenomic datasets

Analysing the BGC prediction results across different biomes, we found that the human digestive system, marine, mammal digestive system, wastewater, bioreactor, and human skin contained the highest numbers of predicted BGCs, as illustrated in Supplementary Figure 5. This outcome is expected, as these biomes have the most assemblies (refer to Supplementary Table 9). Notably, within well sampled biomes (represented by over 50 assemblies), engineered biomes such as bioreactors and wastewater displayed the highest average number of BGCs per assembly (166.4 and 113.1, respectively). This observation is in agreement with previous studies, which demonstrated that wastewater treatment plants, characterised by antibiotic residues and high biomass concentrations, serve as hotspots for horizontal gene transfer, potentially including polyketides and NRPs (Guo et al. 2017; Ziemert et al. 2014). Interestingly, the human skin biome exhibited a particularly low average number of BGCs per assembly at 10.2, a reflection of its relatively lower microbial diversity (Donia et al. 2014).

To determine if these samples captured the full diversity of BGCs from the source biome, we constructed rarefaction curves for each biome (Supplementary Figure 8). Our analysis indicated that only the human digestive system, marine, and mammal digestive system displayed some form of levelling off of new BGC discovery, but the lack of asymptotic curves suggests that numerous additional BGCs are likely to be discovered as more sampling is conducted. We further explored the relationship between the number of BGCs and BGC clusters per assembly for each biome. Our analysis revealed a strong correlation between these variables. A closer examination of the biomes that exhibited signs of saturation in the rarefaction curves (marine, mammal digestive system, and human digestive system) showed that marine samples displayed a higher relative redundancy of clusters compared to mammals and human digestive tract

biomes, as reflected in linear regression slopes of 0.74, 0.88, and 0.95, respectively (Figure 5a). In terms of class distribution, we observed that human and mammal digestive systems demonstrated similar patterns, with saccharides and RiPPs dominating the distribution. However, the class distribution of marine samples differed significantly from that of digestive systems, with fewer RiPPs and a higher proportion of terpenes (Supplementary Figure 5). These differences may be attributed to the lower organism density in aquatic environments compared to digestive systems, as RiPPs are short-range interaction molecules, while terpenes are long-range interaction molecules (Tyc et al. 2017).

To investigate whether BGCs are restricted to specific biomes or distributed across multiple biomes, we plotted BGC clusters as a function of the number of biomes in which they occurred, as well as the top-level biome classes, namely Engineered, Environmental, and Host-associated (Figure 5b). Most clusters display a limited biome range, with the distribution centred around two biomes. The majority of the data is found within a single biome class, host-associated, which can be partially explained by the high overlap between closely related biomes such as human and mammal digestive systems. Nevertheless, a significant number of clusters were observed spanning across environmental, host-associated, and engineered biomes.

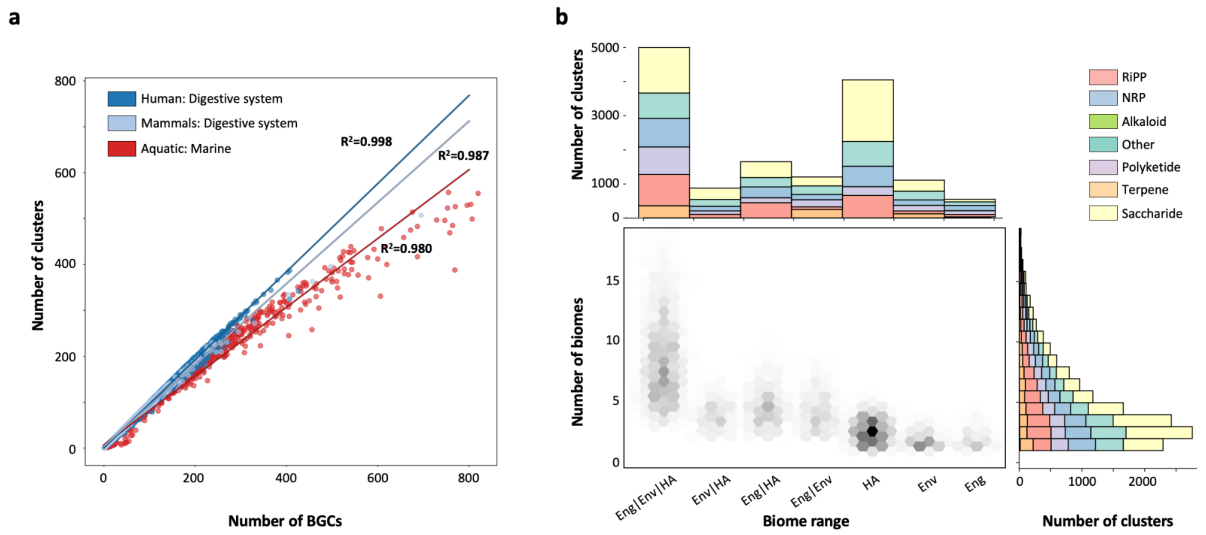


Figure 5. Class distribution and biome range of BGC clusters. **(a)** Relationship between number of detected BGCs and the number of BGC clusters, each dot represents an assembly. **(b)** Distribution of biome range of BGC clusters. X-axis, Broad biome classification in which a BGC cluster is present. Engineered, (Eng). Environmental, (Env). Host-associated, (HA). Y-axis, number of biomes in which a BGC cluster is present.

These clusters were spread across a broad range of biomes, with some exhibiting a narrow range (3-5 biomes), potentially due to ambiguity in biome classification. For example, biomes like Agriculture (Engineered), terrestrial soil (Environmental), and plant rhizosphere

(Host-associated) might display considerable overlap in their microbial compositions. The remaining clusters showed a wide range, suggesting they are ubiquitously conserved across different organisms and/or could be horizontally transferred BGCs. A more in-depth analysis is required to determine the nature of these clusters. Notably, BGC clusters enriched in terpenes demonstrated a broader biome range, particularly in engineered and environmental biomes. This pattern could be attributed to the diffusion properties of terpenes, which are more volatile and soluble, enabling long-distance interactions.

In summary, our findings highlight the heterogeneity of BGCs among diverse biomes, with some biomes exhibiting a higher abundance and diversity of BGCs than others. Moreover, the variation in BGC class distribution across biomes reflects the unique ecological niches and interactions inherent in each habitat. We suggest that this comprehensive dataset could facilitate further research and enhance our understanding of the functional significance of natural products.

Conclusion

In this study, we have demonstrated the application of SanntiS, a novel machine learning-based approach for identifying biosynthetic gene clusters (BGCs) to both genomic and metagenomic datasets. Our results show that SanntiS outperforms existing methods, including core enzyme-dependent detection and other machine learning-based methods, in terms of overall performance, recall, and applicability to diverse datasets. The superior performance of SanntiS is underpinned by the application of range events-specific training and evaluation techniques. These techniques, a departure from classical point-based approaches, effectively capture the sequential nature (a chain of contiguous genes) of BGCs and tackle the issue of length imbalance across classes, both of which have been challenging for existing methods. This strategy, although previously applied in fields such as sound event detection and time-series applications, is a pioneering approach within bioinformatics, specifically for BGC detection. Nevertheless, there remains scope for improvement, such as the recall of RiPPs and overall precision. We believe this can be best addressed in the future by applying active learning techniques to provide curated results that will enhance the training data, an approach that has proven successful in other studies (*e.g.* Reker and Schneider 2015; Zhang and Lee 2019).

The strength of SanntiS is further highlighted by the experimental validation of antimicrobial activity of LS-E, a bacteriocin solely predicted by SanntiS. Further testing of LS-E against different bacterial strains could be conducted to better assess the antimicrobial potential of this RiPP. Although the potency of the LS-E peptide is relatively weak against the tested strain, the fact that no laterosporulin was used in the training dataset, yet it was still predicted, underscores the potential for SanntiS to expand the range of known natural products. In comparing the 1.1 million BGCs from our study against the BiG-SLiCE dataset, the MGnify-SanntiS resource revealed a substantial degree of potential novel BGC diversity. Notably, our analysis aligns with previous research by (Cimermancic et al. 2014), who also observed dominance of a large set of Saccharides in a network analysis of BGCs from a diverse group of prokaryotic genomes. These findings hold significant potential for characterising novel therapeutic molecules and uncovering patterns of genes that could be advantageous for rational design.

Our analysis of metagenomic encoded BGCs unveiled intriguing patterns, such as the differences in BGC redundancy between marine environments and digestive tracts. This observation might be attributed to the well-established redundancy of functions in aquatic environments (Louca et al. 2018). Importantly, these results could aid in function discovery efforts, given that environments with functional redundancy can harbour subtle distinctions in the affinities and kinetics of metabolites (Fuhrman, Cram, and Needham 2015). Furthermore,

the findings might have implications for elucidating the functional roles of natural products, as systems with a high division of labour (low redundancy of function) are known to be easier to engineer and study (Tsoi, Dai, and You 2019).

Another outcome from our analysis was the distribution of BGC clusters across different environmental niches. This distribution may hold significant implications for therapeutic applications of natural products. For example, in human medicine broad-spectrum antibiotics are widely used, but can have deleterious side effects on parts of the microbiome not being targeted and give rise to antibiotic resistance. Thus, there is an increasing demand for more specific, narrow spectrum antibiotics that are specific to a target bacteria. Gaining insights into the microbial range of action of these BGCs could help guide the development of more targeted therapies. The MGnify-SanntiS dataset is publicly available in MGnify and offers a valuable resource for various applications, such as natural product discovery and ecological modelling studies.

Methods

Code and Data availability

The SanntiS software implementation is accessible as a standalone Python package and can be downloaded via the GitHub repository at https://github.com/Finn-Lab/SanntiS. The supplementary materials contain detailed information on data accessions and resource location, as well as the necessary data for replicating the training dataset to ensure reproducibility. The data employed for testing and validation in this study is available in supplementary data. BGC predictions by SanntiS on metagenomic assemblies are readily downloadable from the MGnify website. Assemblies that were analysed using MGnify's pipeline version 5 can be visualised using the portal contig viewer.

Open reading frame (ORF) identification and functional annotation

Prodigal v2.6.3 (Hyatt et al. 2010) was used for ORF identification and protein translation for all the DNA sequences utilised in this work. The resulting proteins were then functionally annotated using a subset of the member databases in InterProScan v5.52-86.0 (Pfam, TIGRFAM, PRINTS, ProSitePatterns, Gene3D) (Jones et al. 2014) and a set of profile hidden Markov models (HMMs) from Pfam version 34.0, which had not yet been included in InterProScan. This subset of member databases was selected based on those employed in MGnify to strike a balance between the speed of annotations and coverage.

Training and validation datasets

A training and validation dataset of known BGCs was constructed using the nucleotide sequences from the MiBIG database version 2.0 (Kautsar et al. 2020), taking BGCs from all classes. To remove redundancy within this dataset and given the known rearrangements of genes within a BGC, the MiBIG BGC sequences were clustered based on the Jaccard similarity of InterProScan entries (cut-off threshold of 0.95), with the longest sequence retained as a representative. The same approach was employed to remove any MiBIG sequences found in the training set that were similar to the test datasets (described below). In total, 1858 BGC (positive) entities were used in the training and validation datasets.

For the development of a negative dataset, representative genomes from GTDB (release 86; Parks et al. 2022) were used as a source of non-BGC regions. Potential BGC sequences were identified and removed using antiSMASH v5.1. To avoid imbalance between the positive and negative datasets, the non-BGC regions were subsampled by randomly selecting 5574 chunks of consecutive ORFs (with length sampled from a normal random distribution with the same mean and standard deviation as the positive dataset).

A synthetic contig was also developed to train the models. The contig was created by first duplicating the positive entities to increase the robustness of the training. The BGC and non-BGC entities were then shuffled to produce a synthetic contig for model training. To assess the performance of the models, a cross-validation approach was used that involved five independent and stratified splits of positive classes for validation, following a similar methodology to the training contig, but excluding the duplication of positive entities. The synthetic contigs generated from these procedures were utilised to train and validate the performance of the models.

Test datasets

To evaluate the method and compare performance with other BGC finding methods, 3 different datasets were used: (i) the "nine genomes" and (ii) "six genomes" datasets, originally used in the work of Cimermancic et al. (2014), and revisited and recollected by Hannigan et al. (2019), with manually annotated BGC and non-BGC regions (294 and 65, respectively); (iii) a new test dataset using published BGC sequences not found in MiBIG.

For the generation of this last dataset, manually curated BGCs from a pre-release version of MiBIG, version 3.0 (Terlouw et al. 2023) that had a nucleotide GenBank/RefSeq accession (so the sequences could be included in the dataset) and a length below 150,000 bp were selected. This process enabled the identification of 637 new BGCs that were not present in the training dataset.

A synthetic genome was created in the following way: to create non-BGC regions, 1 genome per phylum (148) was randomly selected from the representative genomes of GTDB (release 86), and all the CDS annotated as proto-cluster by antiSMASH (v6.0) were removed. The remaining CDS of the 148 genomes and the 637 new BGC regions were shuffled and concatenated to create the "synthetic isolate genome" dataset.

To examine the performance of the BGC detection methods when executed on metagenomic assemblies, a "synthetic metagenomic assembly" was created by randomly sampling 1,000,000 contigs from assemblies in MGnify (Richardson et al. 2023) to determine a contig length probability distribution. Subsequently, the synthetic isolate genome was split into contigs of lengths that matched the metagenomic contig length distribution. This approach was repeated three times with replicas using different fixed seeds (see Supp. table 2).

Feature matrices construction

Feature matrices appropriate for model training and prediction were constructed from a linear sequence of annotation entries, ordered by their start position. Subsequently, the domains were one-hot encoded.

Model training and BGC prediction implementation

An artificial neural network (ANN) model consisting of 4 layers was built using TensorFlow v2.4.1 (Abadi et al. 2016). First, an embedding layer was employed to represent each functional annotation as a vector of 124 features. Second, a one-dimensional Convolutional (Conv1D) layer of 64 filters and a kernel size of 4 was used. Third, a bidirectional long-short term memory (BiLSTM) layer of 32 units was implemented, which conserved the original input dimensions, and a dropout layer of 0.5 was connected to a single-unit dense layer, activated with a sigmoid function. The training used the minibatch setup and applied the Adam optimizer with the learning rate dropping by a factor of 0.1 every 5 epochs of not improving the loss function on the validation dataset. A custom implementation of RLF with a gamma factor of 2.0 was used as the loss function. The training was performed using sequences of 200 timesteps.

The BGC detection process involved predicting the probability of each of the linearized annotation entities and normalising it by retaining the entity with the maximum value for each ORF. Then, given a score threshold, the regions of the sequences above that threshold were identified as BGCs. A Multiclass OneVsRest Random Forest classifier was also trained (using default parameters of scikit-learn v0.24; (Pedregosa et al. 2011) for the 7 MiBIG classes plus a non-BGC class, using the same feature matrices that were used for the ANN model. This classifier served as a post-processing filter, where predicted regions not classified as one of the 7 MiBIG classes were dropped.

Validation of SanntiS performance and comparison with other methods

Performance evaluation metrics, such as recall, precision, F1, and F2 score, were utilised. To avoid known problems of evaluations using short segments as units (Mesaros, Heittola, and Virtanen 2016), two different approaches were employed for deriving score metrics:

Cluster-Based. In this approach, the entire BGC region was used as the unit of measure, demonstrating the method's ability to detect BGCs but omitting the quality of the predictions. A true positive (TP) was a predicted BGC region overlapping a BGC region in the reference, a false positive (FP) was a predicted BGC region not overlapping any BGC in the reference, and a false negative (FN) was a BGC region in the reference not overlapping any predicted BGC.

Average metrics. Micro-averaging of intermediate statistics was implemented in this approach, which is better suited for range-based anomalies (like BGCs) as it can demonstrate the quality of the predictions and avoid biases induced by length differences, giving equal importance to all BGCs (irrespective of their length). For each BGC region in the reference, RecallT was calculated as the number of ORFs predicted in a BGC region divided by the ORFs in the reference region, with the average of this metric then being taken. For each predicted BGC,

PrecisionT was calculated as the number of ORFs that were in a BGC in the reference divided by the total number of predicted ORFs, followed by obtaining the average of the individual precisions (Tatbul et al. 2018).

Validation of the method and comparison of SanntiS v0.9.3 with antiSMASH v6.0, deepBGC v0.1.26, and GECCO v0.6.2 in real genomic contexts was performed by running the methods at different score/thresholds (from 0 to 1 with a step of 0.005) on the "9 genomes" dataset, with all other parameters set to default. Recall metrics for individual BGC classes were also calculated. BGCs in the "9 genomes" and "6 genomes" datasets were classified by obtaining the Jaccard distance (JD), as given by

$$1 - \frac{|X \cap Y|}{|X \cup Y|} \quad (1)$$

where X are the Pfam domains in one region and Y are the Pfam domains in the other, of the BGCs to those in MiBIG.

To evaluate the performance of each method in an independent validation dataset and to assess the differences in detection between a long contig genome assembly and a metagenomic assembly, evaluation metrics for the synthetic isolate genome dataset were derived using the default parameters of each method and the classes assigned by the MiBIG curators. For the prediction in the synthetic metagenomic assembly, the Prodigal "meta mode" was used in antiSMASH (with a minimal length of 500 bp) and deepBGC. For the synthetic metagenomic assembly, a true positive (TP) in the Cluster-based metrics was considered the detection of at least one CDS of a true BGC, irrespective of the number of contigs that particular BGC spans.

BGC prediction in GTDB representative genomes

The 31,910 GTDB representative genomes were downloaded, and BGC regions were predicted using SanntiS, antiSMASH, deepBGC, and GECCO with default parameters. Overlapping predictions were then identified, and regions called only by SanntiS were selected. For every region, the Jaccard distance (JD) to the BGCs in MiBIG was calculated, and the assigned BGC class corresponded to that of the nearest BGC in MiBIG. Some randomly selected BGC regions called only by SanntiS were manually examined and compared with those in MiBIG. For one of these cases, which showed strong evidence of being a true BGC, a multiple sequence alignment of the putative structural protein with those of similar known BGCs was constructed. AlphaFold version 2.0, was used to predict the structure of the putative structural protein.

Experimental validation of predicted BGC

Chemically synthesised LS was ordered from WuXi AppTec (TFA salt, 84 % purity), resuspended in Milli-Q water and filter-sterilised (PVDF; Millex SLGV004SL). Antimicrobial peptide Assay Buffer (AAB; 0.01 % (v/v) acetic acid, 0.2 % (w/v) BSA) was prepared using Milli-Q water, glacial acetic acid (Merck 27225) and BSA powder (Sigma A9418), and filter-sterilised (PES; Millipore SLGP033RS).

Streaked LB-Lennox agar plates (1.6 %; Formedium AGR10, Formedium LBX0102) of *S. aureus* (ATCC 25923) were incubated at 37 C overnight, then individual colonies were picked for overnight culture in liquid LB-Miller medium (Formedium LMM0102). This overnight culture was diluted to optical density (OD_{600}) 0.125 in Mueller-Hinton broth (Millipore 70192) before adding an equal volume of LS stock solution, then establishing a two-fold titration series by repeated dilution in OD 0.125 bacterial culture in a microtitre plate (Eppendorf 0030602102), in triplicate. Plates were sealed (SLS PCR0548) and per-well absorbance at 600 nm was measured *via* spectrophotometer (FLUOstar Omega) to monitor bacterial growth during an overnight incubation at 37 C. All plasticware used was polypropylene.

To fit the data, the following starting parameters were used: effect coefficient, E = 250; Hill coefficient, H = 10; bottom asymptote, B = 0; top asymptote, T = 3; symmetry coefficient, S = 1. A linear approximation to the Hill equation (with fixed B, T and S) was first used to obtain estimates for E and H, which then replaced their corresponding starting parameters for a non-linear model (five-parameter Hill equation; $y = B + (T - B)/(1 + (2^{(1/S)} - 1)((E/x)^{H}))^{S}$), constrained as follows: $150<E<750$; $3<H<100$; $-0.125<B<0.125$; $1.25<T<3.5$; $0<S<1$. The average initial growth rate (dy/dx) for each concentration was determined by dividing the half-maximal OD_{600} by E, then normalised to the initial growth rate measured for untreated culture. The IC_{50} for LS-E was estimated by fitting these initial growth rates via linear regression. Data analysis and plot generation was performed in R.

BGC prediction in metagenomic assemblies

Contigs longer than 3,000 base pairs were extracted from 24,506 publicly accessible metagenomic assemblies, provided by the MGnify resource. BGC regions were predicted within the contigs using SanntiS, withs default parameters.

The predicted regions were integrated with experimentally validated BGCs from the MiBIG Database and BGCs identified in the BiG-SLiCE dataset (Kautsar et al. 2021). This combined dataset was utilised for further analysis.

BGC redundancy and clustering

Redundancy within the BGC collection was identified using the linclust command in MMseqs (Hauser, Steinegger, and Söding 2016), with a minimum sequence identity of 0.95 and a coverage of 66% for the query and subject sequences.

Clusters of BGCs based on functional domain composition were identified in the following manner: Regions containing the same set of Pfams were identified and collapsed into one entity (Supplementary Figure 4). Using the Python package NetworkX (https://github.com/networkx/networkx), a similarity network was built where a node represents a set of Pfams, and the weight of the edges is the Dice coefficient (DC), given by

$$\frac{2|X \cap Y|}{|X|+|Y|} \tag{2}$$

for all pairs in the collection, where |XY| is the longest overlapping string between the two sequences and |X|+|Y| is the sum of both sequences' lengths. To reduce the number of edges and computational expense, a K Neighbours Graph was built, with a K parameter equal to 12 (Luxburg and von Luxburg 2007; Brito, Maria R and Chavez, Edgar L and Quiroz, Adolfo J and Yukich, Joseph E 1997). In total, 700,567 sets of Pfams were clustered into 24,795 groups by retaining the lower leaves of the Louvain modularity method using the Python package Communities (https://github.com/taynaud/python-louvain).

BGC clusters graph

For each BGC cluster, a representative BGC was selected based on two criteria: BGCs not overlapping the edge of the contig and the longest BGC. If more than one BGC had the same top priority, one was randomly selected. The DC for each pair of representative BGCs was then calculated (as above), and Cytoscape (Shannon et al. 2003) was used for visualisation, displaying only those edges with a DC above 0.75.

Rarefaction Curve Analysis

Rarefaction curves were constructed for each biome by randomly sampling metagenomic assemblies in 100 incremental steps evenly spaced over the interval between 1 and the number of assemblies in the sample. At each step, the number of BGC clusters was recorded. This process was repeated five times. The plots were generated using matplotlib.

Correlation Analysis of BGCs and BGC Clusters

The relationship between the number of BGCs and the number of BGC clusters per assembly for biomes that exhibited an asymptotic curve in the rarefaction analysis was investigated. Pearson correlation coefficients were computed to assess the strength and direction of the relationship between these variables.

Biome Range Analysis

In order to examine the range of biomes in which BGC clusters were present, each BGC cluster was plotted as a function of the number of biomes in which it occurred and the broad biome category to which those biomes were classified. To enhance visualisation and prevent overlapping of points, jitter was added to the visualisation.

Funding

This project was financially supported by the Biotechnology and Biological Sciences Research Council [grant reference numbers BB/S009043/1; BB/T000902/1; BB/V01868X/1], the EU Horizon project BlueRemediomics [grant reference number 101082304], Innovate UK [grant reference number 101082304], and European Molecular Biology Laboratory core funds. Funding for open access charge: UK Research and Innovation (UKRI).

Competing interests

The authors declare no competing interests.

Acknowledgements

We extend our deepest gratitude to Olivia Casanueva and Lorna Richardson, both from EMBL-EBI, for their diligent proofreading efforts, which have greatly enhanced the clarity and readability of our work. We particularly appreciate Lorna Richardson for her additional support and insightful contributions throughout the development process.

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "TensorFlow: A System for Large-Scale Machine Learning." http://arxiv.org/abs/1605.08695.

Akiyama, Daichi, Keisuke Imoto, Noriyuki Tonami, Yuki Okamoto, Ryosuke Yamanishi, Takahiro Fukumori, and Yoichi Yamashita. 2020. "Sound Event Detection Using Duration Robust Loss Function," June. https://doi.org/10.48550/arXiv.2006.15253.

Baindara, Piyush, Nisha Singh, Manish Ranjan, Nayudu Nallabelli, Vasvi Chaudhry, Geeta Lal Pathania, Nidhi Sharma, Ashwani Kumar, Prabhu B. Patil, and Suresh Korpole. 2016. "Laterosporulin10: A Novel Defensin like Class IId Bacteriocin from Brevibacillus Sp. Strain SKDU10 with Inhibitory Activity against Microbial Pathogens." *Microbiology* 162 (8): 1286–99.

Blin, Kai, Simon Shaw, Alexander M. Kloosterman, Zach Charlop-Powers, Gilles P. van Wezel, Marnix H. Medema, and Tilmann Weber. 2021. "antiSMASH 6.0: Improving Cluster Detection and Comparison Capabilities." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkab335.

Brito, Maria R and Chavez, Edgar L and Quiroz, Adolfo J and Yukich, Joseph E. 1997. "Connectivity of the Mutual K-Nearest-Neighbor Graph in Clustering and Outlier Detection." *Statistics & Probability Letters* 35 (1): 33–42.

Carroll, Laura M., Martin Larralde, Jonas Simon Fleck, Ruby Ponnudurai, Alessio Milanese, Elisa Cappio, and Georg Zeller. 2021. "Accurate de Novo Identification of Biosynthetic Gene Clusters with GECCO." https://doi.org/10.1101/2021.05.03.442509.

Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. 2014. "Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters." *Cell*. https://doi.org/10.1016/j.cell.2014.06.034.

Donia, Mohamed S., Peter Cimermancic, Christopher J. Schulze, Laura C. Wieland Brown, John Martin, Makedonka Mitreva, Jon Clardy, Roger G. Linington, and Michael A. Fischbach. 2014. "A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics." *Cell* 158 (6): 1402–14.

Fuhrman, Jed A., Jacob A. Cram, and David M. Needham. 2015. "Marine Microbial Community Dynamics and Their Ecological Interpretation." *Nature Reviews. Microbiology* 13 (3): 133–46.

Gaber El-Saber Batiha, Diaa E. Hussein, Abdelazeem M. Algammal, Toyosi T. George, Philippe Jeandet, Ali Esmail Al-Snafi, Achyut Tiwari, Jorge Pamplona Pagnossa, Clara Mariana Lima, Nanasaheb D. Thorat, Muhammad Zahoor, Mohamed El-Esawi, Abhijit Dey, Saad Alghamdi, Helal F. Hetta, Natália Cruz-Martins. 2021. "Application of Natural Antimicrobials in Food Preservation: Recent Views." *Food Control* 126 (August): 108066.

Grover, Minakshi, Shrey Bodhankar, Abha Sharma, Pushpendra Sharma, Jyoti Singh, and Lata Nain. 2021. "PGPR Mediated Alterations in Root Traits: Way Toward Sustainable Crop Production." *Frontiers in Sustainable Food Systems* 4 (January). https://doi.org/10.3389/fsufs.2020.618230.

Guo, Jianhua, Jie Li, Hui Chen, Philip L. Bond, and Zhiguo Yuan. 2017. "Metagenomic Analysis Reveals Wastewater Treatment Plants as Hotspots of Antibiotic Resistance Genes and Mobile Genetic Elements." *Water Research* 123 (October): 468–78.

Hannigan, Geoffrey D., David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. 2019. "A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction." *Nucleic Acids Research* 47 (18): e110.

Hauser, Maria, Martin Steinegger, and Johannes Söding. 2016. "MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets." *Bioinformatics* 32 (9): 1323–30.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.

Ibrahim, Maziya, Lavanya Raajaraam, and Karthik Raman. 2021. "Modelling Microbial Communities: Harnessing Consortia for Biotechnological Applications." *Computational and Structural Biotechnology Journal* 19 (July): 3892–3907.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification."

Bioinformatics 30 (9): 1236–40.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.
Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J. J. van der Hooft, Jeffrey A. van Santen, et al. 2020. "MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function." *Nucleic Acids Research* 48 (D1): D454–58.
Kautsar, Satria A., Justin J. J. van der Hooft, Dick de Ridder, and Marnix H. Medema. 2021. "BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters." *GigaScience* 10 (1). https://doi.org/10.1093/gigascience/giaa154.
Kierończyk, Bartosz, Mateusz Rawski, Zuzanna Mikołajczak, Sylwester Świątkiewicz, and Damian Józefiak. 2020. "Nisin as a Novel Feed Additive: The Effects on Gut Microbial Modulation and Activity, Histological Parameters, and Growth Performance of Broiler Chickens." *Animals : An Open Access Journal from MDPI* 10 (1). https://doi.org/10.3390/ani10010101.
Luxburg, Ulrike von, and Ulrike von Luxburg. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing*. https://doi.org/10.1007/s11222-007-9033-z.
Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. 2016. "Metrics for Polyphonic Sound Event Detection." *Applied Sciences*. https://doi.org/10.3390/app6060162.
Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19.
Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2021. "A Genomic Catalog of Earth's Microbiomes." *Nature Biotechnology* 39 (4): 499–509.
Newman, David J., and Gordon M. Cragg. 2020. "Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019." *Journal of Natural Products* 83 (3): 770–803.
Paoli, Lucas, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar, Agneya Bhushan, Alessandro Lotti, et al. 2022. "Biosynthetic Potential of the Global Ocean Microbiome." *Nature* 607 (7917): 111–18.
Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. "GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy." *Nucleic Acids Research* 50 (D1): D785–94.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *The Journal of Machine Learning Research* 12: 2825–30.
Reker, Daniel, and Gisbert Schneider. 2015. "Active-Learning Strategies in Computer-Assisted Drug Discovery." *Drug Discovery Today* 20 (4): 458–65.
Richardson, Lorna, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L. Bileschi, Tony Burdett, Josephine Burgin, et al. 2023. "MGnify: The Microbiome Sequence Data Analysis Resource in 2023." *Nucleic Acids Research* 51 (D1): D753–59.
Sassone-Corsi, Martina, Sean-Paul Nuccio, Henry Liu, Dulcemaria Hernandez, Christine T. Vu, Amy A. Takahashi, Robert A. Edwards, and Manuela Raffatellu. 2016. "Microcins Mediate Competition among Enterobacteriaceae in the Inflamed Gut." *Nature* 540 (7632): 280–83.
Sayed, A. M., M. H. A. Hassan, H. A. Alhadrami, H. M. Hassan, M. Goodfellow, and M. E. Rateb. 2020. "Extreme Environments: Microbiology Leading to Specialized Metabolites." *Journal of Applied Microbiology* 128 (3): 630–57.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.
Singh, Pradip Kumar, Chittpurna, Ashish, Vikas Sharma, Prabhu B. Patil, and Suresh Korpole. 2012. "Identification, Purification and Characterization of Laterosporulin, a Novel Bacteriocin Produced by Brevibacillus Sp. Strain GI-9." *PloS One* 7 (3): e31498.
Singh, Pradip Kumar, Vipul Solanki, Shalley Sharma, Krishan Gopal Thakur, Beena Krishnan, and Suresh Korpole. 2015. "The Intramolecular Disulfide-Stapled Structure of Laterosporulin, a Class IId Bacteriocin, Conceals a Human Defensin-like Structural Module." *The FEBS Journal* 282 (2):

203–14.
Skinnider, Michael A., Nishanth J. Merwin, Chad W. Johnston, and Nathan A. Magarvey. 2017. "PRISM 3: Expanded Prediction of Natural Product Chemical Structures from Microbial Genomes." *Nucleic Acids Research* 45 (W1): W49–54.
Spradlin, Jessica N., Xirui Hu, Carl C. Ward, Scott M. Brittain, Michael D. Jones, Lisha Ou, Milton To, et al. 2019. "Harnessing the Anti-Cancer Natural Product Nimbolide for Targeted Protein Degradation." *Nature Chemical Biology* 15 (7): 747–55.
Tatbul, Nesime, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. "Precision and Recall for Time Series." http://arxiv.org/abs/1803.03639.
Terlouw, Barbara R., Kai Blin, Jorge C. Navarro-Muñoz, Nicole E. Avalon, Marc G. Chevrette, Susan Egbert, Sanghoon Lee, et al. 2023. "MIBiG 3.0: A Community-Driven Effort to Annotate Experimentally Validated Biosynthetic Gene Clusters." *Nucleic Acids Research* 51 (D1): D603–10.
Tietz, Jonathan I., Christopher J. Schwalen, Parth S. Patel, Tucker Maxson, Patricia M. Blair, Hua-Chia Tai, Uzma I. Zakai, and Douglas A. Mitchell. 2017. "A New Genome-Mining Tool Redefines the Lasso Peptide Biosynthetic Landscape." *Nature Chemical Biology* 13 (5): 470–78.
Tsoi, Ryan, Zhuojun Dai, and Lingchong You. 2019. "Emerging Strategies for Engineering Microbial Communities." *Biotechnology Advances*. https://doi.org/10.1016/j.biotechadv.2019.03.011.
Tyc, Olaf, Chunxu Song, Jeroen S. Dickschat, Michiel Vos, and Paolina Garbeva. 2017. "The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria." *Trends in Microbiology* 25 (4): 280–92.
Wolf, Thomas, Vladimir Shelest, Neetika Nath, and Ekaterina Shelest. 2016. "CASSIS and SMIPS: Promoter-Based Prediction of Secondary Metabolite Gene Clusters in Eukaryotic Genomes." *Bioinformatics* 32 (8): 1138–43.
Zhang, Yao, and Alpha A. Lee. 2019. "Bayesian Semi-Supervised Learning for Uncertainty-Calibrated Prediction of Molecular Properties and Active Learning." *Chemical Science* 10 (35): 8154–63.
Ziemert, Nadine, Anna Lechner, Matthias Wietz, Natalie Millán-Aguiñaga, Krystle L. Chavarria, and Paul Robert Jensen. 2014. "Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus Salinispora." *Proceedings of the National Academy of Sciences of the United States of America* 111 (12): E1130–39.

Supplementary Data

Supplementary Table 1

Randomly selected GTDB representative genomes to generate non-BGC regions.

Supplementary Table 2

Detected BGC regions on the manually annotated genomes using the benchmarked methods.

Supplementary Table 3

Origin information of the 637 validated BGCs from the unreleased MiBIG DB v.3.

Supplementary Table 4

Detected BGC regions on the synthetic isolate genome using the benchmarked methods.

Supplementary Table 5

Detected BGC regions on the synthetic metagenomic assembly using the benchmarked methods.

Supplementary FASTAS

FASTA files of synthetic isolate genome and synthetic metagenomic assembly.

Supplementary Table 6

BGCs from the GTDB Representative Genomes Detected Exclusively by SanntiS.

Supplementary Table 7

Selected Examples of BGCs from the GTDB Representative Genomes Detected Exclusively by SanntiS.

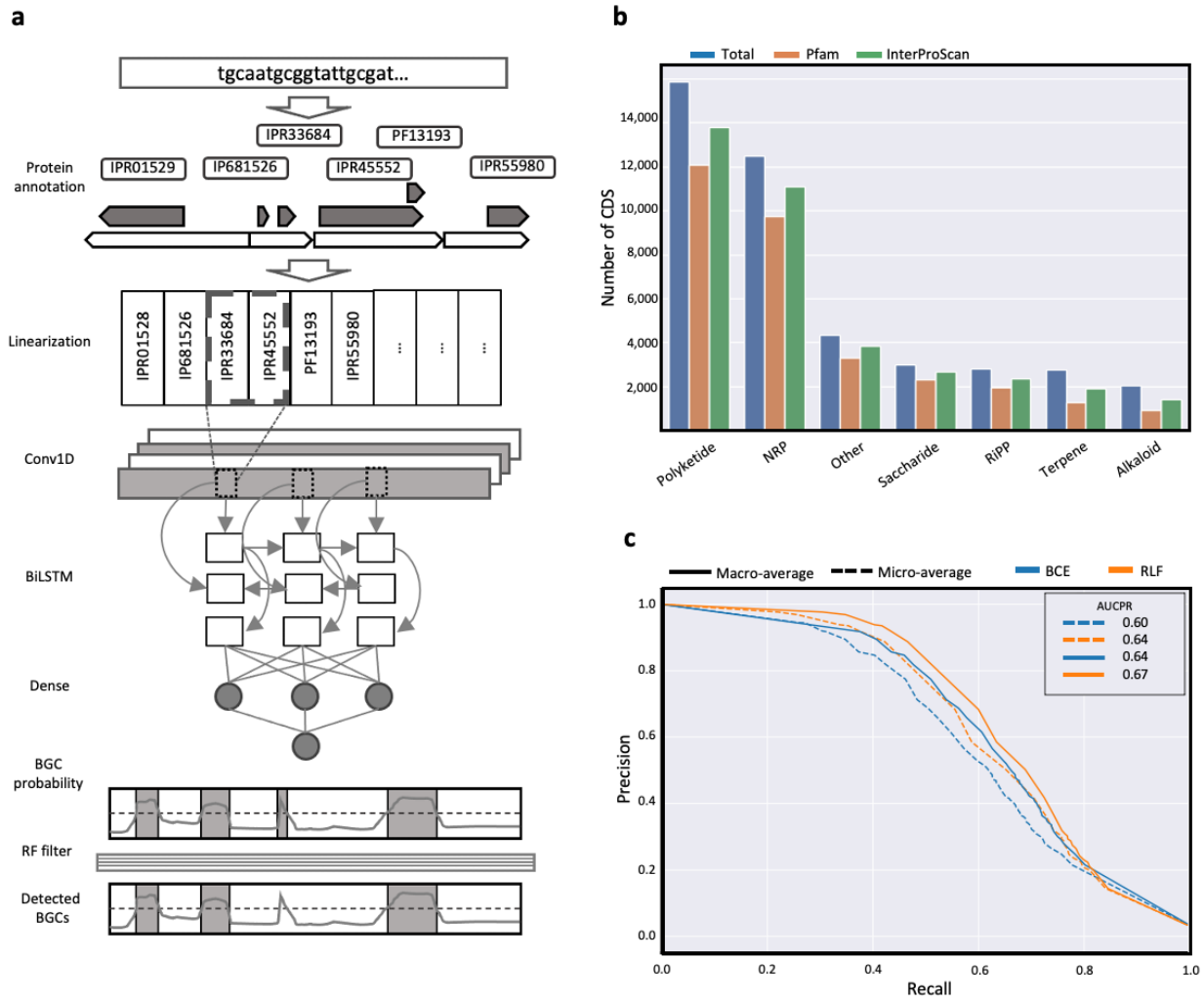
Supplementary Table 8

Summary of detected BGCs in relation to their corresponding GTDB class.

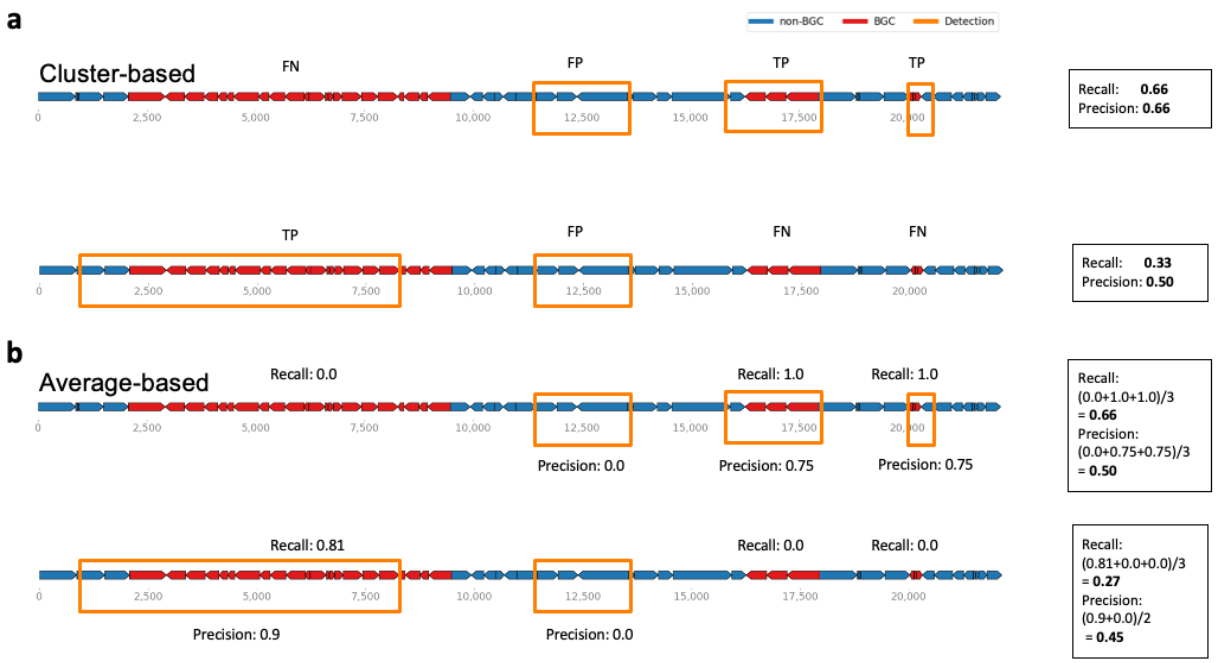
Supplementary Table 9

Summary statistics of BGC prediction in metagenomic assemblies.

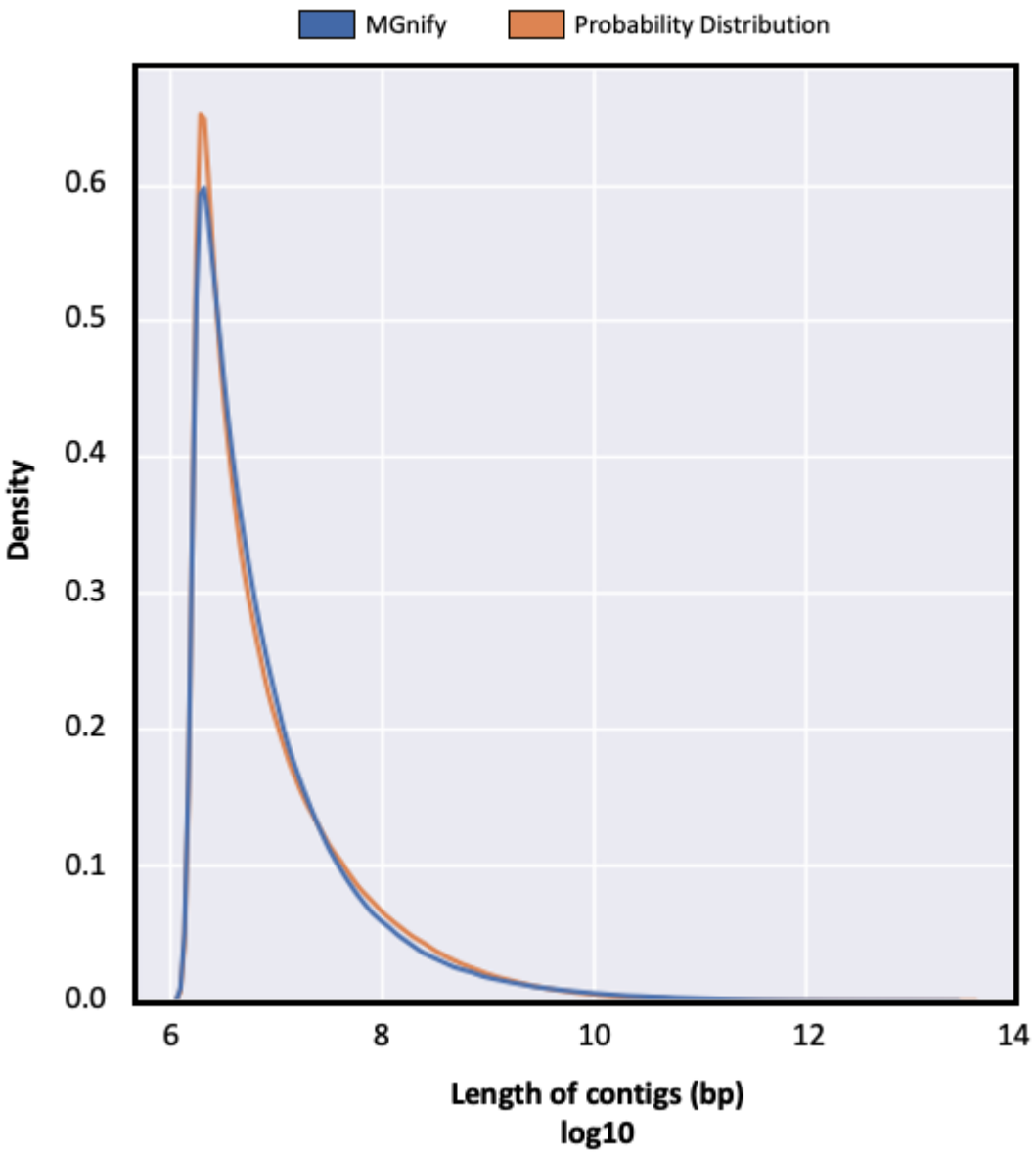
Supplementary Figures



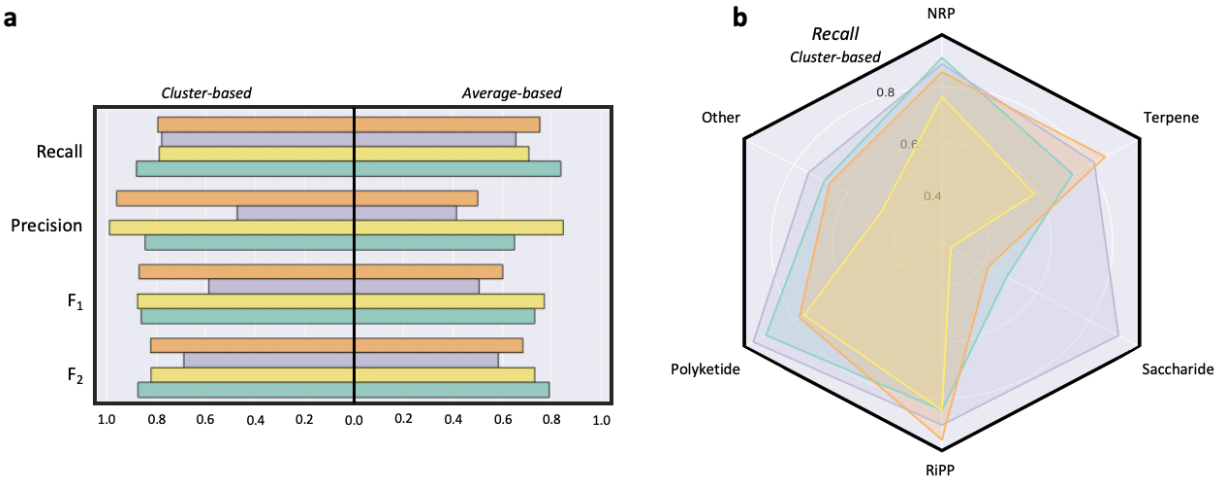
Supplementary Figure 1: SanntiS data flow and unique features of the method. **(a)** Data flow in SanntiS detection method. Nucleotide sequences are used as input, then proteins are predicted to be functionally annotated. The annotations are linearized to go into an artificial neural network (ANN) with a 1D convolutional (Conv1D) layer and a bidirectional long-short term memory (BiLSTM). Regions with a probability above the threshold are filtered by the RF classifier. **(b)** Proteins per BGC class in MiBIG database and their annotation. Blue, total number of proteins. Orange, proteins annotated by Pfam. Green, proteins annotated by the subset of InterProScan. **(c)** 3x5 fold cross-validation precision-recall curves. Colour indicates the loss function used for training. Standard binary cross-entropy loss function (BCE) and the Robust loss function (RLF). Line style indicates the type of metric used.



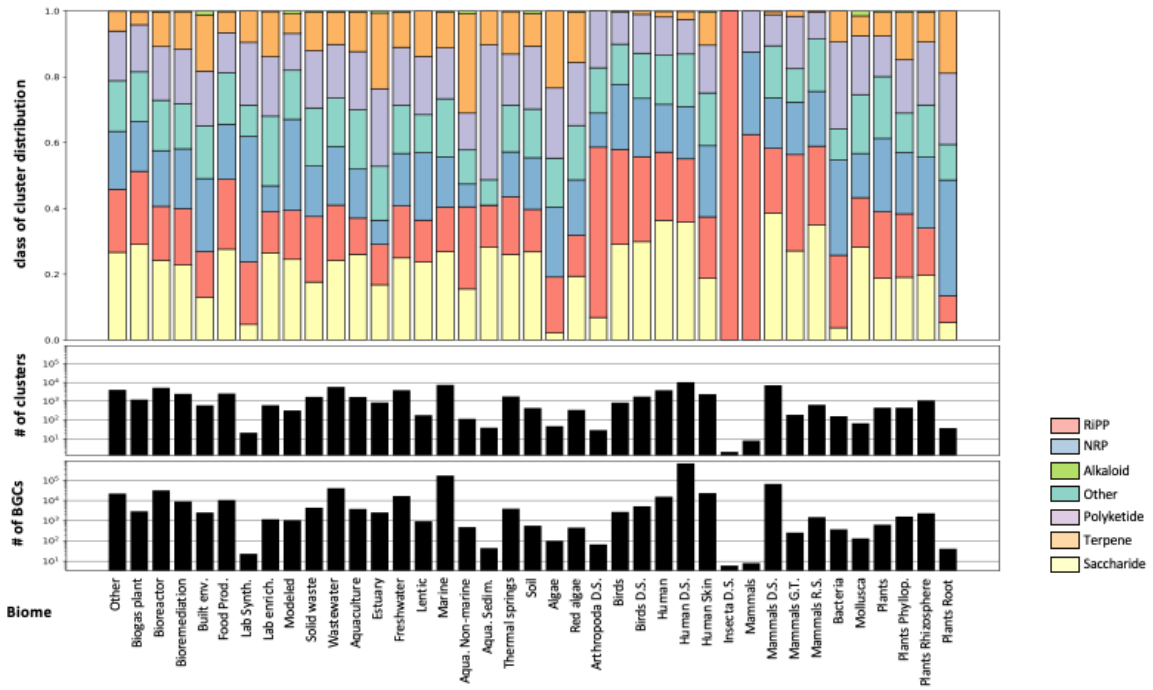
Supplementary Figure 2: Evaluation metrics. Illustration of the hypothetical DNA fragments with coding sequences (CDS) of ground truth BGCs (red) and non-BGC (blue). Yellow boxes represent the predicted BGC regions. **(a)** demonstrates the calculation of metrics using the cluster-based approach, where the overlap of a predicted region with the ground truth BGC is considered as a true positive (TP) and a predicted region not overlapping any ground truth BGC regions is a false positive (FP). **(b)** average-based approach, where precision and recall are calculated for each predicted BGC and ground truth BGC by dividing the number of TP CDS by the total CDSs in the prediction and BGC, respectively, and then averaging the results across all predicted regions. This approach is utilised to avoid ambiguity caused by unbalanced datasets and uneven length distributions of BGCs.



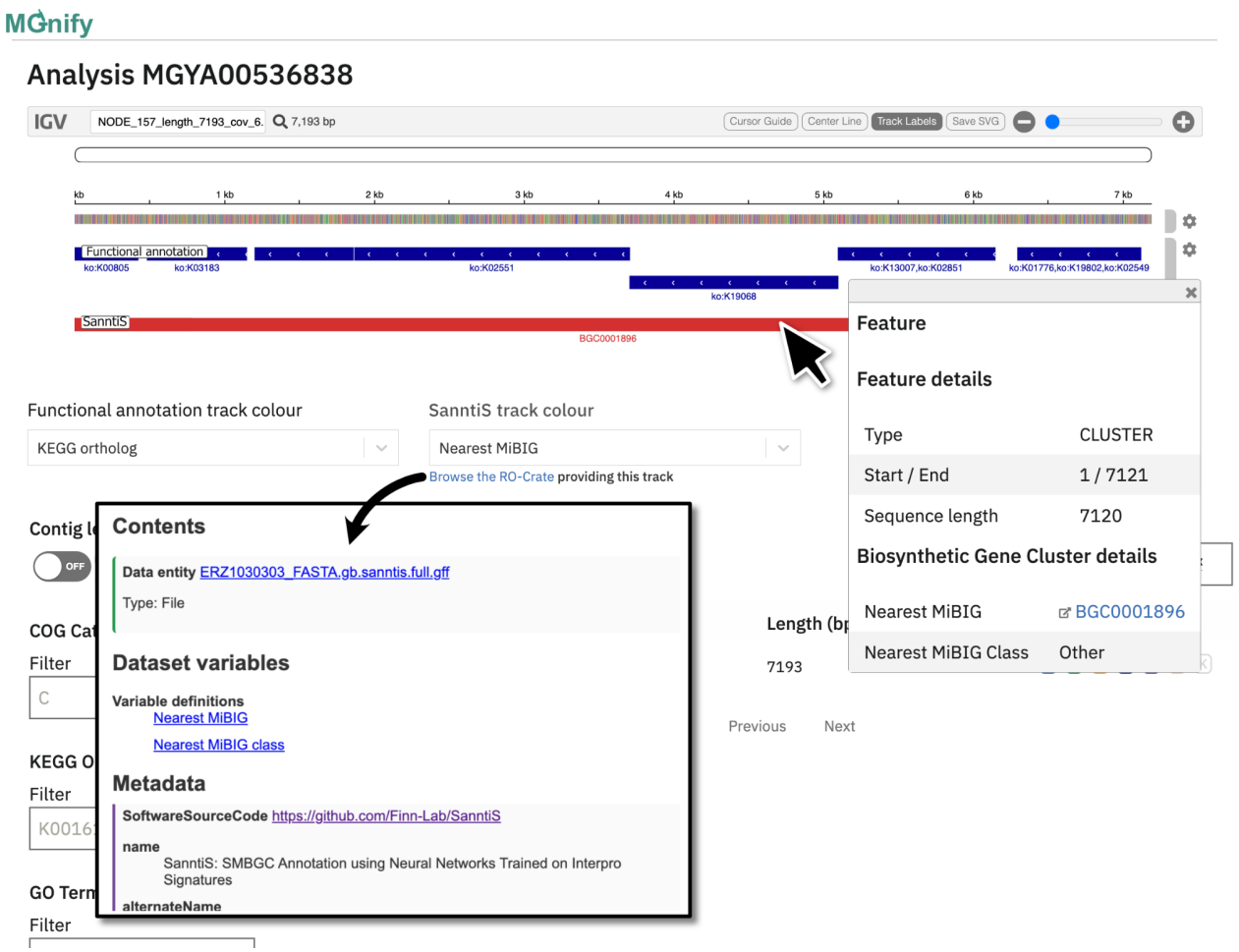
Supplementary Figure 3: Length distribution of metagenomic contigs, presented as a density plot. The x-axis represents the length of contigs in base pairs on a log10 scale, while the y-axis denotes the density. The blue line represents the observed data from MGnify, which illustrates the distribution of metagenomic assemblies from diverse environments. The orange line shows the modelled random probability distribution, which is used to compare and analyse the relationship between the observed data and the expected distribution. This figure helps to provide insights into the contig length distribution in metagenomic datasets and its potential impact on BGC detection methods.



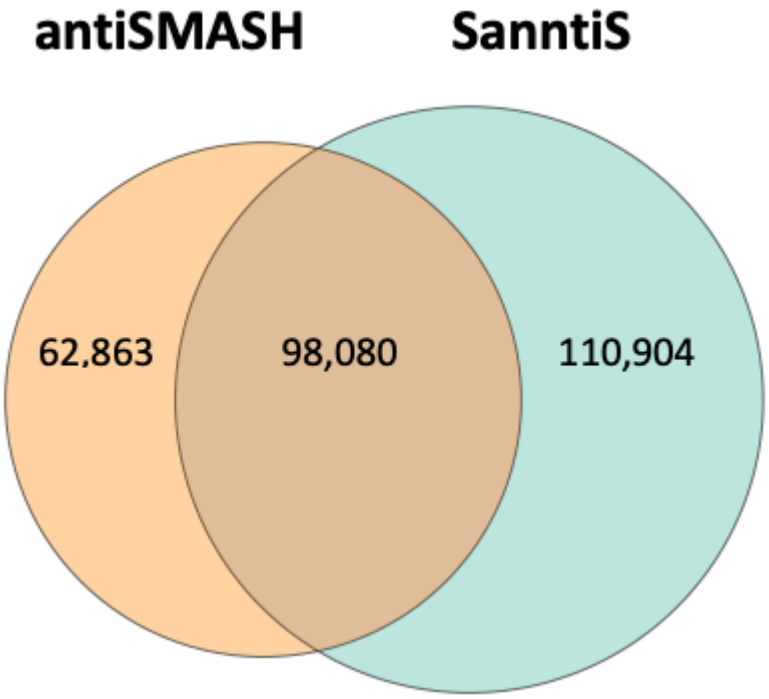
Supplementary Figure 4: Benchmark results**. (a)** Cluster-based and average-based recall, precision, F1 and F2 scores of the compared methods in the synthetic isolate genome. **(b)** Cluster-based recall for each BGC class in the 15 genomes dataset. The following correspond to the number of BGCs in each class: NRP, *n*=119; Other, *n*=43; Polyketide, *n*=107; RiPP, *n*=15; Saccharide, *n*=30; Terpene, *n*=40.



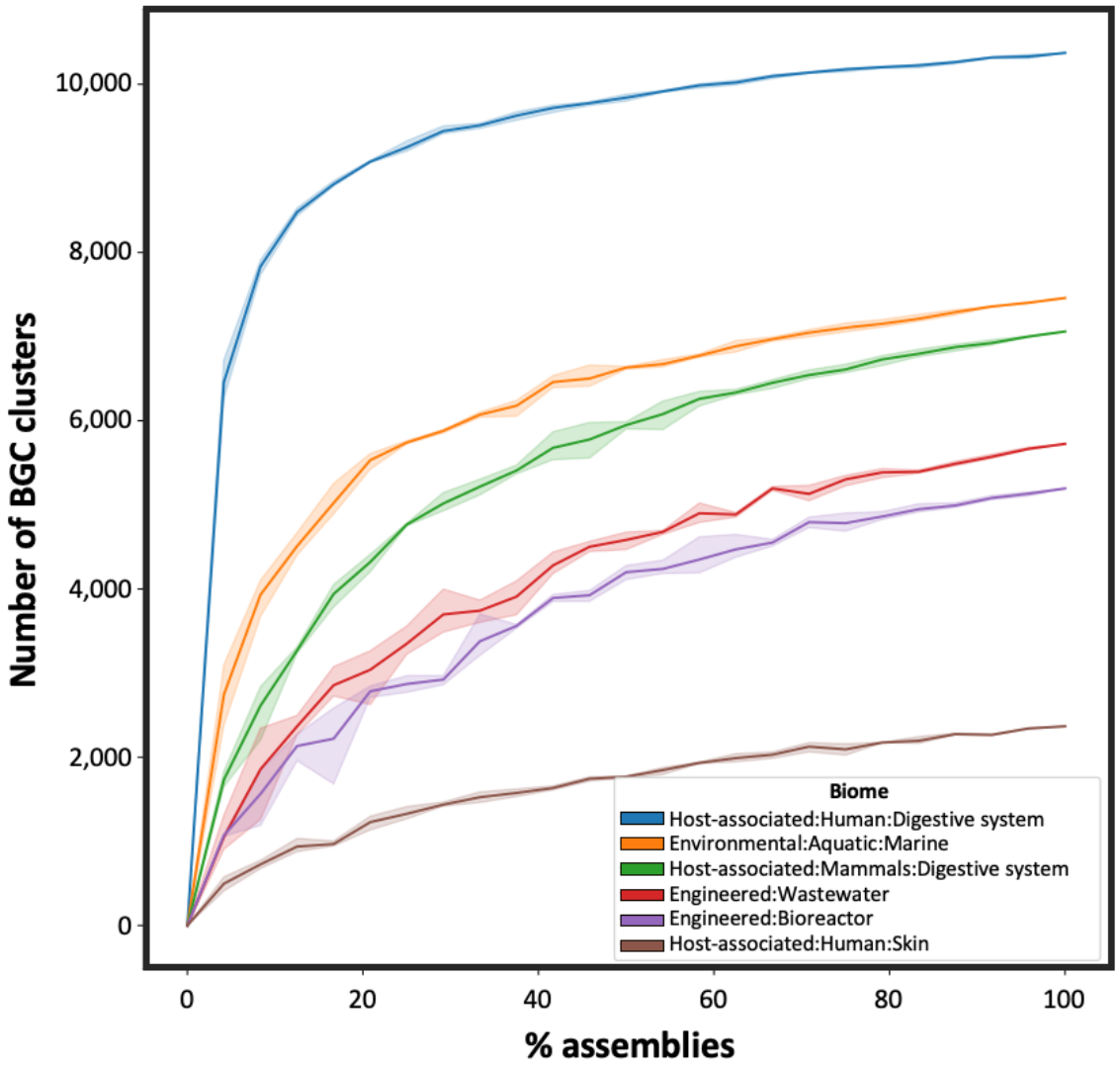
Supplementary Figure 5: MGnify-SanntiS BGC class distributions and biome range of BGC clusters. Bottom, BGC counts per biome. Middle, number of BGC clusters per biome. Top, fraction per class of BGC clusters per biome.



Supplementary Figure 6: Screenshot of MGnify Contig Browser showing in red a region detected as a BGC by SanntiS. Data associated with the prediction is shown in the feature box. The screenshot pertains to assembly accession ERZ1030303 and contig accession NODE_157_length_7193_cov_6.278509.



Supplementary Figure 7: Detection methods intersection, presented as a Venn diagram illustrating the overlap of BGCs detected by antiSMASH (orange) and SanntiS (cyan) on GTDB representative genomes. This figure provides insights into the similarities and differences between the BGC detection capabilities of antiSMASH and SanntiS, highlighting the extent of unique and overlapping predictions for each method.



Supplementary Figure 8: Rarefaction curves of the number of BGC clusters detected as a function of the number of sampled assemblies for each biome. The x-axis represents the percentage of assemblies for each biome, while the y-axis denotes the number of BGC clusters. Only the top 6 biomes are displayed, with the colours of the lines representing: blue, Host-associated:Human:Digestive system; orange, Environmental:Aquatic:Marine; green, Host-associated:Mammals:Digestive system; red, Engineered:Wastewater; purple, Engineered:Bioreactor; brown, Host-associated:Human:Skin.