Sequence modeling tools to decode the biosynthetic diversity of the human microbiome

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

Check for updates

| Computational Biology | Minireview

Sequence modeling tools to decode the biosynthetic diversity of the human microbiome

Mohammed Salim Dason,[1] Davide Corà,[2,3] Angela Re[1]

AUTHOR AFFILIATIONS See affiliation list on p. 19.

ABSTRACT Understanding the biosynthetic potential of the human microbiome remains a significant challenge with far-reaching scientific and translational implications. Analyses of human-associated (meta)genomic sequencing data undeniably show that the biosynthetic diversity encoded in these genomes is largely underexplored. A crucial step in studying specialized metabolites involves the sequence-based identification of genes encoding biosynthetic pathways, typically organized into biosynthetic gene clusters (BGCs). In this review, we provide a concise and updated overview of the widening range of computational approaches that have effectively addressed the sequence-based identification of BGCs across both isolated genomes and complex microbial communities. These advancements are set to deepen our understanding of the biosynthetic potential and diversity of microorganisms residing in different human body sites.

KEYWORDS biosynthetic gene cluster, genome mining, self-supervised learning, hidden Markov model, graph analysis, natural product discovery

The human-associated microbiome is a complex and dynamic consortium of microorganisms residing in and on the human body, fulfilling several crucial roles in maintaining health (1–3). Changes in its functional capacity have been implicated in various pathological states, encompassing metabolic disorders, autoimmune diseases, and neurodegenerative conditions (4, 5). The human-associated microbiome harbors remarkable biosynthetic capabilities, producing a wide array of bioactive molecules, including vitamins, short-chain fatty acids, secondary metabolites, and signaling molecules (6). Among these, secondary metabolites—encoded by the genetic repertoire of human-associated microbial ecosystems (7)—have become a focal point for their therapeutic potential and diverse biological activities.

Advances in sequencing technologies, such as activity-guided genomics, amplicon sequencing, and shotgun metagenomics, have significantly improved our ability to investigate the biosynthetic capabilities of the human microbiome (8), and large-scale initiatives, such as the Human Microbiome Project (HMP), have further accelerated this process (9).

What we have learned from these early studies is that specialized metabolic pathways are often encoded within biosynthetic gene clusters (BGCs), which are distinct genomic loci consisting of two or more co-localized and functionally interconnected genes (Fig. 1) (10). Thus, the systematic identification and functional characterization of these BGCs is set to enhance our understanding of human genetics and biochemistry, leading to the development of new preventive strategies, diagnostic tools, and therapeutics, such as antimicrobials and immunomodulatory agents (8, 11, 12).

The critical role of BGCs is further underscored by their differential representation in health- vs. disease-associated microbiomes (13). Recent progress in annotating gene clusters derived from metagenomic data sets (Table 1) has led to the creation of

Editor Alexander Mahnert, Medizinische Universitat Graz, Graz, Austria

Address correspondence to Angela Re, angela.re@polito.it.

Mohammed Salim Dason and Angela Re contributed equally to this article. Author order was determined both alphabetically and in order of increasing seniority.

The authors declare no conflict of interest.

See the funding table on p. 20.

Published 30 June 2025

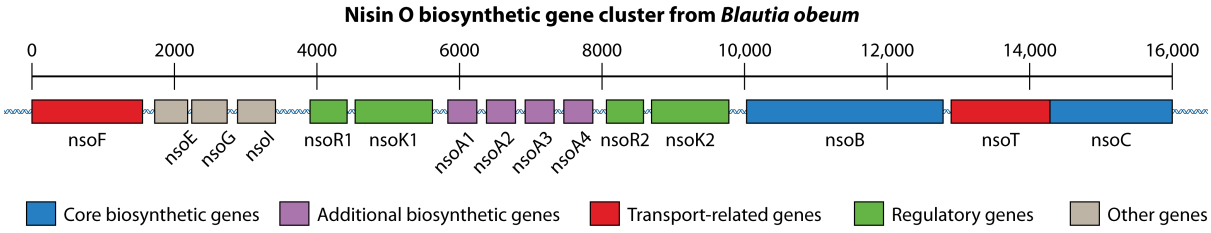


FIG 1 Example of a biosynthetic gene cluster (BGC) encoding Nisin O in *Blautia obeum*. A BGC consists of two or more genes physically clustered on the chromosome of a certain genome, collectively encoding a biosynthetic pathway responsible for the production of a specialized metabolite.

specialized databases cataloging thousands of human-associated BGCs. Notable examples include (i) the Atlas of Biosynthetic gene Clusters in the Human Microbiome (ABC-HuMi) (14), an interactive platform that facilitates navigation through gene clusters inferred using robust genome mining tools and metagenomic data across five distinct human body sites; and (ii) the Atlas of Secondary Metabolite Biosynthetic Gene Clusters from the Human Gut Microbiome (sBGC-hm) (15), a specialized resource that catalog BGCs exclusively derived from the human gut. An overview of these and other generalist and human-associated BGC databases is available in Table 2, which underscores how the diversity among BGC classes and their associated natural products mirrors the substantial taxonomic variability inherent to the human microbiome across different physiological and pathological conditions (16, 17). Intriguingly, a growing body of evidence indicates that a significant number of identified BGCs remains functionally uncharacterized (12, 13).

Current efforts to discover and catalog BGCs in the human microbiome mostly rely on traditional heuristics derived from the accumulated knowledge of specialized metabolic pathways, such as peptide modifications, acyltransferase activities, and adenylation domain substrate specificities. These approaches are often supported by supervised machine learning models, which are trained on heuristic-based data sets to identify BGCs. However, the emergence of protein language models (67, 68) has led to a paradigm shift in this field. These models inherently learn the structural and functional constraints encoded in the billions of protein sequences they are exposed to. The information derived from these models can then be represented as mathematical embeddings, enabling their application to specific tasks of interest, such as the accurate prediction and functional annotation of BGCs.

Against this backdrop, the present review provides a synopsis of both established and emerging methodologies for mining the human microbiome in the context of gene clusters responsible for synthesizing secondary metabolites (Fig. 2). We begin by detailing the primary BGC classes and their defining biosynthetic traits. To further elucidate these concepts, we then examine the BGC classes identified to date within human microbiome genomes and metagenomes, providing meaningful examples of biologically active compounds with relevant functional activities. Lastly, we analyze computational sequence-based tools designed for BGC detection, focusing on their algorithmic frameworks, the implications on inferred gene clusters, their suitability for analyzing single genomes and/or metagenomes, and their ability to characterize previously unclassified BGC classes.

Overall, this review provides researchers with a practical guide to the fast-evolving field of microbiome-derived natural product discovery, underscoring the need for better computational tools capable of integrating multi-omic data.

BGC CLASSES

The Minimum Information about the Biosynthetic Gene Cluster (MIBiG) serves as a reference repository for manually curated data on secondary metabolites, currently exceeding 3,000 entries. Among these, roughly 50% are associated with over 3,600 documented biological activities, and more than 85% are linked to 5,000 distinct

TABLE 1 Overview of tools for both class-specific and broad-spectrum BGC detection[a]

Class	Software tool	Purpose	Algorithmic features	Reference
Generalist	BGCFlow	BGC identification; BGC organization and visualization	Snakemake-based multi-functional BGC analyzer including in-house and third-party analytics for data selection (GTDB-Tk, MASH or fastANI-based genomic distance), functional annotation (Prodigal, eggNOG-mapper), phylogenomic placement (autoMLST), BGC genome mining (antiSMASH, GECCO, and ARTS2), BGC comparative analysis (BiG-SLICE, BiG-SCAPE, BiG-FAM, and Roary)	18
Generalist	lsaBGC (Lineage-specific analysis of BGCs	BGC genome identification; BGC clustering; gene cluster family (GCF) exploration; evolutionary and population genetic analysis of biosynthetic genes	Analytical workflow including in-house and third-party analytics such as Prodigal, KOfam (a customized Hidden Markov Models [HMM] database of KEGG Orthologs) and PGAP (NCBI Prokaryotic Genome Annotation Pipeline) HMMs for gene calling and annotation, antiSMASH, GECCO, and DeepBGC for BGC identification, OrthoFinder2 for phylogenetic orthology inference, Markov chain clustering for BGC clustering into GCFs, custom HMM-based algorithm for GCF identification, GToTree for genome-level evolutionary inference, and gene-based metagenomic analysis	19
Generalist	ClusterFinder	BGC identification	Two-state HMM-based probabilistic algorithm (with hidden states representing BGC and non-BGC)	20
Generalist	antiSMASH (antibiotics and Secondary Metabolite Analysis SHell) 7.0	BGC identification	Two-state HMM-based probabilistic algorithm; BCG class-specific functionalities: trans-AT PKSs-specific profile HMMs; NRPS adenylation domain substrate specificity prediction by NRPyS; CompaRiPPson analysis for novelty assessment in ribosomally synthesized and post-translationally modified peptide (RiPP) precursor peptides	21
Generalist	GECCO (GEne Cluster prediction with COnditional random fields)	BGC identification	Conditional random forests including a feature selection approach based on the two-sided Fisher's exact test to identify domains associated with BGC presence/absence	22
Generalist	Deep-BGCpred	BGC identification	Stacked bidirectional long short-term memory (Bi-LSTM) neural network for biosynthetic genes prediction combined with random forest multi-label classifier for false positive reduction and BGC class assignment	23
Generalist	BiGCARP (Biosynthetic Gene Convolutional Autoencoding Representations of Proteins)	BGC identification	Embedding of Pfam domain sequences with the ESM-1b protein masked language model; BiGCARP architecture consists of a dilated 1D-convolutional neural network masked language model based on ByteNet and CARP	24
Generalist	TaxiBGC (Taxonomy-guided Identification of Biosynthetic Gene Clusters)	BGC identification	MetaPhlAn3-based species-level taxonomic profiling of shotgun metagenomic sequence data; Minimum Information about the Biosynthetic Gene Cluster (MiBiG) query with identified species; sequence aligner for MiBiG BGC confirmation in metagenomic sequences	25
Generalist	DeepBGC	BGC identification; NP[b] classification	Bi-LSTM recurrent neural network; word2vec-like word embedding skip-gram neural network (pfam2vec); random forest multi-label classifier for BGC product class prediction	26
Generalist	PRISM (PRediction Informatics for Secondary Metabolomes)	BGC identification; prediction of NP structure	Profile HMMs for enzymatic domains identification; greedy algorithm for BGC identification; combinatorial linear scaffold-based approach for NP structure prediction; BGC/NP dereplication	27
Generalist	PRISM 3	BGC identification; prediction of NP structure	Profile HMMs for enzymatic domains identification; chemical graph-based approach for structure prediction	28
Generalist	PRISM 4	BGC identification; prediction of NP structure	Profile HMMs, conserved protein motifs, and machine-learning classifiers for enzymatic domains identification; combinatorial chemical graph-based approach for structure prediction	29
Generalist	NPOmix (Natural Products Mixed Omics)	MS-guided BGC identification	K-nearest neighbor algorithm based on similarity BGC fingerprints and similarity MS/MS fingerprints to classify GCFs for each MS/MS spectrum; antiSMASH and BiG-SCAPE for BGC and GCF discovery; GNPS-based cosine score for MS/MS spectra similarity computation	30
Generalist	NPLinker	MS-guided BGC/GCF identification	BGCs clustering into GCFs by BiG-SCAPE; GNPS-based spectral clustering of MS/MS spectra into molecular families (MFs); feature-based approach to link BGCs to MS/MS spectra via input-output kernel regression or correlation-based approach to link GCFs to MFs or MS/MS spectra or the combination of correlation- and feature-based approaches	31
Class specific	TrRiPP	RiPP BGC identification	Transformer encoder generating the input of two layers of Bi-LSTM; concatenated maximum, mean, and last outputs from the last Bi-LSTM layer are subjected to a feed-forward network consisting of two dense layers with a rectified linear unit as activation function to carry out RiPP classification	32

(*Continued on next page*)

TABLE 1 Overview of tools for both class-specific and broad-spectrum BGC detection[a] (*Continued*)

Class	Software tool	Purpose	Algorithmic features	Reference
Class specific	Pep2Path	MS-guided NRP/RiPP BGC identification	Bayesian algorithm NRP2Path matching short amino acid sequence tags to BGC-encoded NRPS assembly lines and employing colinearity index computation to interpret multiple BGC assignments; RiPP2Path matching putative RiPP amino acid sequence tags to the translation frames of a set of (meta)genomic sequences for precursor peptide identification	33
Class specific	ARTS (Antibiotic Resistant Target Seeker) 2.0	Antibiotic-resistant gene identification	antiSMASH-based detection of BGCs for resistance gene identification in (meta)genomic sequences; TIGRFAM protein model for core gene detection; resistance gene and core gene screening based on physical proximity, and detection of gene duplication and horizontal gene transfer events	34
Class specific	BAGEL4 (BActeriocin GEnome mining tooL)	Bacteriocin identification	Protein motif search BLAST against core peptides in the bacteriocin database; Glimmer-based ORF call; BLAST against annotation database; TransTermHP terminator prediction; motif-based promoter prediction	35
Class specific	RODEO (Rapid ORF[b] Description and Evaluation Online)	Lasso peptide BGC identification	Profile HMM local genomic analysis; precursor peptide/structure prediction by a combination of heuristic scoring, motif analysis, and SVM[b] classifier	36
Class specific	RiPPquest	MS-guided lanthipeptide (RiPP category) BGC identification	Pfam domain-based lanthipeptide gene clusters identification; prediction of MS/MS lanthipeptide spectra corresponding to core peptides predicted according to biosynthetic transformations and gas phase reactions in lanthipeptides; matching between predicted MS/MS lanthipeptide spectra and MS/MS spectra obtained by microbial extracts analysis; peptide homologs identification by spectra alignment	37
Class specific	DeepRiPP	MS-guided RiPP BGC identification	Bipartite algorithms adapted from natural language processing for identification of precursor peptides and their cleavage prediction (NLPPrecursor); Basic Alignment of Ribosomal Encoded Products Locally (BARLEY) combining retrobiosynthetic processing of known RiPP structures with local alignment to genomic sequences for candidate RiPP novelty computation; Computational Library for Analysis of Mass Spectra (CLAMS) integrating MS information to identify putative RiPPs within metabolomics datasets	38
Class specific	evoMining 2.0	NP biosynthetic enzyme identification	BLASTP-based search for expansion and recruitment events in enzyme families; phylogenies inference by approximately maximum-likelihood trees for assignment of metabolic origin and fate to enzyme family members	39
Class specific	decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER)	RiPP BGC identification	Support vector machine classifier for RiPP precursor identification; pan-genomic analyses for assessment of their location within operon-like structures encoding accessory genes in a genus; evolutionary conservation- and enzymatic novelty-based ranking of precursors	40
Class specific	RiPPER (RiPP Precursor Peptide Enhanced Recognition)	RiPP BGC identification	RODEO-based identification of putative RiPP tailoring enzymes (RTEs); prodigal-short-based evaluation of the peptide-coding potential of regions surrounding RTEs; peptide optional similarity network approach for RiPP families identification	41
Class specific	NeuRiPP (Neural network identification of RiPP precursor peptides)	RiPP BGC identification	Deep neural network classifier trained on high-confidence ribosomally encoded precursor peptide sequences	42
Class specific	RiPP-PRISM	RiPP BGC identification	Profile HMM- and motif-based identification of RiPP BGCs, prediction of precursor peptide cleavage events, virtual reconstruction of post-translationally modifying reactions for combinatorial structure prediction across RiPP families	43
Class specific	MetaMiner	MS-guided RiPP BGC identification	Profile HMM-based identification of RiPP BGCs and their precursor peptides; target (via HMMer) and decoy putative RiPP structure databases construction; dereplicator-based matching between MS/MS spectra and the constructed target/decoy RiPP structures; spectral alignment for streamlining the peptide modifications identification	44
Class specific	RiPPMiner	RiPP BGC identification; prediction of NP structure	Support vector machine classifiers for prediction of RiPP class and cleavage sites; support vector machine or random forest classifiers for crosslinks prediction	45
Class specific	RiPPMiner-Genome	RiPP BGC identification; prediction of NP structure	Profile HMM-based RiPP BGC identification; random forest classifiers and support vector machine classifier for the prediction of precursor peptides, leader cleavage sites and crosslinks	46

(*Continued on next page*)

TABLE 1 Overview of tools for both class-specific and broad-spectrum BGC detection[a] (*Continued*)

Class	Software tool	Purpose	Algorithmic features	Reference
Class specific	SANDPUMA (Specificity of Adenylation Domain Prediction Using Multiple Algorithms)	Non-ribosomal peptides (NRPS) adenylation domain specificity	Ensemble method based on active site motif sequence signatures, support vector machines, profile hidden Markov models, phylogenetically driven algorithm (prediCAT) calculating a confidence score for each A-domain based on comparative metrics against A-domains of known specificity	47
Class specific	NRPSPredictor2	NRPS adenylation domain specificity	Transductive support vector machines based on sequence and structural information about the active site of the adenylation domain	48
Class specific	CLUSEAN (CLUster SEquence ANalyzer)	Polyketide (PK)/NRP BGC annotation	Analytical workflow built on basic annotation (BLAST), protein domain identification against generalist profile databases (Pfam) and specialized databases for the identification of domains and conserved motifs of PKS/NRPS enzymes, classification of C-domain types and prediction of the specificity of NRPS adenylation domains	49
Class specific	BiosyntheticSPAdes	PK/NRP BGC identification	SPAdes- or metaSPAdes-based (meta)genomic assembling; identifying domain edges in the assembly graph; BGC subgraph extraction from the assembly graph; restoring collapsed domains in the assembly graph; constructing the scaffolding graph; constructing putative BGCs by solving the Rural Postman problem in the scaffolding graph	50
Class specific	NRPminer	MS-guided NRP BGC identification	antiSMASH for NRPS BGC prediction; NRPSpredictor2 for prediction of amino acid for each A-putative NRP structures domain; delineation of NRPS assembly lines accounting for modification enzymes; construction of putative NRP structures; matching between predicted NRPs and experimental spectra with score assignment	51
Class specific	Nerpa	Structure-guided NRP BGC identification	Retro-biosynthesis-based transformation of input structures into monomer graphs; generation of linear representations of monomer graphs; antiSMASH-based NRPS BGC prediction; BGC processing and heuristics-based generation of BGC monomer sequence according to collinear or non-collinear NRPS assembly lines; global alignment of NRP and BGC monomer sequences	52
Class specific	GRAPE (Generalized Retrobiosynthetic Assembly Prediction Engine)	Structure-guided PK/NRP assembly prediction	Retro-biosynthesis of PK and NRP by macrocycles opening, heterocycles opening, monomer linkages removal, identification of tailored additions, and PK processing	53
Class specific	SBSPKS (Structure Based Sequence analysis of PKS and NRPS)	Structure-guided PK/NRP BGC identification	Search for PKs and NRPs chemically similar to query molecule in SMILES format; search for tailoring reactions; linking chemical and genomic space	54
Class specific	GARLIC	Structure-guided PK/NRP BGC identification	Global alignment between monomers from antiSMASH-based BGC cluster predictions and GRAPE-based small molecule breakdowns (fatty acyl units, sugars, amino acids, and carboxylic acids) by random sampling of permutations	53
Miscellaneous	BiG-MAP (Biosynthetic Gene cluster Meta'omics Abundance Profiler)	BGC abundance and expression profiling	Mutual sequence similarity-based redundancy filtering on predicted BGCs (e.g., antiSMASH); BiG-SCAPE for BGCs clustering into GCFs; Bowtie2-based read mapping to non-redundant BGCs; differential abundance/expression analysis by parametric zero-inflated Gaussian distribution mixture model (ZIG model) or non-parametric Kruskal-Wallis test	55
Miscellaneous	BIG-SLiCE (Biosynthetic Genes Super-Linear Clustering Engine)	BGC clustering	BGC vectorization for the transformation of input BGCs into numerical feature vectors based on boolean values and bit scores of hits obtained querying BGC gene against Pfam and sub-Pfam profile hidden Markov models; superlinear clustering of BGCs into GCFs	56
Miscellaneous	BIG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine)	BGC clustering	Exploration of antiSMASH or MIMiB BGC sequence similarity networks based on combined metrics (Jaccard index, adjacency index, and domain sequence similarity index); metrics weights calibration for accounting for different evolutionary modes of BGC classes; affinity propagation clustering algorithm	57
Miscellaneous	CAGECAT (CompArative GEne Cluster Analysis Toolbox)	BGC comparative analysis	Search module relying on cblaster, which utilizes remote BLAST searches and accelerated local hidden Markov models; visualization module relying on the clinker pipeline	58
Miscellaneous	ModulesDetection	BGC module detection	Biosynthetic module identification in antiSMASH-predicted BGCs by orthoMCL-based evolutionary analysis and network analysis where nodes are clusters of orthologous genes, and edges are drawn based on adjacency and colocalization interactions	59
Miscellaneous	CORASON (CORe Analysis of Syntenic Orthologs to prioritize Natural product biosynthetic gene clustersBGC)	BGC phylogenetic analysis	Homology analysis of query genes in input BGCs (via antiSMASH and MIBiG); identification of the core genes in the genomic contexts harboring the identified homologous genes; multi-locus approximate-maximum-likelihood phylogenetic tree for prediction of clades synthesizing structurally different molecular products	57

(*Continued on next page*)

TABLE 1 Overview of tools for both class-specific and broad-spectrum BGC detection[a] (*Continued*)

Class	Software tool	Purpose	Algorithmic features	Reference
Miscellaneous	NaPDoS2 (Natural Product Domain Seeker)	PK/NRP BGC phylogenetic analysis	Phylogeny-based classification scheme of ketosynthase and condensation domains	60
Miscellaneous	LASSOHTP	Lasso peptide structure prediction	Translation of input lasso peptide sequences and annotations (ring, loop, and tail) into conformational ensembles by scaffold construction, random mutagenesis, and molecular dynamics	61

[a]For each software tool, the table reports the purpose(s) and a summary of its algorithmic components.
[b]SVM, support vector machine; ORF, open reading frame; NP, natural product; MS, mass spectrometry.

TABLE 2 Overview of BGC-related databases[a]

BGC database	BGC type	General features	Reference
MIBiG (Minimum Information about a Biosynthetic Gene cluster) v.3.0	Experimentally validated BGC database	Accession mode: web interface. Query options: simple query by keyword; complex query builder. Construction method: manual curation. Additional notes: researchers are enabled to submit new BGC data to the database, adhering to MIBiG standards.	62
antiSMASH Database v.4.0	Computationally predicted BGC database	Accession mode: web interface, API. Query options: query by ribosomally synthesized and post-translationally modified peptide precursor based on NCBI BLAST+; search by protein sequence based on DIAMOND; search by NRPS/PKS module; search by single/multiple categories. Construction method: automated prediction of genomes obtained from NCBI RefSeq database by antiSLASH (v.7.1)	63
IMG-ABC (Integrated Microbial Genomes Atlas of Biosynthetic gene Clusters) v.5.0	Computationally predicted and experimentally validated BGC database	Accession mode: web interface. Query options: query by compound name, genome or accession ID, secondary metabolite ID, BGC ID or gene ID, comment, collection, taxonomy search, sequence similarity, and complex query builde. Construction method: automated prediction by antiSMASH v.5.0 and MiBIG content integration. Additional notes: researchers enabled to submit new BGC data to the database through specified system	64
ABC-HuMi (Atlas of Biosynthetic gene Clusters in the Human Microbiome)	Computationally predicted BGC database	Accession mode: web interface. Query options: query by user-provided nucleotide sequences based on BLAST+; query by translated nucleotide sequences based on tBLASTn; cblaster enabling the search of custom BGCs; filtering by metadata (taxa, body site, and product). Construction method: automated prediction of BGCs from MAGs from JGI GEM, isolates and MAGs from EMBL-EBI MGnify Catalogs and metagenomes from HMP using antiSMASH v.7.0	14
BIG-FAM	Computationally predicted GCF database	Accession mode: web interface. Query options: multi-criterion GCF search and GCF annotation of user-supplied BGCs. Construction method: automated prediction of GCFs by BiG-SLICE clustering of BGCs predicted by antiSAMH v5.1.1	65
sBGC-hm	Computationally predicted BGC database	Accession mode: web interface. Query options: query by cluster, family, organism, Genbank accession, and type. Construction method: automated prediction of BGCs from genomes obtained from the HumGut database and the HMP using antiSMASH v6.1.1	15
BGC Atlas	Computationally predicted BGC database	Accession mode: web interface. Query options: query by user-provided BGC sequence. Construction method: automated prediction of BGCs from genomes obtained from the MGnify database using antiSMASH v6.1.1 and subsequent clustering using BiG-SLiCE	66

[a]For each BGC-related database, the table reports the experimental and/or computational origin of the included BGCs, accession modes, query options, and a summary of its construction method.

chemical structures (62). In this context, this review focuses on the predominant BGC classes found within the human microbiome, as cataloged in specialized databases (14, 15), and narrows the discussion to key biosynthetic traits (Fig. 3A and B) necessary to understand their functional relevance and the principles inspiring supervised BGC mining tools.

Non-ribosomal peptides (NRPs) constitute an important class of secondary metabolites synthesized by non-ribosomal peptide synthetases (NRPSs), which are large enzymatic complexes distinct from ribosomes (69). NRPSs function as modular assembly lines, where each module adds a specific amino acid to the growing peptide chain. Each NRPS module typically consists of at least three core domains: (i) the adenylation (A) domain, responsible for selecting and activating the amino acid to be lined up (70); (ii) the thiolation (T) domain, which allows the activated amino acid to travel within and between modules; and (iii) the condensation (C) domain, which catalyzes the peptide bond formation between the amino acid and the elongating chain (71). Because they are ribosome independent, NRPSs can introduce not only proteinogenic but also non-proteinogenic amino acids into the produced peptides. Post-assembly modifications by specialized tailoring enzymes further increase this diversity. The remarkable variability in NRPs is thought to arise from (i) the modular architecture of NRPSs, which allows autonomous operation of each module and combinatorial amino acid incorporation; (ii) the broader substrate range of monomers compared to ribosomal peptides; and (iii) the extensive peptide modifications during and after chain assembly.

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are initially synthesized as precursor peptides, which are typically larger than the final

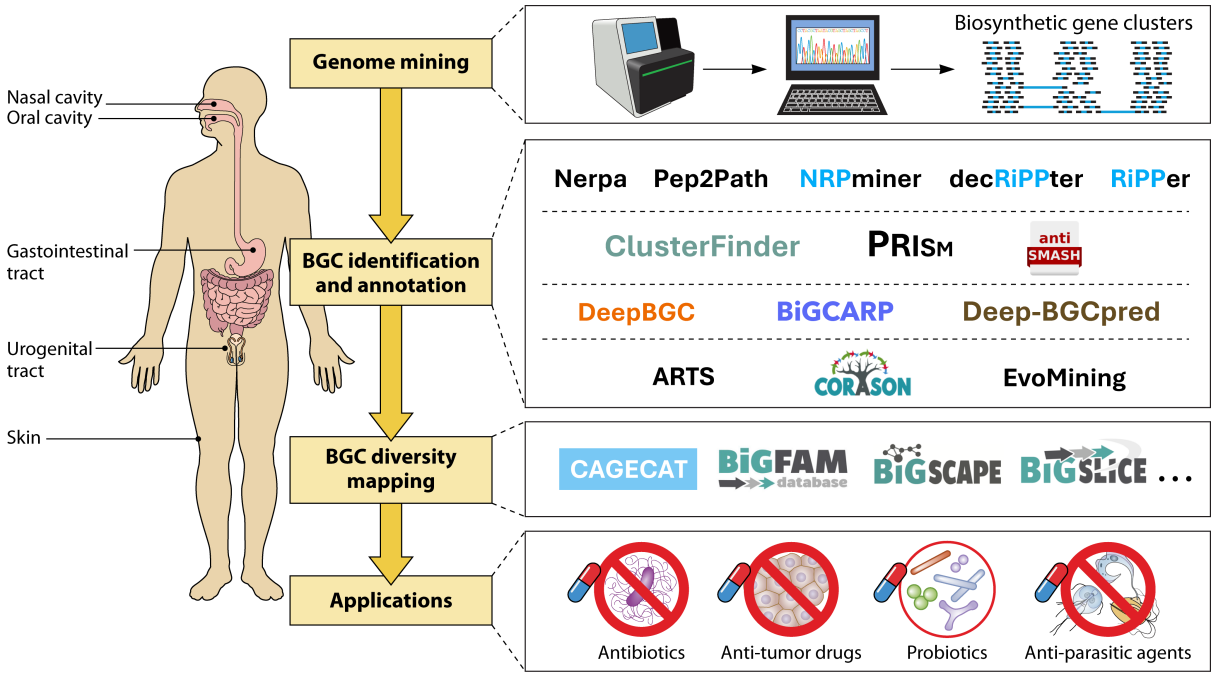


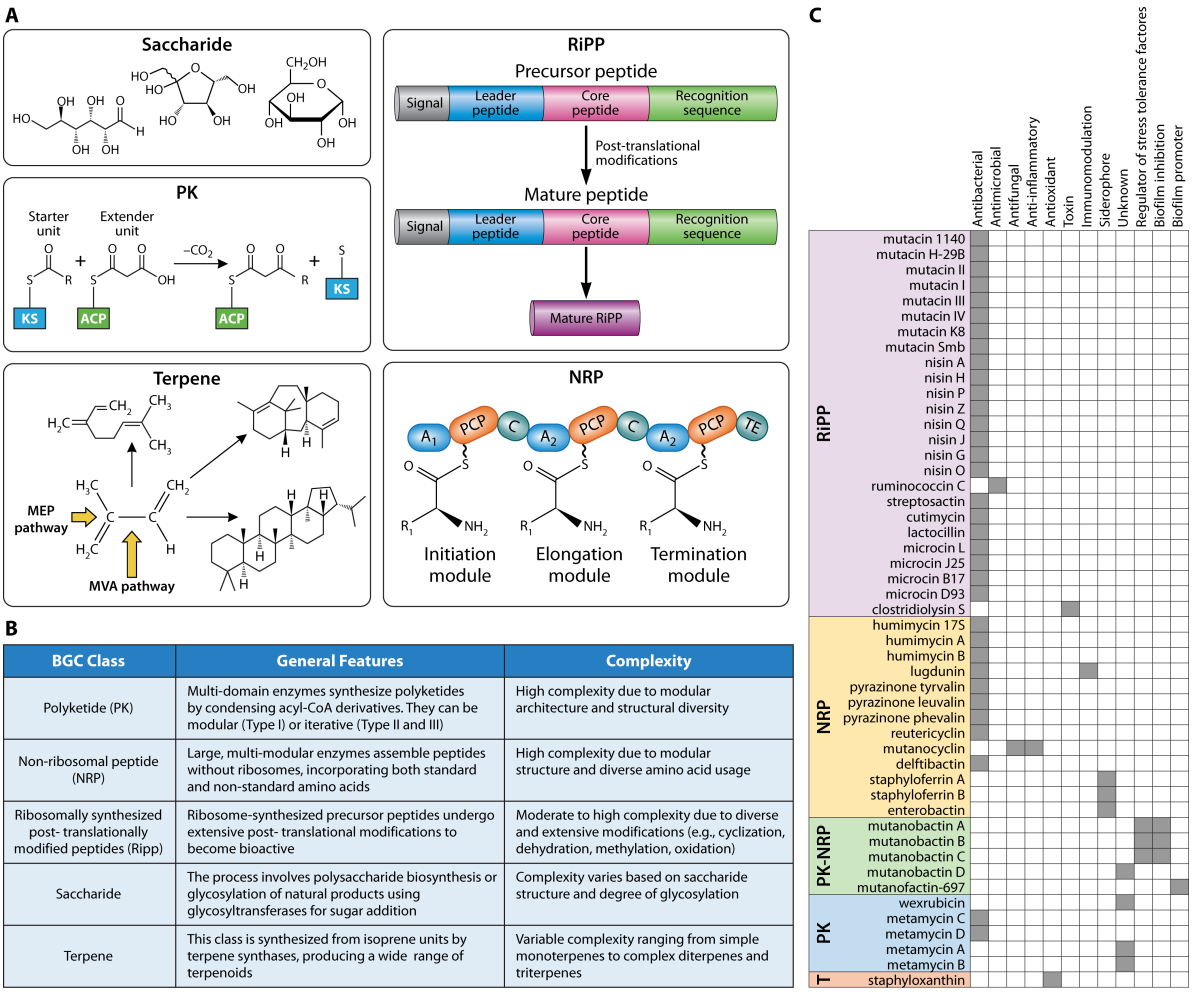
FIG 2 Decoding the biosynthetic diversity of the human microbiome through sequence-based modeling. Outline of sequence-based software tools for the characterization of BGCs endowed with applicative interest in (meta)genomes.

products resulting from the maturation process. The modifying processes release the precursor peptides from the structural and functional constraints characteristic of ribosomal products while restricting conformational flexibility to enhance target recognition and increase stability (72). In general, a precursor peptide consists of an N-terminal leader peptide and a C-terminal core peptide (73). Even though the roles of the leader peptide can vary depending on the RiPP sub-class, its primary functions include facilitating the recognition of the precursor peptide by the RiPP post-translationally modifying enzymes via the RiPP precursor recognition element (RRE) and promoting the export of the RiPP out of the cell (74).

The core peptide undergoes post-translational modification by the RiPP biosynthetic machinery, is proteolytically cleaved from the leader peptide to yield the mature RiPP, and is subsequently exported out of the cell by transporters. RiPPs encompass around 40 sub-classes (75), including polycyclic peptide antibiotics, bacteriocins, cytolysins, amatoxins, cyclotides, microviridins, and conopeptides (74). These examples point to the significant chemical diversity among RiPPs, underscoring their potential for translational applications.

Polyketides (PKs) are versatile in structure and function (76). Examples of PK compounds that have entered the drug market include erythromycin, known for its antimicrobial properties, rapamycin, an immunosuppressant, and doxorubicin, used as an anticancer agent. Such diversification is ascribable to the intrinsic adaptability of PK biosynthesis, which functions through several critical stages. The assembly of the PK carbon scaffold, driven by the choice of the starter units and—to a limited degree—extender modules, provides the basic framework for diversity. In addition, catalytic domains within the biosynthetic machinery act as decision gates that determine the final structure and release of the product. Tailoring enzymes further enhance chemical complexity through scaffold modifications, such as cyclization and dimerization of the β-keto-acyl carbon chain, as well as other changes that affect PK activity.

Terpenes, a ubiquitous class of natural compounds, are synthesized from two building blocks: isopentenyl diphosphate and its isomer dimethylallyl diphosphate. These precursors are synthesized through either the methylerythritol phosphate (MEP)



B

BGC Class	General Features	Complexity
Polyketide (PK)	Multi-domain enzymes synthesize polyketides by condensing acyl-CoA derivatives. They can be modular (Type I) or iterative (Type II and III)	High complexity due to modular architecture and structural diversity
Non-ribosomal peptide (NRP)	Large, multi-modular enzymes assemble peptides without ribosomes, incorporating both standard and non-standard amino acids	High complexity due to modular structure and diverse amino acid usage
Ribosomally synthesized post- translationally modified peptides (Ripp)	Ribosome-synthesized precursor peptides undergo extensive post- translational modifications to become bioactive	Moderate to high complexity due to diverse and extensive modifications (e.g., cyclization, dehydration, methylation, oxidation)
Saccharide	The process involves polysaccharide biosynthesis or glycosylation of natural products using glycosyltransferases for sugar addition	Complexity varies based on saccharide structure and degree of glycosylation
Terpene	This class is synthesized from isoprene units by terpene synthases, producing a wide range of terpenoids	Variable complexity ranging from simple monoterpenes to complex diterpenes and triterpenes

FIG 3 Biosynthesis of main secondary metabolite classes. (A) The panels present an overview of common features in the biosynthetic pathways of major BGC classes. Saccharide biosynthesis relies on monosaccharides such as glucose, fructose, and galactose, which serve as the fundamental building blocks that are enzymatically linked to form carbohydrates ranging from simple disaccharides to complex polysaccharides. The polyketide (PK) biosynthesis is carried out by modular PK synthases which consist of linearly assembled modules. Each module contains an acyl carrier protein (ACP) domain and a ketosynthase (KS) domain, along with a variable number of optional domains. During PK chain elongation, an ACP-bound acyl group condenses with a KS-bound substrate, releasing CO_2 and forming an elongated chain on ACP. Ribosomally synthesized and post-translationally modified peptide (RiPP) biosynthesis starts with a precursor peptide consisting of a signal sequence, a leader peptide at the N-terminus of an unmodified core peptide, and a recognition sequence. The signal and the recognition sequence are optionally present in the precursor peptide. The precursor peptide undergoes post-translational modification, followed by leader peptide removal, resulting in the mature bioactive peptide. NRPS biosynthesis operates through modules containing adenylation (A), condensation (C), peptidyl carrier protein, and thioesterase domains, which sequentially activate, elongate, and release the peptide chain. Terpene biosynthesis relies on isoprene units which are produced through either the methylerythritol phosphate (MEP) or the mevalonate (MVA) pathway as universal precursors that are enzymatically assembled and modified to generate diverse terpenes. (B) The table summarizes the general features of the BGCs depicted in panel A, highlighting the biosynthetic mechanisms that contribute to the complexity of secondary metabolites biosynthesis. (C) The heatmap shows secondary metabolites genetically encoded by BGCs identified in human-associated microorganisms. The secondary metabolites are grouped according to their BGC class and functionally annotated by the broad categories depicted in the legend.

pathway or the mevalonate (MVA) route. Most bacteria predominantly utilize the MEP pathway, while some employ the MVA pathway or a combination of both, reflecting metabolic adaptations to diverse ecological niches (77, 78).

BGC DISCOVERY IN THE HUMAN MICROBIOME

Breakthroughs in molecular techniques, such as amplicon sequencing, whole microbial genome sequencing, and metagenomics, have substantially expanded our knowledge

of the complex microbial communities inhabiting various human body sites (79). The resulting sequencing data, assembled within collaborative frameworks, such as the US National Institutes of Health-funded HMP (80, 81) and the Human Gastrointestinal Bacteria Genome Collection (82), as well as broader initiatives like the Genomic Catalog of Earth's Microbiomes (JGI GEM) (83) and the EMBL-EBI MGnify catalogs (84), have been extensively mined to investigate BGC content in human-associated genomes (12).

A healthy human microbiome features an extensive assortment of BGCs. For instance, the ABC-HuMi database currently catalogs 19,218 BGCs computationally predicted from the (meta)genomic sequences derived from 14 body sites, including the gut, oral, skin, respiratory, and urogenital systems. This biosynthetic wealth is further underscored by the classification of BGCs into 8,989 gene cluster families (GCFs) and 294 gene cluster clans. The main BGC classes identified are saccharides, RiPPs, NRPs, PKs, and terpenes, each variably distributed across body sites (12, 14). In this regard, the gut (8, 12, 85, 86) and oral (13, 87, 88) microbiomes have been the most extensively analyzed for their BGC content.

Saccharides constitute the most abundant BGCs in the human microbiome, with significant enrichment observed in gut and oral samples (89). One of the best-characterized examples is polysaccharide A (90), encoded by the ubiquitous gut microorganism *Bacteroides fragilis*, which is renowned for its immunomodulatory activity.

BGCs encoding specialized metabolites have consistently proven to be a reliable source of compounds with impactful biomedical applications. Examples include the PK antibiotics erythromycin and tetracycline; the NRP antibiotics penicillins and cephalosporins; and the glycopeptides from the vancomycin family (91). Despite this success, the sequence-guided discovery of specialized metabolites encoded in the human microbiome has progressed at a relatively slow pace, yielding a limited number of bioactive compounds derived from the identified BGCs.

Most specialized metabolites that have been characterized so far belong to the RiPPs type (Fig. 3C), encompassing lanthipeptides (92), sactipeptides (93, 94), thiopeptides (12), and microcins (95). RiPPs are particularly interesting antimicrobial compounds due to their narrow-spectrum activities and their ability to employ multiple simultaneous mechanisms of action, thereby reducing off-target effects and minimizing the risk of resistance development (96). However, none of the RiPPs identified from human microbiome-derived metagenomic sequencing data have yet been approved for clinical use in humans. A promising candidate is LFF571, a derivative of thiopeptide GE2270-A, produced by the vaginal isolate *Lactobacillus gasseri* JV-V03, which showed potent activity against *Clostridium difficile* infections and successfully completed a phase II clinical trial (97). Another notable compound is lactocillin, identified in the vaginal isolate of *L. gasseri* JV-V03 as well during a systematic analysis of BGCs in the HMP metagenomic samples. Lactocillin has been proposed to play a role in protecting the vaginal microbiota against pathogen invasion (12).

Mining metagenomic sequences has also uncovered a large family of 47 NRP BGCs, widely distributed in HMP stool samples and highly prevalent in the gut microbiome (12). Preliminary analysis of these gene clusters in RNA-seq data sets from stool samples has revealed robust expression levels for at least one BGC in nearly all samples. A subset of the identified BGCs was shown to be peptide aldehydes, which bear protease inhibitory activity with pronounced selectivity toward a subset of cathepsins involved in immune responses (98). Conversely, relatively few polyketides have been characterized from human microbiome sequences. One of them is the potent genotoxic colibactin (99), which is encoded by various *Escherichia coli* strains belonging to the B2-phylogroup and has been strongly implicated in the development of colorectal cancer (100).

Albeit scarcely applied to human microbiome-related BGCs, approaches pairing sequence-based computational imputation of BGCs with synthetic biology have led to the discovery of molecules with promising bioactivity (101, 102). For example, the discovery of two closely related NRPS BGCs by mining the human microbiome sequences in the HMP (80) and HOMD (103) repositories led to the chemical synthesis of

two humimycins, which were found to be broadly active against *Firmicutes* and, to a lesser extent, *Actinobacteria*, the dominant phylum in the human gut microbiome (104). More recently, lasso peptides, discovered in human commensal organisms, have been refactored using *E. coli* codon-optimized genes for *E. coli* expression, resulting in the production of novel lasso peptides, whose bioactivities warrant further investigation (11).

Despite these notable examples, the discovery of the chemical repertoire encoded by the human microbiome remains significantly underdeveloped compared to its well-documented taxonomic richness, as corroborated by multiple human microbiome studies (13, 90). This disparity calls for the systematic interrogation of the microbiome as a source of pharmacologically and, more broadly, biotechnologically valuable molecules. A wide variety of genome mining approaches, which we address in detail below, have become readily accessible (105). Moreover, with the growing availability of matched genomic and metabolomic data sets on platforms such as the Paired Omics Data Platform (106), the prioritization, functional analysis, and structural characterization of BGCs synthesizing natural products can now be considerably enhanced by linking mass spectral data to gene clusters (30, 107, 108).

BGC IDENTIFICATION AND ANNOTATION

BGC discovery involves a wide range of information processing tools designed to automate the identification and functional annotation of BGCs from genomic data. This crucial task presents significant challenges as it entails solving several complex sub-tasks that require innovative algorithmic solutions. Specifically, BGC discovery encompasses not only the identification of BGCs in (meta)genomic sequences but also their classification, dereplication, prioritization, connection with metabolite data, and functional characterization.

This review focuses on the sequence-based identification and classification of gene clusters in different settings (Tables 1 and 3).

BGC-class independent genome mining

To identify putative BGCs from isolate genomes or metagenomes, two basic approaches can be employed: principled methods, which leverage accumulated knowledge on biosynthetic logic, and data-driven methods, which use machine learning and statistical models to infer patterns from large data sets. The former excels at predicting BGCs encoding known classes of biosynthetic enzymes, yielding low false positive rates, but its ability to discover novel classes is limited. Conversely, data-driven approaches facilitate the identification of previously unknown BGCs but often exhibit higher false positive rates for uncharacterized clusters and increased false negative rates for known ones.

Biosynthetic logic-assisted tools rely on the observation that BGCs often share common properties, particularly enzyme families responsible for catalyzing key biochemical transformations involved in specialized metabolite biosynthesis. Popular tools such as antibiotics and Secondary Metabolite Analysis SHell (antiSMASH) (21) and PRediction Informatics for Secondary Metabolomes (PRISM) (27) employ profile hidden Markov models (pHMMs) of protein domains generated from multiple sequence alignments to identify gene combinations encoding biosynthetic pathway signatures. The use of pHMMs is very reliable for identifying BGCs encoding many well-established types of biosynthetic machinery, including PKSs, NRPSs, and known classes of RiPPs. For instance, antiSMASH 7.0 (21) can recognize up to 81 BGC types. While both antiSMASH and PRISM generally provide very similar results, antiSMASH has increasingly emphasized functional and comparative genomic analyses, whereas PRISM has refined a combinatorial approach for chemical structure prediction, enabling automated matching with mass-spectral data.

Data-driven tools play a key role in discovering new BGC types, thereby expanding the range of natural products. Various procedures have been implemented to achieve this goal (Table 1). One pioneering method, ClusterFinder (20), avoids reliance on

TABLE 3 Mapping frequently asked questions concerning BGC identification to representative software tools[a]

Tool	URL	Are you interested in any BGC class?	Are you searching for specific BGC classes?	Are you searching for a tool assisted by biosynthetic logic?	Are you searching for a supervised ML[b] tool?	Are you searching for an NLP-inspired ML tool?	Are metagenomic data available?	Would you like to link BGCs to NPs?	Would you like to compare BGCs?	Would you like to avail yourself of evolutionary principles?	Are you interested in analytics workflows?	Are structural data available?	Are expression data available?
Antibiotics and Secondary Metabolite Analysis SHell (antiSMASH)	https://antismash.secondarymetabolites.org	x		x	x		x						
Antibiotic-Resistant Target Seeker	http://arts.ziemertlab.com		x	x	x		x			x			
BAGEL4	http://bagel4.molgenrug.nl		x		x		x						
BGCFlow	https://github.com/NBChub/bgcflow	x		x	x		x		x	x	x		
BGC-Prophet	https://github.com/HUST-NingKang-Lab/BGC-Prophet	x			x	x							
BiGCARP	https://github.com/microsoft/bigcarp	x			x	x							
BiG-MAP	https://github.com/medema-group/BiG-MAP	x		x	x		x				x		x
BiG-SCAPE	https://bigscape-corason.secondarymetabolites.org/	x		x	x		x		x				
BiG-SLiCE	https://github.com/medema-group/bigslice	x		x	x		x		x				
Biosynthetic-SPADES	https://genome.cshlp.org/content/suppl/2019/07/24/gr.243477.118.DC1		x				x						
CLUSEAN	https://bitbucket.org/tilmweber/clusean/src/master/		x	x									
ClusterFinder	https://github.com/petercim/ClusterFinder	x			x		x						
decRiPPter	https://github.com/Alexamk/decRiPPter		x		x					x			
DeepBGC	https://github.com/Merck/deepbgc	x			x	x	x						
Deep-BGCpred	https://github.com/pmobio/Deep-BGCpred	x			x	x							
DeepRiPP	https://github.com/magarveylab/NLPPrecursor/tree/master		x		x	x			x				
EvoMining	https://github.com/nselem/evomining	x								x			
GARLIC	https://github.com/magarveylab/garlic-release	x		x	x		x	x				x	
GECCO	https://gecco.embl.de	x			x		x						
GRAPE	https://github.com/magarveylab/grape-release		x	x				x				x	
IsaBGC	https://github.com/Kalan-Lab/IsaBGC	x		x	x	x	x		x		x		
MetaMiner	https://github.com/ablab/npdtools		x	x	x		x	x					
Nerpa	https://github.com/ablab/nerpa	x		x	x		x						
NeuRiPP	https://github.com/emzodls/neuripp		x		x		x						
NPLinker	https://zenodo.org/records/4680579	x		x	x		x	x	x				

(*Continued on next page*)

TABLE 3 Mapping frequently asked questions concerning BGC identification to representative software tools[a] (*Continued*)

Tool	URL	Are you interested in any BGC class?	Are you searching for specific BGC classes?	Are you searching for a tool assisted by biosynthetic logic?	Are you searching for a supervised ML[b] tool?	Are you searching for an NLP-inspired ML tool?	Are metagenomic data available?	Would you like to link BGCs to NPs?	Would you like to compare BGCs?	Would you like to avail yourself of evolutionary principles?	Are you interested in analytics workflows?	Are structural data available?	Are expression data available?
NPOmix	https://github.com/tiagolbiotech/NPOmix_python	x		x	x		x	x	x				
NRPminer	https://github.com/mohimanilab/NRPminer		x	x	x			x					
NRPSPredictor2	https://github.com/roettig/NRPSpredictor2		x		x		x						
Pep2Path	https://pep2path.sourceforge.net	x		x	x		x	x					
PRISM3	https://doi.org/10.1101/2023.05.23.540769	x		x	x								
PRISM4	http://prism.adapsyn.com	x		x	x		x						
RiPPER	https://github.com/streptomyces/ripper		x	x	x		x		x				
RiPPMiner	http://www.nii.ac.in/~priyesh/lantipepDB/new_predictions/index.php		x		x								
RiPPMiner-Genome	http://www.nii.ac.in/~priyesh/lantipepDB/new_predictions1/cyclizationPrediction.php		x		x								
RiPP-PRISM	https://doi.org/10.1101/2023.05.23.540769		x	x									
RiPPquest	http://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp		x	x				x					
RODEO	https://webtool.ripp.rodeo/		x	x	x		x						
SANDPUMA	https://bitbucket.org/chevrm/sandpuma		x	x	x					x			
SBSPKS	http://www.nii.ac.in/sbspks2.html		x	x				x				x	
TaxiBGC	https://github.com/danielchang2002/TaxiBGC_2022		x				x	x					
TrRiPP	https://github.com/zzhongzz/TrRiPP		x		x	x	x						

[a]The table displays common questions which arise when carrying out BGC genome mining along with the software tools serving as examples of the available solutions to answer the intended questions.
[b]ML, machine learning.

predefined biosynthetic gene profiles by relying on domain-based functional profiles. More specifically, ClusterFinder identifies gene clusters of both known and unknown classes using a two-state HMM-based probabilistic algorithm, where one hidden state corresponds to BGCs (BGC state) and the other represents non-BGC genome (non-BGC state). This approach converts a nucleotide sequence of a certain genome into a structured series of Pfam domains, assigning probabilistic scores to each domain based on its frequency distribution in BGC and non-BGC training data sets, while also considering the identities of adjacent domains for improved accuracy. After computing BGC probabilities for all domains processed by the algorithm, ClusterFinder identifies gene clusters as sets of genes separated by no more than one intervening gene and containing at least one domain with a BGC probability above a predefined threshold. Since the ClusterFinder procedure relies on functional domains, which can assemble in various ways to produce distinct natural products, it comes with minimal training set bias and is particularly effective in guiding the discovery of hybrid and novel BGC classes.

Evolutionary principles have been instrumental in guiding the development of several BGC detection tools. Some of them incorporate synteny analysis, taking advantage of the fact that biosynthetic genes are often physically clustered. Indeed, this analysis enables the detection of co-localized genetic loci that appear in phylogenetically related organisms but are absent from core genomes, making it an effective method for identifying BGCs (109).

Other evolution-based approaches do not rely on synteny analysis. A key example is EvoMining (39), which identifies metabolism-associated non-syntenic gene blocks based on the premise that genes involved in secondary metabolism originate as paralogs of primary metabolic enzymes. Indeed, an enzyme family can undergo accelerated evolution through duplication or horizontal gene transfer events, with retained gene copies conferring an evolutionary advantage by enabling secondary metabolite biosynthesis.

EvoMining requires a genome database, an enzyme database, and a database of natural product biosynthetic enzymes. Initially, EvoMining detects expansion and recruitment events within enzyme families. An expanded family consists of all enzyme copies retrieved by querying the genome database with the entries of the enzyme database. Subsequently, the expanded enzymatic family is cross-referenced with the database of natural product biosynthetic enzymes to identify its potential role in specialized metabolism. Enzyme families associated with expansion and recruitment events are then subjected to phylogenetic analysis. The ensuing phylogenetic trees are examined to pinpoint enzymes more closely related to those predicted to be involved in natural product biosynthesis rather than those engaged in central metabolism. Such enzymes are deemed putative candidates for novel functions in specialized metabolic pathways.

As briefly outlined in the previous section, HMMs are widely employed in various tools, including antiSMASH, PRISM, and ClusterFinder (Table 1). A known disadvantage of HMMs is represented by their inability to recognize position dependencies between genomically distant entities (110). DeepBGC was designed to address this algorithmic limitation by implementing a deep learning approach based on bidirectional long short-term memory (Bi-LSTM) recurrent neural networks and vector representations of Pfam protein family domains (pfam2vec), which inherently capture dependencies between both adjacent and distant entities in the genome (26). Furthermore, a prominent feature of DeepBGC is its post-processing stage, where a random forest classifier, trained and tested on the MIBiG database, enables the classification of putative BGCs based on their corresponding product classes and broad molecular mechanisms of action. Of note, the DeepBGC random forest classifier can detect the most influential Pfam domains, thereby providing a data-driven alternative to expert knowledge-based classification approaches that rely on human-defined rules.

Deep-BGCpred, inspired by DeepBGC, boasts several technical innovations to mitigate false positives and enhance model stability (23). This method consists of two

stages: a Pfam protein family domain encoder and a stacked Bi-LSTM model. Unlike DeepBGC, Deep-BGCpred encodes not only Pfam domain identifiers into pfam2vec vectors but also integrates additional neural network vectors incorporating Pfam domain annotations, such as the domain summaries and high-level classification information. Similarly to DeepBGC, these embedding vectors are concatenated and fed into a stacked Bi-LSTM neural network, which predicts Pfam domain scores. These scores are then averaged per gene, and genes with scores exceeding a predefined threshold are grouped into BGCs. Finally, Deep-BGCpred employs a random forest classifier that assigns BCGs to one of eight categories consisting of a non-BGC class and seven BGC ones. To reduce the number of false positives in predicted BGCs, Deep-BGCpred enables the creation of negative samples using an augmentation technology that exploits the Pfam domain similarity network in the European Molecular Biology Laboratory (EMBL) database.

A source of model instability is the imbalance between the number of protein family domains in artificially created training data and the real genomic data. To address this inconsistency, Deep-BGCpred applies a sliding window procedure to the training data, allowing multiple Pfam sequence fragments to be derived from continuous Pfam sequences for improved Pfam score computation.

The language processing neural network BGC-Prophet has recently been introduced to detect both known and novel BGCs and classify their natural products (111). This tool is trained on thousands of genomes to learn gene location dependencies using the evolutionary scale modeling (ESM) method, which relies on the fact that protein sequences at the evolutionary scale offer a representation of biological structure and function. A subsequent fine-tuning phase ensures that the representational capacity of the embedding is specifically honed to grasp context information relevant to biosynthetic gene clusters. The embedding vectors are then input to a transformer encoder language model that employs a multi-head self-attention mechanism to speed up the training phase and improve accuracy in BGC prediction and multi-label classification. The increase in computational speed makes BGC-Prophet suitable for processing also metagenomic data sets and pan-phylogenetic screenings.

The machine learning tools discussed so far are supervised models. As such, their predictive quality is tied to the accuracy and depth of the training data sets. The difficulty of accessing massive amounts of high-quality and balanced labeled data is known to significantly impact model performance. This limitation can be addressed by self-supervised protein language models, which are gaining growing attention. In such approaches, models are pretrained on large volumes of unlabeled data to capture fundamental knowledge underlying proteins in such a way that they can then use the captured principles to carry out diverse analytical tasks. Alternatively, they can be fine-tuned on downstream supervised tasks for enhanced specificity. Although still in the early stages of application, self-supervised masked language models have recently been employed for BGC identification. These models train neural networks to reconstruct missing tokens in a corrupted sequence or predict the next element in a sequence based on preceding ones. One such method is Biosynthetic Gene Convolutional Autoencoding Representations of Proteins (BiGCARP) (24), which models BGCs as chains of functional protein domains and uses the ESM-1b transformer to obtain pretrained embeddings of these protein domains. BiGCARP trains a convolutional neural network (CNN)-based masked language model, built upon ByteNet (112) and CARP (113), on such domains to acquire BiGCARP, which can be used to predict BGCs and classify their products. As mounting evidence suggests that the adoption of self-supervised training schemes can effectively expand the detection capabilities of BGCs beyond known classes, further investigation into self-supervised language modeling approaches is warranted. The computational advances in BGC prediction enabled by CNNs, which—unlike transformers—scale linearly with the input sequence length, highlight the potential for further exploration of alternative CNN architectures.

Mapping BGC diversity

The substantial diversity and distribution of BGCs and their associated secondary metabolites are now being recognized thanks to the utilization of genome mining techniques across various levels of biological organization. Sequence-similarity-based networks have eased the exploration of relationships among BGC architectures, leading to the identification of biosynthetic GCFs. Organizing BGCs into GCFs can provide valuable insights into the recognition and classification of newly identified BGCs. Furthermore, it can help link the identified BGCs to their corresponding peptide products by detecting statistically significant correlations between the presence of GCF members in genomic data sets and that of chemotypes in MS data sets (114).

Conventional sequence similarity-based approaches start by comparing BGC protein sequences, an all-vs-all problem that scales quadratically with increasing data set size. Given the sheer volume of BGCs, such methods require lengthy CPU time, making them impractical for routine research. In addition, sequence-similarity-based methods may incorrectly quantify the similarity between fragmented gene clusters, which are common in metagenomic and large-scale pan-genome sequencing data sets generated through short-read technologies.

The Biosynthetic Gene Similarity Clustering and Prospecting Engine (BIG-SCAPE) tool is a sequence-similarity-based method employed to reconstruct GCFs. It addresses the aforementioned shortcomings by reframing the all-vs-all sequence comparisons into Pfam-based pairwise comparisons of pHMMs to obtain a similarity network. GCFs are then extracted using an affinity propagation clustering algorithm (57). Moreover, BIG-SCAPE becomes more informative by facilitating the analysis of evolutionary relationships among BGCs through CORe Analysis of Syntenic Orthologs to prioritize Natural Product Biosynthetic Gene Clusters (CORASON), which computes high-resolution multi-locus phylogenies of BGCs within and across GCFs (57).

Unlike BIG-SCAPE, which relies on pairwise comparisons to build GCF networks, Biosynthetic Genes Super-Linear Clustering Engine (BiG-SLiCE) queries a library of curated pHMMs with input BGCs. The presence or absence of hits, along with their bit scores, is used to generate numerical feature vectors corresponding to the input BGCs. BiG-SLiCE then projects these vectors into Euclidean space where it runs a partitional clustering algorithm in a near-linear time complexity. This computational efficiency allows BiG-SLiCE to handle increasingly large data sets as they become available (56).

Metagenomic BGC mining

The emergence of shotgun metagenomic sequencing has revolutionized BGC prediction, extending its application beyond culture-dependent genomes to the broader biosynthetic potential of interacting microbial communities. Whether heuristic or machine learning based, most of the existing tools assume that each BGC is confined within a single contig in genome or metagenome assemblies. While this assumption presents challenges for sequenced microbial genomes—where BGCs are often scattered through multiple contigs—it is even less feasible in shotgun metagenomics, where contigs are often fragmented and short.

An alternative approach is represented by exploiting genome assembly graphs, exemplified by biosyntheticSPAdes (50), a tool that allows users to assemble NRPS BGCs, PKS BGCs, and mixed NRPS-PKS BCCs from assembly graphs generated by SPAdes (115) and metaSPAdes (116) assemblers. The core premise of BiosyntheticSPAdes is that reconstructing the arrangement of biosynthetic domains within a BGC is often sufficient to predict the core scaffold of the secondary metabolite encoded by such a cluster. The algorithm identifies domain motifs along assembly graph edges, extracts BGC assembly subgraphs by selecting all edges within a preset distance from previously detected domain edges, and subsequently generates a scaffolding graph linking contigs with closely positioned domains. The final step consists of solving the Rural Postman problem—consisting of finding a closed walk traversing a given subset of edges with

minimum total cost—to reconstruct putative BGC arrangements in the scaffolding graph.

The Metagenomic identifier of Biosynthetic Gene Clusters (MetaBGC) offers an assembly-independent framework for BGC detection in microbiomes by directly analyzing metagenomic reads, which are scored according to their alignment with predefined pHMMs (101). To reconcile the discrepancy between the full-length proteins used in pHMM training and the short metagenomic reads (typically ~100 bp long), MetaBGC transforms these models into segmented pHMMs (spHMMs). Metagenomic reads are then scored against the spHMMs and retained for further processing provided that their scores exceed predefined cutoffs. Since reads originating from the same BGC are reasonably expected to display similar coverage across metagenomic samples, biosynthetic reads are subsequently dereplicated, quantified, and binned according to their abundance profiles, ultimately facilitating BGC prediction. Despite the ability of MetaBGC to bypass biases caused by reliance on cultured isolates or metagenomic assemblies, its implementation is constrained by the need for fine-tuning multiple parameters for spHMM-based biosynthetic read identification, quantification, and clustering. Addressing this limitation through automated optimization and adaptive parameter selection may ameliorate the accuracy and scalability of BGC prediction in metagenomic data sets.

Another genome assembly-independent method is the Taxonomy-guided Identification of Biosynthetic Gene Clusters (TaxiBGC), which operates a structured three-stage workflow (25). First, it performs species-level taxonomic profiling of metagenomic samples. Second, it queries the TaxiBGC reference database to infer the presence of experimentally characterized BGCs in the identified species. Third, it evaluates the predicted BGCs by aligning the metagenomic reads to these clusters, ensuring they meet minimum criteria for the presence of a BGC gene and overall BGC coverage. Once validated, the corresponding secondary metabolites are retrieved from the TaxiBGC reference database. The MIBiG database provides annotated information on experimentally characterized BGCs and their associated products, which are stored in the TaxiBGC reference database.

As dictated by its pipeline design, the performance of TaxiBGC is strictly dependent on the accurate assignment of predicted BGCs to microbial species identified through metagenome taxonomic profiling, as well as the depth of structural and functional characterization available for known BGC classes. A key feature of the TaxiBGC method is its intrinsic ability to identify specific microbial species harboring the predicted BGC genes.

The Secondary Metabolite Gene Cluster Annotations using Neural Networks Trained on InterPro Signatures (SanntiS) tool represents a more recent machine learning-based approach suitable to both genomic and metagenomic data sets (117). It consists of an artificial neural network with a one-dimensional convolutional layer and a Bi-LSTM to classify BGCs based on representations learned from the MIBiG database.

Despite the advancements achieved through the computational approaches mentioned, a fundamental challenge remains in identifying yet-to-be-characterized BGC classes in metagenomic samples, as each tool ultimately relies on known BGC models. Consequently, the integration of unsupervised learning algorithms into BGC detectors is expected to boost our ability to fully understand the biosynthetic potential of microbial communities.

BGC class-specific mining approaches

Besides general-purpose tools for BGC mining, a broad range of specialized computational approaches have been developed to help discover specific BGC classes. These tools leverage extensive knowledge of biosynthetic pathways, chemical scaffolds, and bioactivities to improve prediction accuracy. A comprehensive survey of the relevant literature, focusing on studies associated with tools that have been carefully curated and maintained over time, is summarized in Table 1. Several of these specialized tools

serve as BGC class-oriented detectors, providing targeted solutions for RiPP-, NRP-, and PK-producing BGCs.

RiPPs are characterized by the absence of conserved biosynthetic features and extensive structural diversity, which makes them highly promising for drug discovery applications (118–120) but also particularly challenging to discover (121). As a result, genome mining for RiPP gene clusters has become a focal point in biomedicine and cheminformatics, leading to the development of several specialized tools.

One of the earliest tools is RiPP-PRISM (43), which extends the PRISM framework by incorporating: (i) motif assembly for precursor cleavage prediction, (ii) HMMs for detecting a broad range of RiPP classes, and (iii) virtual tailoring reactions to infer the chemical structure of RiPP products. In general, class-dependent RiPP genome mining relies on sequence similarity to known biosynthetic enzymes (75), whereas class-independent approaches leverage conserved RiPP RREs. For example, the RRE-Finder (122) combines pHMMs with a truncated HHpred pipeline (123) to facilitate the detection of divergent RRE sequences. Rapid ORF Description and Evaluation Online (RODEO) (36) can then analyze RRE-containing proteins, reducing false positives by identifying co-occurring ORFs encoding biosynthetic enzymes.

Another approach employed to identify first-in-class RiPPs is represented by RiPPER (41), which scans genomic regions flanking putative tailoring enzymes for ORFs with lengths consistent with known RiPP precursor peptides. RiPPMiner uses a support vector machine (SVM) classifier to distinguish RiPP precursor peptides from other small proteins, while also predicting RiPP classes, leader peptide cleavage sites, and potential crosslinks in the core peptide (45). RiPPMiner-Genome builds on this approach by accepting genomic sequences as input—instead of relying on RiPP precursor peptide sequences—and predicting crosslinked RiPP structures (46). This update was made possible by systematically revising BGC information for RiPPs with known chemical structures and updating the prediction rules for RiPP precursor identification, cleavage motifs, and core modifications.

Deep learning-based methods have also contributed to RiPP genome mining. For instance, DeepRiPP (38) adapts architectures from natural language processing, whereas NeuRiPP uses a neural network structure (42). Furthermore, TrRiPP (32) combines transformer encoders with Bi-LSTM layers to discriminate RiPP precursors from non-RiPP short peptides and classify RiPPs into subclasses. Since these tools rely on the precursor peptides for predicting RiPPs, they do not suffer from limitations in RiPP identification associated with fragmented assemblies or the presence of distantly encoded modification enzymes. As such, they are well suited for RiPP identification in metagenomic sequences.

The discovery of NRPs and PKs poses significant challenges due to the modular assembly of these systems (76, 124), the broad substrate specificity of adenylation domains responsible for amino acid recognition and activation, and their extensive post-assembly modifications. Addressing this biosynthetic complexity has led to the emergence of tools that often couple general-purpose BGC mining tools, such as antiSMASH, with NRPS- and PKS-specific predictors. Examples include NRPSpredictor2 (48), the NRPyS library (125), and SANDPUMA (47), all of which are capable of predicting the substrate specificity of the adenylation domains. In addition, NRPminer (51), Nerpa (52), and SBSPKS (54) employ distinct computational strategies to integrate (meta)genomic and metabolomic data, with the common goal of linking predicted BGCs to their corresponding NRPs and PKs.

Unlike the previous tools, NaPDoS (60) utilizes a phylogeny-based classification of ketosynthase (KS) and condensation domains to infer PK and NRP biosynthetic potential. Since it does not require fully assembled BGCs, it becomes especially useful when analyzing large metagenomic and PCR amplicon data sets.

Finally, DeepT2 (126) applies principles from protein natural language processing to predict bacterial type II PKs. It converts protein sequences of key ketosynthases into vector embeddings using EMS-2 and employs semi-supervised learning to associate KS

embeddings with the PK class labels, facilitating the prediction of both known and novel type II PKs.

CONCLUSIONS

The present review provides a detailed assessment of sequence-based BGC mining tools that have significantly advanced our understanding of the biosynthetic potential of both isolate genomes and complex microbial communities across diverse ecological environments (127). Based on existing studies, the systematic characterization of the specialized metabolites produced by human-associated microorganisms has not only revealed the mechanistic interplay between human microbiome and health—an association that continues to gain recognition (128–130)—but has also contributed to the discovery of bioactive compounds with promising biomedical applications (6).

The genetic basis of specialized metabolites in the human microbiome has been extensively investigated using established tools like ClusterFinder and antiSMASH. Nonetheless, the rapid accumulation of (meta)genomic sequencing data has far outpaced our ability to gain new insights into gene clusters synthesizing biologically active molecules. To address this gap, it is worth pursuing the integration of complementary computational methodologies—such as the unsupervised machine learning algorithms outlined in this review—to enhance discovery rates and facilitate the identification of novel biosynthetic pathways.

As our overview points out, scientists keen to decipher microbial secondary metabolism face a burgeoning growth of software tools, which differ in input requirements, reconstructed attributes, and algorithmic frameworks (Table 3). Despite the inclusion of performance evaluations in most newly introduced BGC predictors, navigating these tools—whether across different categories or within the same category with varying settings—remains a substantial challenge.

The underlying causes of improved tool performance are often unclear and cannot always be ascribed to improved algorithmic architecture or training data. This underscores the need for standardized software evaluation using data sets with established biological ground truths. Such efforts would support researchers in selecting the most suitable tools and enable developers to better understand what design features contribute to successful modeling and prediction. The wealth of genetically encoded secondary metabolites suggests that any benchmarking study should take into account the intended use of the software, apply metrics that are appropriate for the question being asked—as no single evaluation metric fits all purposes—and adopt the most unbiased evaluation setting possible. We put forward that the benchmarking of tools developed to mine (meta)genomes for BGCs should be subjected to a community-wide effort in the wake of the initiative known as DREAM, standing for Dialog for Reverse Engineering Assessments and Methods (131).

Finally, we emphasize the importance of exploring options beyond the scope of our review that are particularly relevant to translational applications. Integrating sequence-based BGC prediction with complementary data types, such as transcriptomic (55), proteomic and metabolomic features (30), and structural modeling data (53), has already shown its value in providing deeper insights into the biosynthetic potential of the human microbiome and advancing natural product discovery.

ACKNOWLEDGMENTS

This publication is part of the project PRIN 2022 PNRR with code P2022AFS8P, which has received funding from NextGeneration EU-MUR–M4C2 1.1, CUP C53D23007570001.

AUTHOR AFFILIATIONS

[1]Department of Applied Science and Technology (DISAT), Politecnico di Torino, Torino, Italy

[2]Department of Translational Medicine (DIMET), University of Piemonte Orientale, Novara, Italy
[3]Center for Translational Research on Autoimmune and Allergic Disease (CAAD), Novara, Italy

AUTHOR ORCIDs

Angela Re http://orcid.org/0000-0002-3179-6967

FUNDING

Funder	Grant(s)	Author(s)
Next Generation EU - MUR	P2022AFS8P	Davide Corà
		Angela Re

AUTHOR CONTRIBUTIONS

Mohammed Salim Dason, Data curation, Investigation, Visualization, Writing – original draft, Writing – review and editing | Davide Corà, Data curation, Funding acquisition, Investigation, Writing – review and editing | Angela Re, Conceptualization, Data curation, Funding acquisition, Investigation, Supervision, Visualization, Writing – original draft, Writing – review and editing

REFERENCES

1. Aggarwal N, Kitano S, Puah GRY, Kittelmann S, Hwang IY, Chang MW. 2023. Microbiome and human health: current understanding, engineering, and enabling technologies. Chem Rev 123:31–72. https://doi.org/10.1021/acs.chemrev.2c00431
2. Esvap E, Ulgen KO. 2021. Advances in genome-scale metabolic modeling toward microbial community analysis of the human microbiome. ACS Synth Biol 10:2121–2137. https://doi.org/10.1021/acssynbio.1c00140
3. Liu YY. 2023. Controlling the human microbiome. Cell Syst 14:135–159. https://doi.org/10.1016/j.cels.2022.12.010
4. Manor O, Borenstein E. 2017. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. Cell Host Microbe 21:254–267. https://doi.org/10.1016/j.chom.2016.12.014
5. Castells-Nobau A, Mayneris-Perxachs J, Fernández-Real JM. 2024. Unlocking the mind-gut connection: Impact of human microbiome on cognition. Cell Host Microbe 32:1248–1263. https://doi.org/10.1016/j.chom.2024.07.019
6. Milshteyn A, Colosimo DA, Brady SF. 2018. Accessing bioactive natural products from the human microbiome. Cell Host Microbe 23:725–736. https://doi.org/10.1016/j.chom.2018.05.013
7. Joice R, Yasuda K, Shafquat A, Morgan XC, Huttenhower C. 2014. Determining microbial products and identifying molecular targets in the human microbiome. Cell Metab 20:731–741. https://doi.org/10.1016/j.cmet.2014.10.003
8. Wang L, Ravichandran V, Yin Y, Yin J, Zhang Y. 2019. Natural products from mammalian gut microbiota. Trends Biotechnol 37:492–504. https://doi.org/10.1016/j.tibtech.2018.10.003
9. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. Nature 449:804–810. https://doi.org/10.1038/nature06244
10. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, et al. 2015. Minimum information about a biosynthetic gene cluster. Nat Chem Biol 11:625–631. https://doi.org/10.1038/nchembio.1890
11. King AM, Zhang Z, Glassey E, Siuti P, Clardy J, Voigt CA. 2023. Systematic mining of the human microbiome identifies antimicrobial peptides with diverse activity spectra. Nat Microbiol 8:2420–2434. https://doi.org/10.1038/s41564-023-01524-6
12. Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG, Fischbach MA. 2014. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158:1402–1414. https://doi.org/10.1016/j.cell.2014.08.032
13. Aleti G, Baker JL, Tang X, Alvarez R, Dinis M, Tran NC, Melnik AV, Zhong C, Ernst M, Dorrestein PC, Edlund A. 2019. Identification of the bacterial biosynthetic gene clusters of the oral microbiome illuminates the unexplored social language of bacteria during health and disease. mBio 10:e00321-19. https://doi.org/10.1128/mBio.00321-19
14. Hirsch P, Tagirdzhanov A, Kushnareva A, Olkhovskii I, Graf S, Schmartz GP, Hegemann JD, Bozhüyük KAJ, Müller R, Keller A, Gurevich A. 2024. ABC-HuMi: the atlas of biosynthetic gene clusters in the human microbiome. Nucleic Acids Res 52:D579–D585. https://doi.org/10.1093/nar/gkad1086
15. Zou H, Sun T, Jin B, Wang S. 2023. sBGC-hm: an atlas of secondary metabolite biosynthetic gene clusters from the human gut microbiome. Bioinformatics 39:btad131. https://doi.org/10.1093/bioinformatics/btad131
16. Pita S, Myers PN, Johansen J, Russel J, Nielsen MC, Eklund AC, Nielsen HB. 2024. CHAMP delivers accurate taxonomic profiles of the prokaryotes, eukaryotes, and bacteriophages in the human microbiome. Front Microbiol 15:1425489. https://doi.org/10.3389/fmicb.2024.1425489
17. Manghi P, Blanco-Míguez A, Manara S, NabiNejad A, Cumbo F, Beghini F, Armanini F, Golzato D, Huang KD, Thomas AM, Piccinno G, Punčochář M, Zolfo M, Lesker TR, Bredon M, Planchais J, Glodt J, Valles-Colomer M, Koren O, Pasolli E, Asnicar F, Strowig T, Sokol H, Segata N. 2023. MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. Cell Rep 42:112464. https://doi.org/10.1016/j.celrep.2023.112464
18. Nuhamunada M, Mohite OS, Phaneuf PV, Palsson BO, Weber T. 2024. BGCFlow: systematic pangenome workflow for the analysis of biosynthetic gene clusters across large genomic datasets. Nucleic Acids Res 52:5478–5495. https://doi.org/10.1093/nar/gkae314
19. Salamzade R, Cheong JZA, Sandstrom S, Swaney MH, Stubbendieck RM, Starr NL, Currie CR, Singh AM, Kalan LR. 2023. Evolutionary investigations of the biosynthetic diversity in the skin microbiome using *lsa*BGC. Microb Genom 9:mgen000988. https://doi.org/10.1099/mgen.0.000988
20. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Linington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421. https://doi.org/10.1016/j.cell.2014.06.034

21. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res 51:W46–W50. https://doi.org/10.1093/nar/gkad344
22. Carroll LM, Larralde M, Fleck JS, Ponnudurai R, Milanese A, Cappio E, Zeller G. 2021. Accurate *de novo* identification of biosynthetic gene clusters with GECCO. bioRxiv. https://doi.org/10.1101/2021.05.03.442509
23. Yang Z, Liao B, Hsieh C, Han C, Fang L, Zhang S. 2021. Deep-BGCpred: a unified deep learning genome-mining framework for biosynthetic gene cluster prediction. bioRxiv. https://doi.org/10.1101/2021.11.15.468547
24. Rios-Martinez C, Bhattacharya N, Amini AP, Crawford L, Yang KK. 2023. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. PLoS Comput Biol 19:e1011162. https://doi.org/10.1371/journal.pcbi.1011162
25. Gupta VK, Bakshi U, Chang D, Lee AR, Davis JM 3rd, Chandrasekaran S, Jin Y-S, Freeman MF, Sung J. 2022. TaxiBGC: a taxonomy-guided approach for profiling experimentally characterized microbial biosynthetic gene clusters and secondary metabolite production potential in metagenomes. mSystems 7:e00925-22. https://doi.org/10.1128/msystems.00925-22
26. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, Durcak J, Wurst M, Kotowski J, Chang D, Wang R, Piizzi G, Temesi G, Hazuda DJ, Woelk CH, Bitton DA. 2019. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res 47:e110. https://doi.org/10.1093/nar/gkz654
27. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster ALH, Wyatt MA, Magarvey NA. 2015. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 43:9645–9662. https://doi.org/10.1093/nar/gkv1012
28. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. 2017. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res 45:W49–W54. https://doi.org/10.1093/nar/gkx320
29. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, Li H, Ranieri MRM, Webster ALH, Cao MPT, Pfeifle A, Spencer N, To QH, Wallace DP, Dejong CA, Magarvey NA. 2020. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun 11:6058. https://doi.org/10.1038/s41467-020-19986-1
30. Leão TF, Wang M, da Silva R, Gurevich A, Bauermeister A, Gomes PWP, Brejnrod A, Glukhov E, Aron AT, Louwen JJR, Kim HW, Reher R, Fiore MF, van der Hooft JJJ, Gerwick L, Gerwick WH, Bandeira N, Dorrestein PC. 2022. NPOmix: a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. PNAS Nexus 1:gac257. https://doi.org/10.1093/pnasnexus/pgac257
31. Hjörleifsson Eldjárn G, Ramsay A, van der Hooft JJJ, Duncan KR, Soldatou S, Rousu J, Daly R, Wandy J, Rogers S. 2021. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLoS Comput Biol 17:e1008920. https://doi.org/10.1371/journal.pcbi.1008920
32. Gao Y, Zhong Z, Zhang D, Zhang J, Li YX. 2024. Exploring the roles of ribosomal peptides in prokaryote-phage interactions through deep learning-enabled metagenome mining. Microbiome 12:94. https://doi.org/10.1186/s40168-024-01807-y
33. Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, Breitling R. 2014. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comput Biol 10:e1003822. https://doi.org/10.1371/journal.pcbi.1003822
34. Mungan MD, Alanjary M, Blin K, Weber T, Medema MH, Ziemert N. 2020. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. Nucleic Acids Res 48:W546–W552. https://doi.org/10.1093/nar/gkaa374
35. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. 2018. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res 46:W278–W281. https://doi.org/10.1093/nar/gky383
36. Tietz JI, Schwalen CJ, Patel PS, Maxson T, Blair PM, Tai HC, Zakai UI, Mitchell DA. 2017. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol 13:470–478. https://doi.org/10.1038/nchembio.2319
37. Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, Brewer HM, Pasa-Tolic L, Bandeira N, Moore BS, Pevzner PA, Dorrestein PC. 2014. Automated genome mining of ribosomal peptide natural products. ACS Chem Biol 9:1545–1551. https://doi.org/10.1021/cb500199h
38. Merwin NJ, Mousa WK, Dejong CA, Skinnider MA, Cannon MJ, Li H, Dial K, Gunabalasingam M, Johnston C, Magarvey NA. 2020. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. Proc Natl Acad Sci USA 117:371–380. https://doi.org/10.1073/pnas.1901493116
39. Sélem-Mojica N, Aguilar C, Gutiérrez-García K, Martínez-Guerrero CE, Barona-Gómez F. 2019. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. Microb Genom 5:e000260. https://doi.org/10.1099/mgen.0.000260
40. Kloosterman AM, Cimermancic P, Elsayed SS, Du C, Hadjithomas M, Donia MS, Fischbach MA, van Wezel GP, Medema MH. 2020. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. PLoS Biol 18:e3001026. https://doi.org/10.1371/journal.pbio.3001026
41. Santos-Aberturas J, Chandra G, Frattaruolo L, Lacret R, Pham TH, Vior NM, Eyles TH, Truman AW. 2019. Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool. Nucleic Acids Res 47:4624–4637. https://doi.org/10.1093/nar/gkz192
42. de los Santos ELC. 2019. NeuRiPP: Neural network identification of RiPP precursor peptides. Sci Rep 9:13406. https://doi.org/10.1038/s41598-019-49764-z
43. Skinnider MA, Johnston CW, Edgar RE, Dejong CA, Merwin NJ, Rees PN, Magarvey NA. 2016. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. Proc Natl Acad Sci USA 113:E6343–E6351. https://doi.org/10.1073/pnas.1609014113
44. Cao L, Gurevich A, Alexander KL, Naman CB, Leão T, Glukhov E, Luzzatto-Knaan T, Vargas F, Quinn R, Bouslimani A, et al. 2019. MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. Cell Syst 9:600–608. https://doi.org/10.1016/j.cels.2019.09.004
45. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D. 2017. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res 45:W80–W88. https://doi.org/10.1093/nar/gkx408
46. Agrawal P, Amir S, Barua D, Mohanty D. 2021. RiPPMiner-genome: a web resource for automated prediction of crosslinked chemical structures of RiPPs by genome mining. J Mol Biol 433:166887. https://doi.org/10.1016/j.jmb.2021.166887
47. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. 2017. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across *Actinobacteria*. Bioinformatics 33:3202–3210. https://doi.org/10.1093/bioinformatics/btx400
48. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. 2011. NRPSpredictor2--a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res 39:W362–W367. https://doi.org/10.1093/nar/gkr323
49. Weber T, Rausch C, Lopez P, Hoof I, Gaykova V, Huson DH, Wohlleben W. 2009. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol 140:13–17. https://doi.org/10.1016/j.jbiotec.2009.01.007
50. Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res 29:1352–1362. https://doi.org/10.1101/gr.243477.118
51. Behsaz B, Bode E, Gurevich A, Shi YN, Grundmann F, Acharya D, Caraballo-Rodríguez AM, Bouslimani A, Panitchpakdi M, Linck A, Guan C, Oh J, Dorrestein PC, Bode HB, Pevzner PA, Mohimani H. 2021. Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery. Nat Commun 12:3225. https://doi.org/10.1038/s41467-021-23502-4
52. Kunyavskaya O, Tagirdzhanov AM, Caraballo-Rodríguez AM, Nothias LF, Dorrestein PC, Korobeynikov A, Mohimani H, Gurevich A. 2021. Nerpa: a tool for discovering biosynthetic gene clusters of bacterial nonribosomal peptides. Metabolites 11:693. https://doi.org/10.3390/metabo11100693

53. Dejong CA, Chen GM, Li H, Johnston CW, Edwards MR, Rees PN, Skinnider MA, Webster ALH, Magarvey NA. 2016. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol 12:1007–1014. https://doi.org/10.1038/nchembio.2188
54. Khater S, Gupta M, Agrawal P, Sain N, Prava J, Gupta P, Grover M, Kumar N, Mohanty D. 2017. SBSPKSv2: structure-based sequence analysis of polyketide synthases and non-ribosomal peptide synthetases. Nucleic Acids Res 45:W72–W79. https://doi.org/10.1093/nar/gkx344
55. Pascal Andreu V, Augustijn HE, van den Berg K, van der Hooft JJJ, Fischbach MA, Medema MH. 2021. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. mSystems 6:e00937-21. https://doi.org/10.1128/mSystems.00937-21
56. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. 2021. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. Gigascience 10:giaa154. https://doi.org/10.1093/gigascience/giaa154
57. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. https://doi.org/10.1038/s41589-019-0400-9
58. van den Belt M, Gilchrist C, Booth TJ, Chooi Y-H, Medema MH, Alanjary M. 2023. CAGECAT: The CompArative GEne Cluster Analysis Toolbox for rapid search and visualisation of homologous gene clusters. BMC Bioinform 24:181. https://doi.org/10.1186/s12859-023-05311-2
59. Del Carratore F, Zych K, Cummings M, Takano E, Medema MH, Breitling R. 2019. Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. Commun Biol 2:83. https://doi.org/10.1038/s42003-019-0333-6
60. Klau LJ, Podell S, Creamer KE, Demko AM, Singh HW, Allen EE, Moore BS, Ziemert N, Letzel AC, Jensen PR. 2022. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. J Biol Chem 298:102480. https://doi.org/10.1016/j.jbc.2022.102480
61. Juarez RJ, Jiang Y, Tremblay M, Shao Q, Link AJ, Yang ZJ. 2023. LassoHTP: a high-throughput computational tool for lasso peptide structure construction and modeling. J Chem Inf Model 63:522–530. https://doi.org/10.1021/acs.jcim.2c00945
62. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, Bailey CB, Barra L, Booth TJ, Bozhüyük KAJ, et al. 2025. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res 53:D678–D690. https://doi.org/10.1093/nar/gkae1115
63. Blin K, Shaw S, Medema MH, Weber T. 2024. The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. Nucleic Acids Res 52:D586–D589. https://doi.org/10.1093/nar/gkad984
64. Palaniappan K, Chen I-MA, Chu K, Ratner A, Seshadri R, Kyrpides NC, Ivanova NN, Mouncey NJ. 2020. IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. Nucleic Acids Res 48:D422–D430. https://doi.org/10.1093/nar/gkz932
65. Kautsar SA, Blin K, Shaw S, Weber T, Medema MH. 2021. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res 49:D490–D497. https://doi.org/10.1093/nar/gkaa812
66. Bağcı C, Nuhamunada M, Goyat H, Ladanyi C, Sehnal L, Blin K, Kautsar SA, Tagirdzhanov A, Gurevich A, Mantri S, von Mering C, Udwary D, Medema MH, Weber T, Ziemert N. 2025. BGC Atlas: a web resource for exploring the global chemical diversity encoded in bacterial genomes. Nucleic Acids Res 53:D618–D624. https://doi.org/10.1093/nar/gkae953
67. Bepler T, Berger B. 2021. Learning the protein language: evolution, structure, and function. Cell Syst 12:654–669. https://doi.org/10.1016/j.cels.2021.05.017
68. Topol EJ. 2025. Learning the language of life with AI. Science 387:eadv4414. https://doi.org/10.1126/science.adv4414
69. Süssmuth RD, Mainz A. 2017. Nonribosomal peptide synthesis-principles and prospects. Angew Chem Int Ed Engl 56:3770–3821. https://doi.org/10.1002/anie.201609079
70. Heard SC, Winter JM. 2024. Structural, biochemical and bioinformatic analyses of nonribosomal peptide synthetase adenylation domains. Nat Prod Rep 41:1180–1205. https://doi.org/10.1039/d3np00064h
71. Bloudoff K, Schmeing TM. 2017. Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. Biochim Biophys Acta Proteins Proteom 1865:1587–1604. https://doi.org/10.1016/j.bbapap.2017.05.010
72. Richter D, Piel J. 2024. Novel types of RiPP-modifying enzymes. Curr Opin Chem Biol 80:102463. https://doi.org/10.1016/j.cbpa.2024.102463
73. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, et al. 2013. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat. Prod. Rep 30:108–160. https://doi.org/10.1039/C2NP20085F
74. Oman TJ, van der Donk WA. 2010. Follow the leader: the use of leader peptides to guide natural product biosynthesis. Nat Chem Biol 6:9–18. https://doi.org/10.1038/nchembio.286
75. Montalbán-López M, Scott TA, Ramesh S, Rahman IR, van Heel AJ, Viel JH, Bandarian V, Dittmann E, Genilloud O, Goto Y, et al. 2021. New developments in RiPP discovery, enzymology and engineering. Nat Prod Rep 38:130–239. https://doi.org/10.1039/d0np00027b
76. Cummings M, Breitling R, Takano E. 2014. Steps towards the synthetic biology of polyketide biosynthesis. FEMS Microbiol Lett 351:116–125. https://doi.org/10.1111/1574-6968.12365
77. Hoshino Y, Gaucher EA. 2018. On the origin of isoprenoid biosynthesis. Mol Biol Evol 35:2185–2197. https://doi.org/10.1093/molbev/msy120
78. Cheng S, Wang X, Deng Z, Liu T. 2025. Innovative approaches in the discovery of terpenoid natural products. Curr Opin Microbiol 83:102575. https://doi.org/10.1016/j.mib.2024.102575
79. Yi X, Lu H, Liu X, He J, Li B, Wang Z, Zhao Y, Zhang X, Yu X. 2024. Unravelling the enigma of the human microbiome: evolution and selection of sequencing technologies. Microb Biotechnol 17:e14364. https://doi.org/10.1111/1751-7915.14364
80. Consortium HMP. 2012. A framework for human microbiome research. Nature 486:215–221. https://doi.org/10.1038/nature11209
81. Consortium HMP. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234
82. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. 2019. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol 37:186–192. https://doi.org/10.1038/s41587-018-0009-7
83. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen IM, Huntemann M, et al. 2021. A genomic catalog of Earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-0718-6
84. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell LJ, Curtis T, Escobar-Zepeda A, Gurbich TA, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson DJ, Finn RD. 2023. MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res 51:D753–D759. https://doi.org/10.1093/nar/gkac1080
85. Hu D, Zhang T, He S, Pu T, Yin Y, Hu Y. 2024. Mining metagenomic data to gain a new insight into the gut microbial biosynthetic potential in placental mammals. Microbiol Spectr 12:e00864-24. https://doi.org/10.1128/spectrum.00864-24
86. Liu S, Zhang Z, Wang X, Ma Y, Ruan H, Wu X, Li B, Mou X, Chen T, Lu Z, Zhao W. 2024. Biosynthetic potential of the gut microbiome in longevous populations. Gut Microbes 16:2426623. https://doi.org/10.1080/19490976.2024.2426623
87. Barber CC, Zhang W. 2021. Small molecule natural products in human nasal/oral microbiota. J Ind Microbiol Biotechnol 48:kuab010. https://doi.org/10.1093/jimb/kuab010
88. Liu L, Hao T, Xie Z, Horsman GP, Chen Y. 2016. Genome mining unveils widespread natural product biosynthetic capacity in human oral microbe *Streptococcus mutans*. Sci Rep 6:37479. https://doi.org/10.1038/srep37479
89. Fobofou SA, Savidge T. 2022. Microbial metabolites: cause or consequence in gastrointestinal disease? Am J Physiol Gastrointest Liver Physiol 322:G535–G552. https://doi.org/10.1152/ajpgi.00008.2022
90. Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. 2005. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell 122:107–118. https://doi.org/10.1016/j.cell.2005.05.007

91. Fischbach MA, Walsh CT. 2006. Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. Chem Rev 106:3468–3496. https://doi.org/10.1021/cr0503097
92. Kim SG, Becattini S, Moody TU, Shliaha PV, Littmann ER, Seok R, Gjonbalaj M, Eaton V, Fontana E, Amoretti L, et al. 2019. Microbiota-derived lantibiotic restores resistance against vancomycin-resistant *Enterococcus*. Nature 572:665–669. https://doi.org/10.1038/s41586-019-1501-z
93. Bushin LB, Covington BC, Rued BE, Federle MJ, Seyedsayamdost MR. 2020. Discovery and biosynthesis of streptosactin, a sactipeptide with an alternative topology encoded by commensal bacteria in the human microbiome. J Am Chem Soc 142:16265–16275. https://doi.org/10.1021/jacs.0c05546
94. Balty C, Guillot A, Fradale L, Brewee C, Boulay M, Kubiak X, Benjdia A, Berteau O. 2019. Ruminococcin C, an anti-clostridial sactipeptide produced by a prominent member of the human microbiota *Ruminococcus gnavus*. J Biol Chem 294:14512–14525. https://doi.org/10.1074/jbc.RA119.009416
95. Baquero F, Lanza VF, Baquero MR, Del Campo R, Bravo-Vázquez DA. 2019. Microcins in *Enterobacteriaceae*: peptide antimicrobials in the eco-active intestinal chemosphere. Front Microbiol 10:2261. https://doi.org/10.3389/fmicb.2019.02261
96. Rowe SM, Spring DR. 2021. The role of chemical synthesis in developing RiPP antibiotics. Chem Soc Rev 50:4245–4258. https://doi.org/10.1039/d0cs01386b
97. Mullane K, Lee C, Bressler A, Buitrago M, Weiss K, Dabovic K, Praestgaard J, Leeds JA, Blais J, Pertel P. 2015. Multicenter, randomized clinical trial to compare the safety and efficacy of LFF571 and vancomycin for *Clostridium difficile* infections. Antimicrob Agents Chemother 59:1435–1440. https://doi.org/10.1128/AAC.04251-14
98. Guo CJ, Chang FY, Wyche TP, Backus KM, Acker TM, Funabashi M, Taketani M, Donia MS, Nayfach S, Pollard KS, Craik CS, Cravatt BF, Clardy J, Voigt CA, Fischbach MA. 2017. Discovery of reactive microbiota-derived metabolites that inhibit host proteases. Cell 168:517–526. https://doi.org/10.1016/j.cell.2016.12.021
99. Nougayrède JP, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. 2006. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science 313:848–851. https://doi.org/10.1126/science.1127059
100. Addington E, Sandalli S, Roe AJ. 2024. Current understandings of colibactin regulation. Microbiology (Reading) 170:001427. https://doi.org/10.1099/mic.0.001427
101. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, Jeffrey PD, Donia MS. 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 366:eaax9176. https://doi.org/10.1126/science.aax9176
102. Chu J, Vila-Farres X, Brady SF. 2019. Bioactive synthetic-bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome. J Am Chem Soc 141:15737–15741. https://doi.org/10.1021/jacs.9b07317
103. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford) 2010:baq013–baq013. https://doi.org/10.1093/database/baq013
104. Chu John, Vila-Farres X, Inoyama D, Ternei M, Cohen LJ, Gordon EA, Reddy BVB, Charlop-Powers Z, Zebroski HA, Gallardo-Macias R, Jaskowski M, Satish S, Park S, Perlin DS, Freundlich JS, Brady SF. 2016. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. Nat Chem Biol 12:1004–1006. https://doi.org/10.1038/nchembio.2207
105. Medema MH. 2021. The year 2020 in natural product bioinformatics: an overview of the latest tools and databases. Nat Prod Rep 38:301–306. https://doi.org/10.1039/d0np00090f
106. Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, Aleti G, Moghaddam JA, Aron AT, Aziz S, et al. 2021. A community resource for paired genomic and metabolomic data mining. Nat Chem Biol 17:363–368. https://doi.org/10.1038/s41589-020-00724-z
107. Louwen JJR, Medema MH, van der Hooft JJJ. 2023. Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. Microbiome 11:13. https://doi.org/10.1186/s40168-022-01444-3
108. Medema MH, Fischbach MA. 2015. Computational approaches to natural product discovery. Nat Chem Biol 11:639–648. https://doi.org/10.1038/nchembio.1884
109. Park HB, Perez CE, Barber KW, Rinehart J, Crawford JM. 2017. Genome mining unearths a hybrid nonribosomal peptide synthetase-like-pteridine synthase biosynthetic gene cluster. Elife 6:e25229. https://doi.org/10.7554/eLife.25229
110. Choo KH, Tong JC, Zhang L. 2004. Recent applications of hidden markov models in computational biology. Genomics Proteomics Bioinformatics 2:84–96. https://doi.org/10.1016/s1672-0229(04)02014-5
111. Lai Q, Yao S, Zha Y, Zhang H, Ye Y, Zhang Y, Bai H, Ning K. 2023. Deciphering the biosynthetic potential of microbial genomes using a BGC language processing neural network model. bioRxiv. https://doi.org/10.1101/2023.11.30.569352
112. Kalchbrenner N, Espeholt L, Simonyan K, Oord A, Graves A, Kavukcuoglu K. 2017. Neural machine translation in linear time. bioRxiv. https://doi.org/10.48550/arXiv.1610.10099
113. Yang KK, Fusi N, Lu AX. 2024. Convolutions are competitive with transformers for protein sequence pretraining. Cell Syst 15:286–294. https://doi.org/10.1016/j.cels.2024.01.008
114. Doroghazi JR, Albright JC, Goering AW, Ju K-S, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW. 2014. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol 10:963–968. https://doi.org/10.1038/nchembio.1659
115. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021
116. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116
117. Sanchez S, Rogers JD, Rogers AB, Nassar M, McEntyre J, Welch M, Hollfelder F, Finn RD. 2023. Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. bioRxiv. https://doi.org/10.1101/2023.05.23.540769
118. Barrett SE, Mitchell DA. 2024. Advances in lasso peptide discovery, biosynthesis, and function. Trends Genet 40:950–968. https://doi.org/10.1016/j.tig.2024.08.002
119. Burkhart BJ, Hudson GA, Dunbar KL, Mitchell DA. 2015. A prevalent peptide-binding domain guides ribosomal natural product biosynthesis. Nat Chem Biol 11:564–570. https://doi.org/10.1038/nchembio.1856
120. Li H, Ding W, Zhang Q. 2024. Discovery and engineering of ribosomally synthesized and post-translationally modified peptide (RiPP) natural products. RSC Chem Biol 5:90–108. https://doi.org/10.1039/d3cb00172e
121. Russell AH, Truman AW. 2020. Genome mining strategies for ribosomally synthesised and post-translationally modified peptides. Comput Struct Biotechnol J 18:1838–1851. https://doi.org/10.1016/j.csbj.2020.06.032
122. Kloosterman AM, Shelton KE, van Wezel GP, Medema MH, Mitchell DA. 2020. RRE-Finder: a genome-mining tool for class-independent RiPP discovery. mSystems 5:e00267-20. https://doi.org/10.1128/mSystems.00267-20
123. Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–8. https://doi.org/10.1093/nar/gki408
124. Challis GL, Naismith JH. 2004. Structural aspects of non-ribosomal peptide biosynthesis. Curr Opin Struct Biol 14:748–756. https://doi.org/10.1016/j.sbi.2004.10.005
125. Stachelhaus T, Mootz HD, Marahiel MA. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol 6:493–505. https://doi.org/10.1016/S1074-5521(99)80082-9
126. Huang J, Gao Q, Tang Y, Wu Y, Zhang H, Qin Z. 2023. A deep learning model for type II polyketide natural product prediction without sequence alignment. Digital Discovery 2:1484–1493. https://doi.org/10.1039/D3DD00107E
127. Gavriilidou A, Kautsar SA, Zaburannyi N, Krug D, Müller R, Medema MH, Ziemert N. 2022. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. Nat Microbiol 7:726–735. https://doi.org/10.1038/s41564-022-01110-2

128. Loh JS, Mak WQ, Tan LKS, Ng CX, Chan HH, Yeow SH, Foo JB, Ong YS, How CW, Khaw KY. 2024. Microbiota-gut-brain axis and its therapeutic applications in neurodegenerative diseases. Signal Transduct Target Ther 9:37. https://doi.org/10.1038/s41392-024-01743-1
129. Tegegne HA, Savidge TC. 2025. Leveraging human microbiomes for disease prediction and treatment. Trends Pharmacol Sci 46:32–44. https://doi.org/10.1016/j.tips.2024.11.007
130. Metwaly A, Reitmeier S, Haller D. 2022. Microbiome risk profiles as biomarkers for inflammatory and metabolic disorders. Nat Rev Gastroenterol Hepatol 19:383–397. https://doi.org/10.1038/s41575-022-00581-2
131. Jarchum I, Jones S. 2015. DREAMing of benchmarks. Nat Biotechnol 33:49–50. https://doi.org/10.1038/nbt.3115