nature reviews microbiology

https://doi.org/10.1038/s41579-024-01141-y

Review article

Check for updates

Microbial secondary metabolites: advancements to accelerate discovery towards application

Jaime Lorenzo N. Dinglasan, Hiroshi Otani, Drew T. Doering, Daniel Udwary & Nigel J. Mouncey

Abstract

Microbial secondary metabolites not only have key roles in microbial processes and relationships but are also valued in various sectors of today's economy, especially in human health and agriculture. The advent of genome sequencing has revealed a previously untapped reservoir of biosynthetic capacity for secondary metabolites indicating that there are new biochemistries, roles and applications of these molecules to be discovered. New predictive tools for biosynthetic gene clusters (BGCs) and their associated pathways have provided insights into this new diversity. Advanced molecular and synthetic biology tools and workflows including cell-based and cell-free expression facilitate the study of previously uncharacterized BGCs, accelerating the discovery of new metabolites and broadening our understanding of biosynthetic enzymology and the regulation of BGCs. These are complemented by new developments in metabolite detection and identification technologies, all of which are important for unlocking new chemistries that are encoded by BGCs. This renaissance of secondary metabolite research and development is catalysing toolbox development to power the bioeconomy.

Sections

US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
e-mail: nmouncey@lbl.gov

Introduction

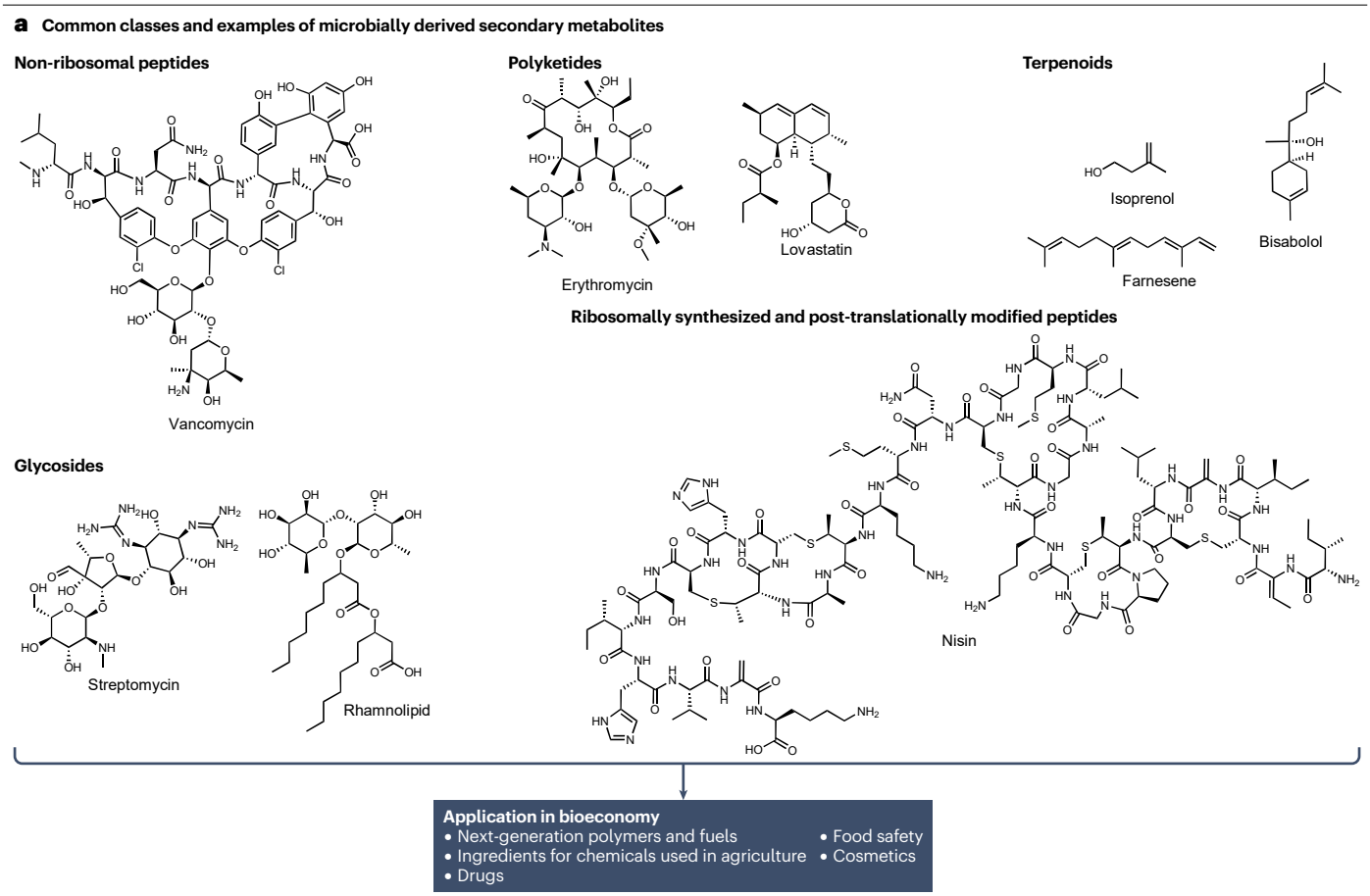
The discovery of new natural products with unique or enhanced bioactivities is integral to the advancement of the bioeconomy (defined here as the knowledge-based production and utilization of biological resources, innovative biological processes and principles to sustainably provide goods and services across all economic sectors[1]). Secondary metabolites are a large group of natural products that organisms synthesize as they interact with their environments[2,3]. Many of these possess bioactivities that are beneficial to society (for example, antibiotics, anti-inflammatory agents and pesticides)[3]. Today, approximately 50% of approved small-molecule therapeutics and active ingredients in crop protection are or are derived from natural products[4,5]. The use of secondary metabolites has also been increasingly explored as non-toxic and eco-friendly alternatives to petrochemicals[6] and are beginning to or can potentially transform petroleum-based industries (for example, cosmetics[7,8], fuels[7,9] or materials[8,10]).

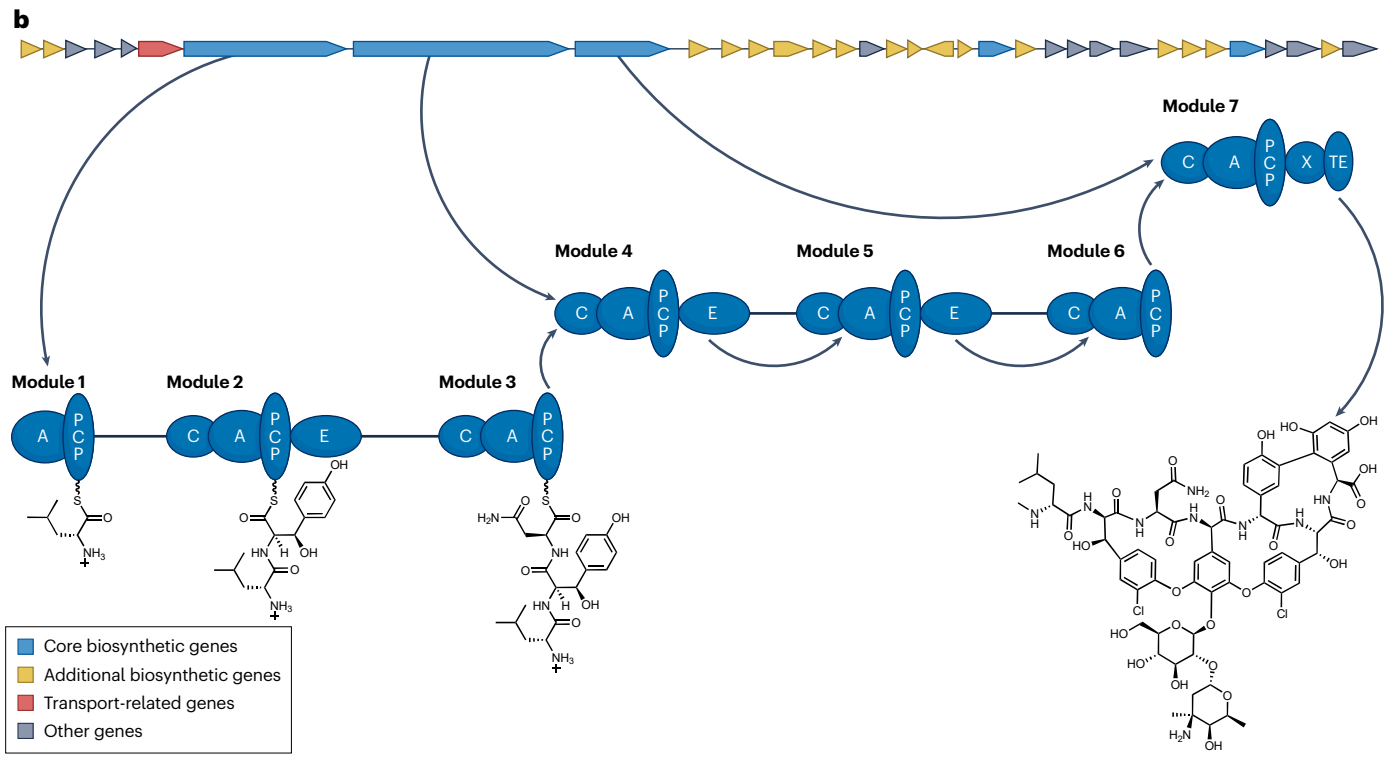
Microorganisms produce diverse secondary metabolites and almost half of those that are known exhibit some biological activity[11,12] (Fig. 1a). The major classes of microbial secondary metabolites include non-ribosomal peptides, polyketides, ribosomally synthesized and post-translationally modified peptides (RiPPs), glycosides, and terpenoids (Fig. 1a), and some of their current or potential market applications are introduced in Box 1. Numerous genes are involved in the production of a given secondary metabolite, with most of these genes often co-localized on bacterial or fungal genomes, forming a biosynthetic gene cluster (BGC)[13] (Fig. 1b). These genes are often co-expressed under tight regulatory control. Generally, BGCs typically comprise genes that encode core enzymes that build the natural product scaffold, tailoring enzymes that modify the biochemistry of compounds, transporters that secrete the product into the environment, and transcription factors that regulate the expression of various genes within the cluster. Depending on the type of metabolite synthesized, additional genes present in the cluster may be responsible for the production of precursor molecules and cofactors or for conferring host resistance to the secondary metabolite (for example, those with antibiotic activity). Around 33,300 secondary metabolites have been isolated from microorganisms, with *Streptomyces*, which are filamentous bacteria, and *Aspergillus*, which are filamentous fungi, as the largest bacterial and fungal sources of secondary metabolites, respectively[14]. The genus *Streptomyces* belongs to a group called actinomycetes, along with other filamentous bacteria like *Saccharopolyspora*, *Micromonospora* and *Amycolaptosis*, that also synthesize known natural products. However, active BGCs have also been characterized in non-actinobacterial genera (for example, *Bacillus*, *Lactococcus* and *Paenibacillus*). Moreover, fungal producers have also been identified within the genus *Penicillium*, from which the well-known secondary metabolite penicillin was isolated. Early efforts mostly discovered these bacteria and fungi from soils but microbial BGCs have since been identified in other terrestrial environments as well as marine, freshwater, host-associated and engineered microbiomes[15–19]. BGCs have also, more recently, been identified in largely uncultured bacteria and unexpected fungal taxa (for example, the budding yeast genus *Kluyveromyces*)[20]. Intriguingly, most of these encode chemistries that are still unknown but are coming to light with the evolution of new computational and empirical strategies for natural product discovery[13].

Notably, during the golden age of antibiotic discovery (1940s–1970s), the process of sourcing natural products from microorganisms typically entailed acquiring and cultivating microorganisms and then testing their fermentation broths or cell extracts for desirable activities[3], which led to the identification of 200–300 new natural products per year in the 1970s[21]. However, this method would eventually yield high rediscovery rates, which led to a downturn in the search for novel natural product structures in the environment[21]. The later introduction of genomics in the 2000s rekindled interest in this area by unveiling the broader biosynthetic potential of microorganisms. Early sequencing efforts revealed that certain microorganisms — especially actinomycetes — possess above tenfold more BGCs than routine cultivation and detection techniques could identify[3]. With the rapid progression of genomic sampling and sequencing technologies (for example, metagenomics) as well as the evolution of platforms that search genomes for BGCs (for example, genome-mining tools), it has become clear that BGCs in the environment remain an abundant yet largely untapped resource for new bio-products. For instance, a recent analysis of a large actinobacterial genomic data set — comprising 4,800 isolate genomes from public databases and 824 newly sequenced actinobacterial genomes from the Genomic Encyclopedia of Bacteria and Archaea (GEBA) initiative — predicts almost 80,947 BGCs of which only an estimated 3.2% have known functions[22]. Such realizations have prompted the development of empirical approaches that prioritize the sampling of unique BGCs from the environment and the characterization of new chemistries in the laboratory. For example, tactics using metagenomics can identify novel BGCs from uncultivated microorganisms[16] and advanced heterologous expression-based techniques are enabling the production of their cognate compounds in a culture-independent fashion[23]. The integration of these improved computational and high-throughput experimental approaches is key to a new age of rapid natural product discovery — the 'Platinum Age' — that will facilitate bioeconomic growth[2,24].

In this Review, we explore innovations that facilitate rapid microbial secondary metabolite discovery, focusing on recent techniques for the prediction of BGCs from genomic databases, for the isolation of novel BGCs from the environment, and to produce and detect the chemical products of these sequences in the laboratory. Some of the topics we cover have been reviewed separately and more comprehensively by others[23,25–31]. We aim to add to these discussions by compiling and describing recent advancements that could be integrated into workflows for rapid natural product discovery. We envision such workflows

Fig. 1 | Overview of microbial secondary metabolites, biosynthetic gene clusters and their biosynthesis. a, Microorganisms are abundant in diverse environments and are a rich source of secondary metabolites. Polyketides, non-ribosomal peptides, ribosomally synthesized and post-translationally modified peptides, glycosides and terpenoids are some of the major classes of microbial natural products. Secondary metabolites have current and potential applications in various sectors of the growing bioeconomy as next-generation polymers and fuels, drugs, and ingredients for chemicals used in agriculture, food safety, and cosmetics, to name a few examples. **b**, An illustration of the biosynthetic gene cluster (BGC) of vancomycin, a non-ribosomal peptide, as an example of BGC architectures. The BGC encodes core enzymes that catalyse the vancomycin peptide backbone as well as additional biosynthetic enzymes for tailoring reactions, transporters and other uncharacterized proteins. Three of the core enzymes are multi-domain, multimodular synthetases[138] and each module uses its domains to incorporate an amino acid monomer into the growing peptide, alter its structure, and deliver it to the succeeding module. A, adenylation domain; C, condensation domain; E, epimerization domain; PCP, peptidyl carrier protein; TE, thioesterase; X, oxygenase-recruiting domain[138].


a Common classes and examples of microbially derived secondary metabolites
Non-ribosomal peptides
Vancomycin
Polyketides
Erythromycin
Lovastatin
Terpenoids
Isoprenol
Bisabolol
Farnesene
Ribosomally synthesized and post-translationally modified peptides
Nisin
Glycosides
Streptomycin
Rhamnolipid
Application in bioeconomy
• Next-generation polymers and fuels
• Ingredients for chemicals used in agriculture
• Drugs
• Food safety
• Cosmetics


b
Module 1
Module 2
Module 3
Module 4
Module 5
Module 6
Module 7
A
PCP
C
E
X
TE
Core biosynthetic genes
Additional biosynthetic genes
Transport-related genes
Other genes

Box 1 | Microbial secondary metabolites and their applications

Transitioning economies to bioeconomies, where biological processes are harnessed across various economic sectors, can address challenges in human health, food security and climate change, among others[139,140]. Microbial secondary metabolites already have a crucial role in this transition given that many of these bioactives have had profound impacts on medicine and agriculture, especially in the past century[3]. These compounds have been increasingly incorporated into cosmetic products[141,142] and emerging applications have the potential to transform other industries[9,143]. Here, we briefly introduce some of the major classes of microbial secondary metabolites and their current or potential market applications. The genetics and enzymology behind the synthesis of these natural products in microorganisms have been comprehensively reviewed elsewhere[144–152].

Non-ribosomal peptides and polyketides are among the most abundant and widely distributed microbial secondary metabolites. Non-ribosomal peptides are synthesized by non-ribosomal peptide synthetases, which are megasynthetases with multi-domain modules that condense amino acids in a sequential fashion, assembling peptide chains that are also modified (for example, oxidation or methylation) in the process[153,154] (Fig. 1b). The broad-spectrum antibiotic vancomycin originally isolated from *Streptomyces orientalis* is a popular example of this class and is a 'last resort' against antibiotic-resistant strains[155]. Polyketides are assembled by polyketide synthases (PKSs), which condense acyl-CoA units to form a chemical scaffold that is subsequently modified and cyclized[145]. Examples of polyketides that are widely used today include the antibiotic erythromycin isolated from *Saccharopolyspora erythraea* and lovastatin from *Aspergillus terreus* (1970s), which is known for its cardioprotective, neuroprotective and anticancer properties[3,145].

Notably, the modules and domains of non-ribosomal peptide synthetases and certain PKSs can be rearranged or genetically engineered, enabling the production of new secondary metabolite structures[156,157]. This paradigm of 'combinatorial biosynthesis' has been successfully implemented by pharmaceutical companies to synthesize new drugs[3] and opens doors for utilizing polyketides and non-ribosomal peptides in other bio-industries[9,154]. For instance, PKS engineering is now being explored for the production of circular bioplastics and was recently applied to combinatorially synthesize bio-polymer variants with altered properties[10].

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are ribosomally synthesized peptides that undergo structural diversification via post-translational modifications, lending to the broad range of bioactivities observed within this group of metabolites. As peptides, RiPPs can have short precursor coding sequences that were previously overlooked but the in silico methodology for identifying RiPP genomic signatures has significantly improved[59]. In turn, their recognition as a class of potent bioactive compounds has only significantly grown over the past few decades and so their commercial adoption remains limited[158]. An exception is nisin from *Lactococcus lactis*, which belongs to the lantibiotic sub-class of RiPPs and is used as a food preservative in more than 48 countries[159]. Substantial ongoing research on RiPPs will advance these metabolites as key natural products in future markets. For example, lasso peptides, a type of RiPP, are being explored as therapeutics because their lasso-like topologies confer them with high heat, pH and proteolytic stability[160].

Glycosides encompass a wide range of secondary metabolites with glycone and aglycone components[161]. Glycosyltransferases are an example of enzymes that transfer a sugar moiety to the aglycone (that is, glycosylation)[162]. Structural differences in the two components influence an assortment of bioactivities. For example, aminoglycosides are common antibiotics comprising amino-modified sugars linked to an aminocyclitol structure[163], whereas glycolipids incorporate a carbohydrate and lipid component. A common aminoglycoside is streptomycin, the main antibiotic for plant disease control[164]. Industrial glycolipids include rhamnolipids, whose amphiphilic structures make them strong biosurfactants in cosmetic and food safety formulations and in oil cleaning and recovery[165].

Terpenoids are another class of secondary metabolites found in current markets and are synthesized from isoprene units. Almost all bacteria carry pathways for synthesizing these building blocks from central precursors and bacterial terpenome exploration is an evolving field[166]. Although terpenoids that are currently commercially available mostly originate from plants, some of these are synthesized in engineered microorganisms. Amyris has successfully scaled up the fermentation of farnesene and bisabolol[8]. Farnesene feeds into various industries as a versatile precursor to adhesives, surfactants, fragrances, polymers and crop protection agents, among others[7]. Bisabolol is an ingredient in cosmetic products like lotions and eye creams. Additionally, microbial pathways have been engineered to synthesize alcohols from terpenoid precursors that may be used in next-generation fuel blends[7].

to facilitate the accelerated discovery of secondary metabolites and, consequently, bioeconomic development (Fig. 2).

Predicting novel BGCs

The post-genomics era is marked by extensive biological data. These are housed in widely utilized repositories, rich with unique BGCs and novel chemical structures that are of particular interest to microbial secondary metabolite researchers. Specifically, the National Center for Biotechnology Information (NCBI) Datasets[32], the Department of Energy Joint Genome Institute (JGI) Integrated Microbial Genomes and Microbiomes (IMG/M) data management system[33], and the European Nucleotide Archive[34] keep and refine genomic sequences from both cultured and uncultured microorganisms. In addition, the Global Natural Product Social Molecular Networking Mass Spectrometry Interactive Virtual Environment (GNPS-MassIVE)[35] is a database for raw and annotated mass spectra, the informational outputs of mass spectrometry (MS)-based metabolomics experiments.

Extracting relevant biosynthetic information from these large databases requires both BGC-oriented databases[36] and genome-mining platforms[27], which typically complement each other in workflows to identify putatively novel BGCs[25]. Genome-mining tools populate BGC-focused repositories with predicted BGCs[33,37], and these databases support comparative analyses with empirically validated BGCs for dereplication[38].

BGC-oriented databases

The Minimum Information about a Biosynthetic Gene Cluster (MIBiG) repository is a widely used, community-driven database that contains standardized and curated annotation information for validated microbial BGCs and their products[38]. MIBiG was recently updated to v4.0 with new entries as well as a large-scale validation and re-annotation of existing entries and now carries ~3,000 secondary metabolite clusters in total. MIBiG is also cross-linked with the Natural Products Atlas (v2.0), a comprehensive database of bacterial and fungal secondary metabolites as described in peer-reviewed literature[14].

Databases like the established Antibiotics and Secondary Metabolite Analysis Shell (antiSMASH)[37] and the new Secondary Metabolites Collaboratory (SMC)[39] house BGCs that have been predicted by genome-mining analyses. AntiSMASH is populated with predictions of BGCs that are mined from RefSeq genomes with antiSMASH (see next section), and currently houses >230,000 secondary metabolite regions. The SMC portal aims to provide long-term support for the

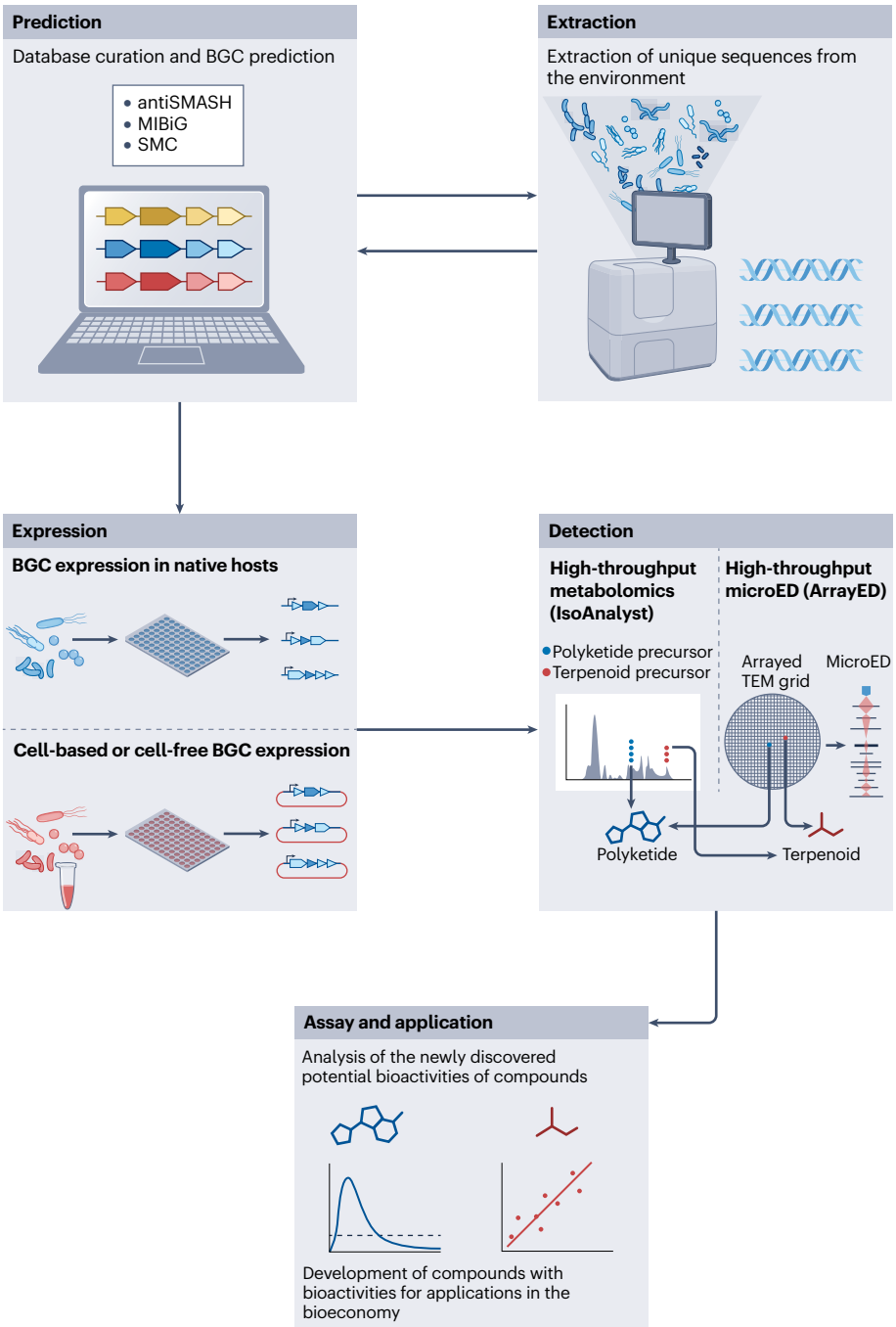


Fig. 2 | Example workflow for microbial secondary metabolite discovery. Prediction: unique biosynthetic gene clusters (BGCs) can be prioritized using in silico genome-mining tools and systems like Antibiotics and Secondary Metabolite Analysis Shell (antiSMASH) or the Secondary Metabolites Collaboratory (SMC) that predict or store unknown or previously uncharacterized sequences from curated genomic databases. Extraction: these analyses inform and are informed by genomic sampling approaches for extracting novel BGCs from the environment, which conversely help populate genomic databases with new information. Expression: unique BGCs are then activated in the laboratory either in their native producers or through heterologous expression in cell-based or cell-free platforms. Detection: analytical chemistry techniques are leveraged to analyse new secondary metabolites that are synthesized from active BGCs. Recent advanced techniques include IsoAnalyst and ArrayED. Assay and Application: the potential bioactivities of newly identified compounds are empirically assayed. Those with valuable functions can be further developed for the market.

natural products research community by providing complete sequence and annotation data for BGCs across all kingdoms of life. The SMC, currently in its first version, houses BGC data from >1,300,000 genomes and employs two different genome-mining tools, which together have predicted >13,000,000 BGC regions. In addition, the BiG-FAM database stores information on gene cluster families (GCFs), which are networks of BGCs with architectural and sequence similarities, and are thus likely to produce related, if not similar, metabolites[40]. Knowledge of GCFs enables the prioritization of BGCs that encode potentially novel compounds or intriguing variants of known chemicals, making BiG-FAM a valuable reference for reducing rediscovery rates.

Metabologenomics is an emerging field, where available metabolomics data are analysed in the context of complementary genomics data and vice versa[41,42]. The Paired Omics Data Platform (PoDP)[43] is a new community-driven resource that pairs available MS data in libraries like GNPS-MassIVE[35] with genome and metagenome assemblies for a given strain. Linking these data can connect previously uncharacterized MS signatures to their potential source BGCs, facilitating more rapid BGC annotation and dereplication[41,44]. The database currently has >100 unique links between BGCs and MS/MS spectra.

Genome-mining tools

AntiSMASH is the most widely used platform for accessing and analysing BGCs from bacterial, fungal and archaeal genomes[45]. AntiSMASH identifies different BGC types in a query sequence using profile hidden Markov models that search genomes based on manually curated and validated 'rules'. These define essential biosynthetic functions that need to be present in a genomic region to qualify as a certain BGC type. AntiSMASH 7.1 can now identify 81 cluster types, includes improved annotations of NRPSs and PKSs and evaluations of novel RiPPs, and integrates a tool for predicting transcription factor-binding sites. NaPDoS2 is a specialized platform to predict PKSs and NRPSs by searching for enzymes with sequence similarity to known PKS-related and NRPS-related domains and then analysing their phylogeny, making it especially useful for predicting these types of BGCs in incomplete genomes, metagenomes and PCR amplicon data[46]. Submissions to NaPDoS2 can be made directly through the results page in antiSMASH 7.0, streamlining the analysis of antiSMASH hits for these particular domains[45].

Increasingly available genomic collections can be analysed with genome-mining tools to gain insights into the broader biosynthetic potential within microbial taxa. A newly developed tool that supports this type of analysis is Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE). It groups BGCs, detected via antiSMASH or extracted from MIBiG, into GCFs[36]. BiG-SCAPE generates GCFs by calculating a pairwise distance matrix for BGCs based on their shared types of domains, pairs of adjacent domains and sequence homologies. While sensitive, BiG-SCAPE is computationally expensive. Biosynthetic Gene Super-Linear Clustering Engine (BiG-SLICE) was later introduced to cluster BGCs based on a non-pairwise calculation that scales better with the current wealth of genomic data[47] and was used to populate the aforementioned BiG-FAM database[40].

With the rapidly growing availability of BGC data in public repositories, evolutionary patterns can also now be deduced from mined BGC sequences, which can in turn be leveraged in novel genome-mining tools to uncover BGCs with novel or enhanced bioactivities compared to their well-characterized counterparts[48]. For instance, CORASON conducts a phylogenomic analysis on BGCs containing homologues of a query gene[36]. This was released with BiG-SCAPE to uncover sequence diversity among BGCs within a GCF and, consequently, the diversity among their possible product structures. Antibiotic Resistance Target Seeker (ARTS) prioritizes BGCs responsible for the synthesis of antibiotics with new mechanisms, following the observation that microorganisms have evolved resistance to the antibiotics they produce[49]. Specifically, it has been previously observed that an antibiotic producer can carry a second copy of the gene for the antibiotic's target protein within the antibiotic-producing BGC that enables host survival[50]. ARTS uses antiSMASH to identify BGCs in a query sequence, then uses a reference set of bacterial genomes to find genes that have undergone duplication events and to phylogenetically screen for genes that have undergone horizontal gene transfer. ARTS initially came with a reference set of only complete genomes from actinobacteria but ARTS 2.0 has higher taxonomic coverage and now includes metagenomes[49]. Fungal Bioactive Compound Resistant Target Seeker (FunARTS) has also been developed to analyse fungal reference genomes through a similar pipeline[51]. In addition, non-canonical BGCs, which are cluster types that are excluded from genome-mining pipelines like antiSMASH, can be identified by searching for common regulatory motifs. Some genes of the same cluster are co-regulated by the same transcription factor and thus share a promoter motif. Following this, a genome-mining pipeline was built to identify BGCs carrying isocyanide synthases (ICS), which are the newest class of backbone enzymes in fungi. The pipeline searches for ICS-specific domains and their associated motifs and then defines BGC boundaries by searching for neighbouring genes with similar motifs. The predicted BGCs are then grouped into GCFs, and highly conserved and co-localized genes within each GCF are defined as the core biosynthetic machinery for that cluster[52].

The aforementioned platforms rely on rule-based methods[48] to predict BGCs, precluding the prediction of truly novel cluster types for which empirical evidence does not yet exist[48,53,54]. New complementary tools enabling the latter take advantage of large genomic data sets to predict potentially new BGC classes without a pre-defined set of rules. We highlight a few recent advancements here that take advantage of artificial intelligence and point the reader to a recent comprehensive review[55]. EvoMining is an evolutionary genome-mining platform that can find novel BGCs by tracking the divergence of genes from a taxon's conserved metabolism and its recruitment into a cluster[54]. Leveraging the observation that many of the core enzymes involved in natural product biosynthesis evolved by gene duplication of enzymes involved in central metabolic pathways, EvoMining identifies novel BGCs by using homology searches and phylogenetic analyses to identify gene expansions in central metabolic enzyme families that have high sequence similarity to a manually curated database of known BGC sequences (like MIBiG) and are thus likely to be recruited into secondary metabolite synthesis. However, the identification of these recruitments is dependent on existing BGC data, disqualifying potential gene expansions to poorly characterized BGC architectures. Meanwhile, DeepBGC[53] and SanntiS[56] use deep-learning approaches to identify novel BGCs. Briefly, they utilize neural networks trained on a positive set of BGC-containing sequences and a negative set of sequences that lack known BGCs. The networks learn complex features and patterns that are associated with BGCs that may be missed by rule-based approaches. After assigning protein domains to genes within a query sequence, the trained algorithms can score candidate BGC-containing regions of the genome. An important shortcoming of these models is that they are trained on available BGC data sets that remain heavily biased towards well-studied taxa, like *Streptomyces*. Incorporating data sets from phylogenetically distant producers can

help uncover novel signatures or deconvolute common genetic associations in one taxon that seem predictive of but do not actually result in secondary metabolite synthesis[53]. Additionally, both DeepBGC and SanntiS were designed to analyse bacterial data. However, DeepBGC underperforms compared to a rule-based approach like fungiSMASH, the fungal version of antiSMASH, presumably because fungal BGC architectures are much more varied. TOUCAN is also a supervised machine-learning approach that outperforms both DeepBGC and fungiSMASH when predicting BGCs in *Aspergillus* genomes[57]. This is partially because it uses a supervised machine-learning approach that discriminates BGCs based on multiple features from fungal genomic data, specifically: characteristic sequence patterns (*k*-mers), protein domains (Pfam domains) and functional information (gene ontology). Similarly, Gene Cluster prediction with COnditional random fields (GECCO) evaluates the context and order of features like gene ontology terms enriched in bacterial BGCs to predict new BGCs. By assigning features for the model to learn from, GECCO advantageously requires less training than models like DeepBGC, which need to be trained on large data sets to learn complex patterns[58].

Novel BGC prediction accuracy can be improved when the putative sequence can be connected to an uncharacterized chemical signature that has already been empirically detected. Metabologenomics is an approach to genome mining that leverages metabolomics data. HypoRiPP Atlas is a new pipeline introduced in 2023 that specifically mines genomes for novel RiPP BGCs using mass spectra[59]. Specifically, a machine-learning algorithm uses rules to detect BGCs within queried genomes and then generates a combinatorial list of plausible RiPP structures for each cluster. Hypothetical mass spectra are then generated from these predicted structures and used to populate the atlas. By comparing these spectra to GNPS data uploaded to PoDP, connections can be made between hypothetical RiPP structures and those that have been observed from experiments but lack annotation. The source BGC of a putative RiPP with a successful alignment in an experimental MS data set can then be prioritized for downstream functional characterization. A pitfall of this approach is that current algorithms for predicting chemical structures from genomic data are optimized for canonical BGCs, again precluding access to truly novel cluster types[26].

Altogether, BGC prediction is increasingly becoming efficient owing to the rapid evolution of relevant computational tools. Rule-based and pattern-based genome-mining approaches for identifying different cluster types (Fig. 3) can be augmented by non-canonical methods and metabologenomics to dereplicate BGCs for downstream empirical efforts. To aid in the former, there is a significant focus on prioritizing the extraction of unique information from the environment that will enrich BGC-oriented databases and can subsequently be mined for new BGCs.

Isolating uncharacterized BGCs

High microbial secondary metabolite rediscovery rates in the 1900s could be partially attributed to the use of routine culture conditions that select for a subset of microorganisms from environmental samples[3]. In effect, BGCs in uncultivated microorganisms have come to represent 'biosynthetic dark matter'[18,19]. Capturing this information from nature requires cultivation-independent strategies, particularly those with expansive reach[60]. A prevailing tactic is untargeted genome-resolved metagenomics, which involves directly sequencing environmental samples, typically via short-read sequencing, then recovering composite genomes (metagenome-assembled genomes; MAGs) through computational analyses[61] (Fig. 4b). MAGs are typically mined for BGCs to elucidate the biosynthetic potential of unculturable microorganisms within the sample[25]. This has been leveraged to provide the largest catalogue of taxonomically diverse genomes of uncultivated bacteria and archaea from various host-associated, aquatic, terrestrial and engineered microbiomes across the globe[16]. In building this catalogue, 52,515 high-quality MAGs were assembled from >10,000 metagenomes in the IMG/M system. Importantly, they recovered 104,211 putative BGCs from these MAGs using antiSMASH, with 87,187 of these not being identified in the NCBI nucleotide sequence collection. Thus, this catalogue represents a valuable resource for the experimental validation of previously uncharacterized secondary metabolites. The diverse chemistries offered by such a catalogue can be further elucidated with other genome-mining tools like BiG-SLiCE. This tool was used to group 1.2 million antiSMASH-predicted BGCs from a set of genomes that included >20,000 MAGs into GCFs, thus offering insights into the functional niches occupied by BGCs within uncultivated microbial genomes[47].

Untargeted metagenomics can also access novel BGCs that existed in ancient environments. MAGs have been recovered from dental calculus metagenomes sampled from Upper and Middle Palaeolithic humans and Neanderthals[62]. This enabled the resurrection of a biosynthetic pathway identified in MAGs of ancient *Chlorobium* species. An antiSMASH analysis predicted terpene BGCs as well as a butyrolactone BGC, whose architecture was found to be similar to related modern butyrolactone BGCs via BiG-SCAPE. Heterologous expression of the reconstructed cluster led to the identification of two previously uncharacterized paleofurans — 5-alkylfuran-3-carboxylic acids that may have a regulatory role in bacterial photosynthesis.

Notably, MAGs built from short-read sequencing can be incomplete for under-represented taxa or contaminated with erroneous gene fragments[63]. Utilizing long-read sequencing on environmental samples can bypass the need to construct MAGs altogether[64]. The flexible genome, where most BGCs reside, is also better assembled with long reads than with short reads[64]. Both approaches were utilized to generate MAGs from biocrust samples, demonstrating that more BGCs could be recovered from one long-read sequencing library (548 BGCs of which 174 are full-length clusters) compared against two short-read sequencing libraries (359 BGCs between the 2 libraries of which only 9 are full-length clusters)[64]. Long-read metagenomics has recently been leveraged to unearth the grander biosynthetic potential of underexplored, non-actinobacterial phyla such as Acidobacteria and Verrucomicrobia from Antarctic soil samples[65]. Previous sequencing efforts had been able to identify PKS and NRPS domains from these taxa[17,66], but the long-read sequencing approach could additionally mine lasso peptides and bacteriocins. The approach also identified highly divergent BGCs in under-represented actinomycetes: Acidimicrobia and Thermoleophilia. A current disadvantage of long-read (compared to short-read) metagenomic sequencing is its relative cost, due in part to the lack of multiplexed sequencing in current instruments[67]. However, costs can be evaluated against the higher quality of long-read metagenomics data, which require less bioinformatic analyses and offer higher taxonomic resolution in microbiomes[67]. High error rates are also associated with long-read sequencing but innovations like circular consensus sequencing, where multiple passes of a single read are used to build a consensus sequence, can remedy this issue[68].

Untargeted metagenomic sequencing also tends to miss low-abundance microorganisms in the sample, precluding access to rare BGCs[63]. Co-occurrence network analysis of targeted sequences (CONKAT-seq) relieves this problem by generating cloned libraries from

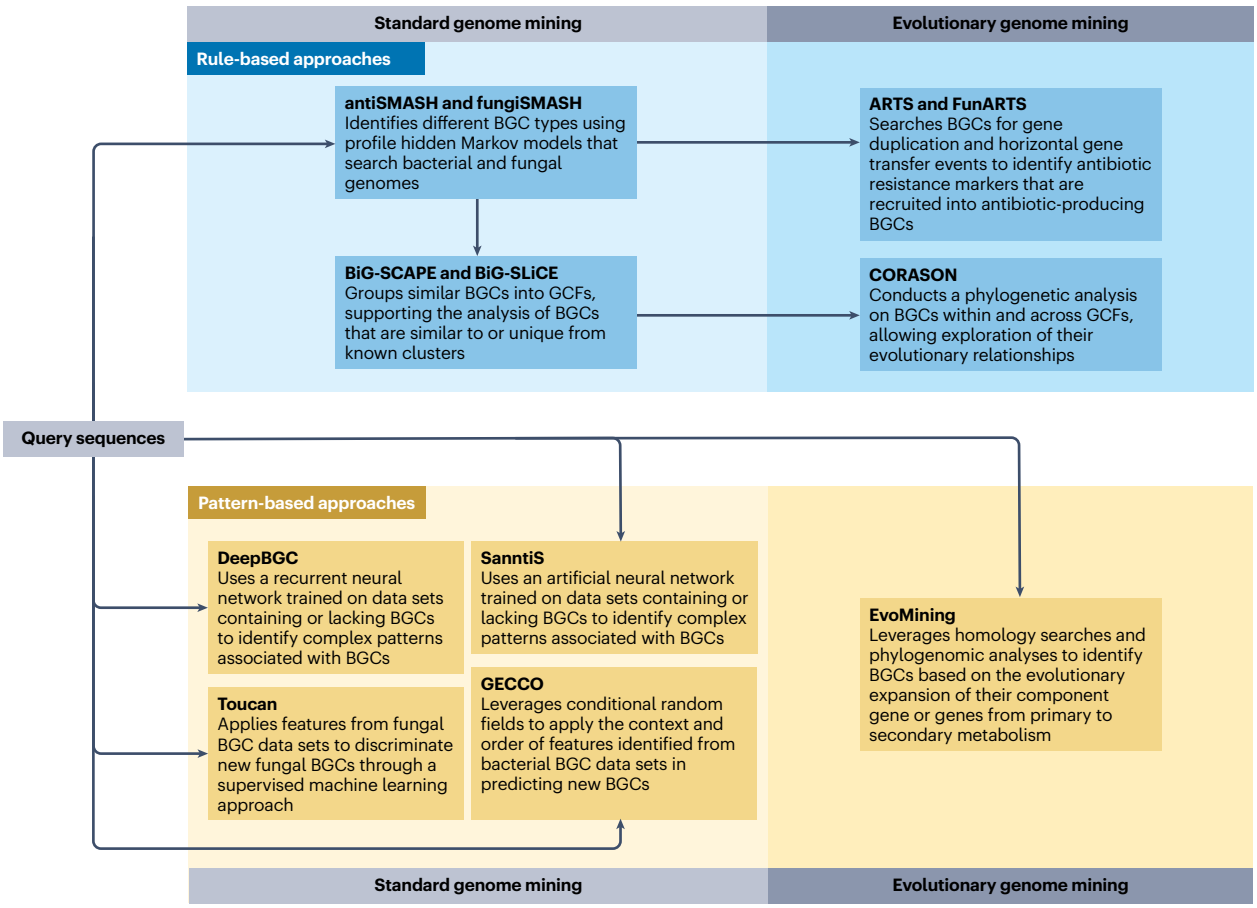


Fig. 3 | Standard and evolutionary genome-mining tools that are currently available for predicting different cluster types. Standard genome-mining tools are designed to predict biosynthetic gene clusters (BGCs) and/or gene cluster families (GCFs) from data sets through pre-defined sets of human-coded rules (rule-based approaches) or by recognizing features in BGC-containing sequences (pattern-based approaches). Evolutionary genome-mining platforms further apply evolutionary principles to discriminate novel BGCs. These may utilize BGC predictions identified through a rule-based standard genome-mining approach like antiSMASH or analyse large genomic data sets for patterns that represent evolutionary events such as EvoMining. antiSMASH, Antibiotics and Secondary Metabolite Analysis Shell; ARTS, Antibiotic Resistance Target Seeker; BiG-SCAPE, Biosynthetic Gene Similarity Clustering and Prospecting Engine; BiG-SLiCE, Biosynthetic Gene Super-Linear Clustering Engine; FunARTS, Fungal Bioactive Compound Resistant Target Seeker; GECCO, Gene Cluster prediction with COnditional random fields.

extracted metagenomes instead[69,70] (Fig. 4a). Briefly, large metagenomic DNA fragments (~40 kbp) are first captured into sub-pooled cosmid cloned libraries. These are then subjected to PCR with degenerate primers that target certain biosynthetic domains and are barcoded for each sub-pool. Given that whole BGCs are likely to be captured into a single large-insert cosmid, clustered genes should co-occur highly in sequenced amplicons. Domains that are likely to belong to a single BGC can therefore be determined by a co-occurrence analysis, and barcodes can be used to locate clones carrying intact BGCs. When compared against an untargeted short-read sequencing approach on soil samples, CONKAT-seq recovers clusters originating from genomes with <0.05% frequency in the metagenomic library[70]. In a more recent study, CONKAT-seq was applied on a collection of *Streptomyces* strains with unsequenced genomes to enrich for rare PKS and NRPS domains[69]. Mobilization of intact BGCs into a heterologous host led to the discovery of conkatamycin, which showed potent antibacterial activity against several multidrug-resistant *Staphylococcus aureus* strains. So far, CONKAT-seq has only been applied with degenerate primers that detect under-represented NRPSs and PKSs, and it would be exciting to see its application towards other BGC classes.

Metagenomic approaches are undoubtedly formidable for recovering novel BGCs from the environment but they do not guarantee the biosynthetic activity of these genes in vivo[71,72]. Carrier protein domains of PKS and NRPS synthases commonly incorporate a pantetheine moiety during polyketide and non-ribosomal peptide synthesis[73] (Fig. 4c). Two recent publications describe hijacking this process to selectively label and sort cells with active synthase activity. Bacteria in tunicate[72] and nudibranch[71] microbiomes were exposed to fluorescently labelled synthetic pantetheine analogues that have higher fluorescence upon incorporation into carrier proteins. Following sample homogenization, a fluorescence-activated cell sorting (FACS) assay was used to isolate synthase-active members of the microbiomes for single-cell genome sequencing and subsequent genome-mining analyses. This workflow detected a novel NRPS–PKS hybrid cluster

from the tunicate microbiome[72] and a beta-lactone cluster in a newly identified genus from the nudibranch skin microbiome[71]. Given that approximately half of the discovered microbial secondary metabolites possess bioactivties[12], probing active enzymes involved in secondary metabolism within the microbiome may narrow down the search for novel natural products that actually confer some ecological function with exploitable bioactivity. However, it is difficult to apply this type of approach towards the synthesis of other secondary metabolites whose modifications require more general enzymatic reactions. That is, aside from NRPSs and PKSs, fatty acid synthases are the only other enzymes known to utilize pantetheine moieties, simplifying selection. Meanwhile, not all other secondary metabolite classes are defined by unique enzymatic reactions but instead require mechanisms broadly applied in other metabolic processes as well (that is, glycosylation and cyclization).

Altogether, these methods are enabling unprecedented access to sequences that encode novel biosynthetic information. While the described workflows effectively identify thousands of BGC sequences, it is typical that only a few (less than five per study) of these undergo downstream characterization[62,69–72]. Empirically connecting BGCs to the molecular structures of their products remains a bottleneck. Innovations that specifically enable higher throughput in the expression of BGCs in the laboratory and the detection of their product metabolites are therefore crucial to accelerate natural product discovery.

Accessing cryptic BGCs

Routine culturing strategies during the golden age of antibiotic discovery also precluded the expression of most BGCs in bacteria and fungi[3]. The synthesis of secondary metabolites can require the complex interaction of several genes within and outside a cluster, all of which are regulated by cues in the native environment of the source organism (for example, nutrient availability, physicochemical conditions or metabolites secreted by other organisms)[74]. Without these cues, BGCs remain inactive or 'silent' during the growth of culturable microorganisms. Newer cultivation techniques and recombinant DNA technologies developed through the late 1900s and 2000s can better access chemistries encoded by BGCs[2,3]. Today, common approaches

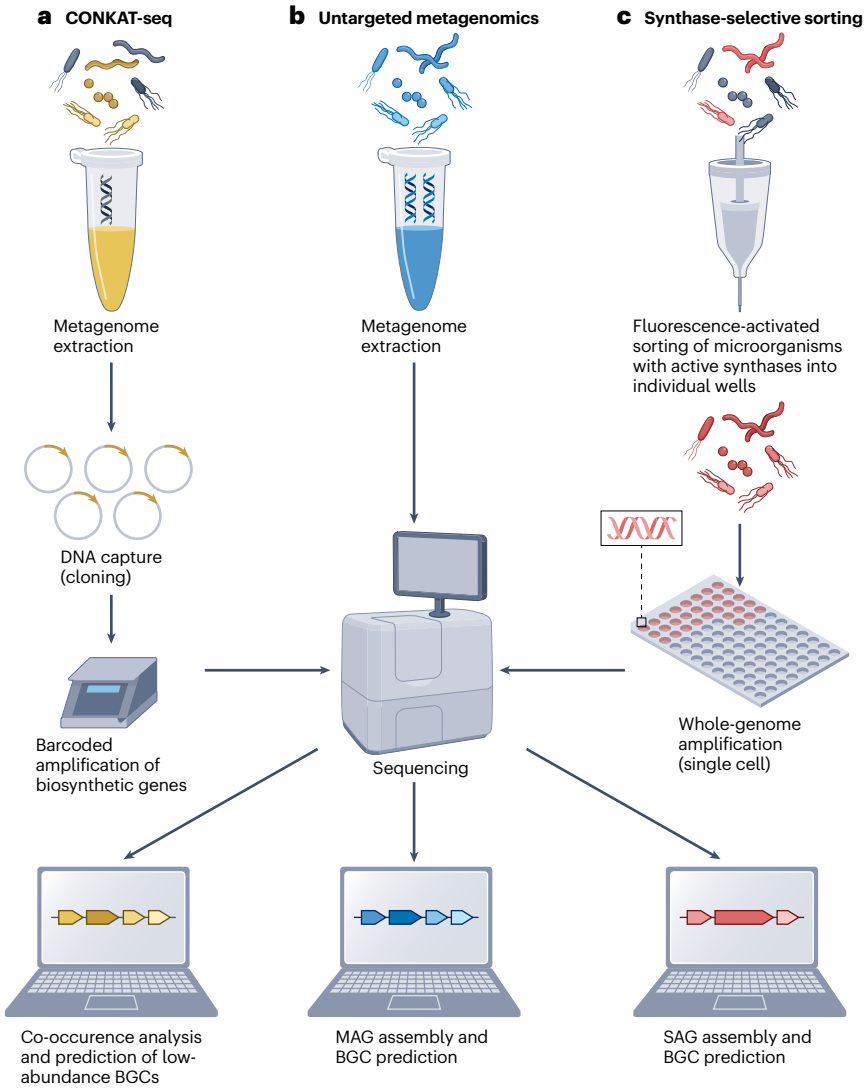
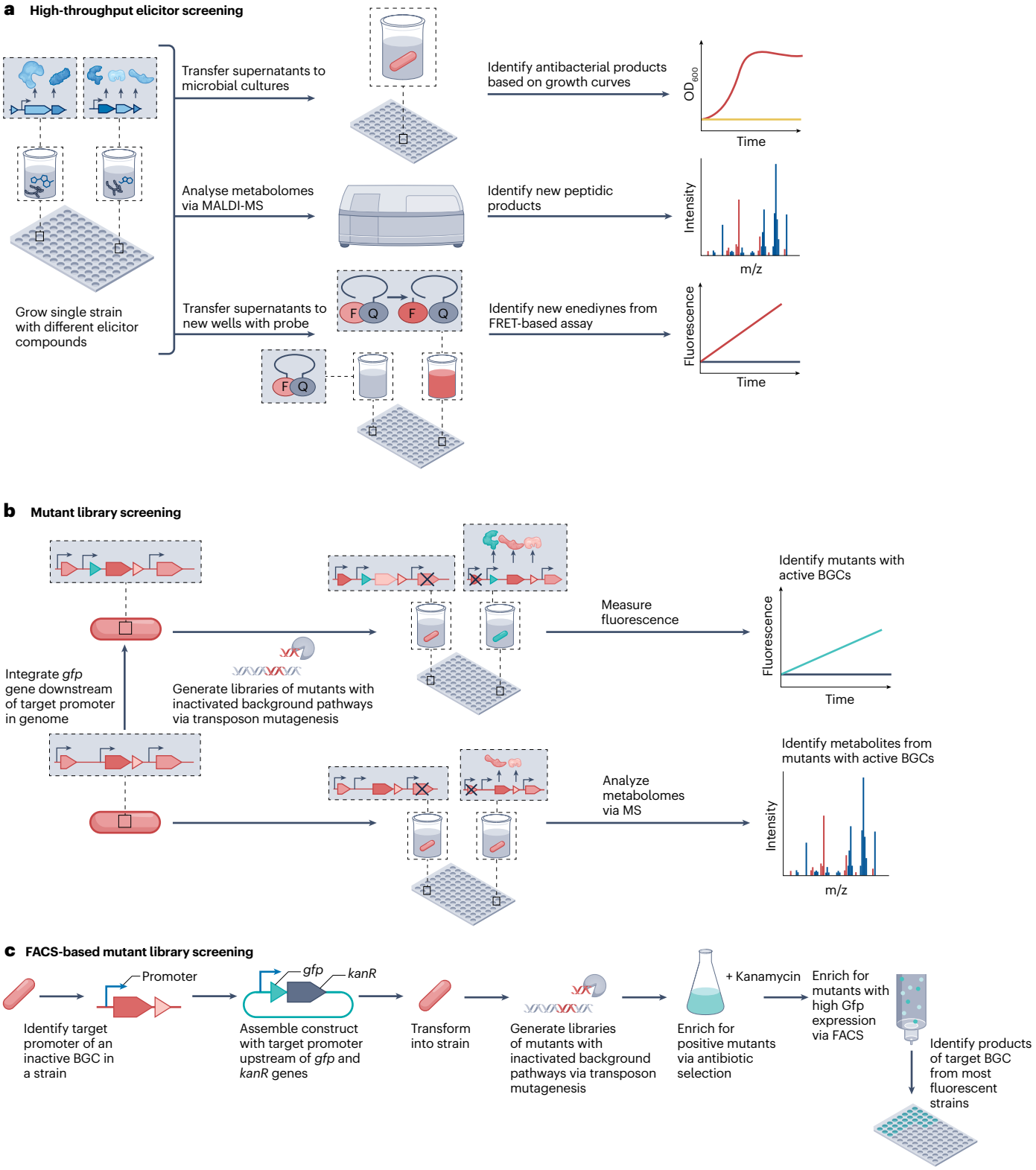


Fig. 4 | Pipelines for extracting BGCs from the environment. a, Co-occurrence network analysis of targeted sequences (CONKAT-seq) involves cloning large DNA sequences directly from metagenomic samples into vectors, which are then sub-pooled. Target biosynthetic domains are PCR amplified from sub-pools of captured DNA with degenerate, barcoded primers. After sequencing amplicons, a statistical analysis is performed on the observation that co-occurring domains should exist on the same amplicons (co-occurrence network analysis). Biosynthetic gene clusters (BGCs) are then constructed in silico and barcodes can be used to locate clones with intact BGCs (yellow arrows). **b**, Untargeted metagenomics involves sequencing extracted metagenomes to generate short-read or long-read libraries. Assembly of these reads into metagenome-assembled genomes (MAGs) enables the prediction of BGCs from genomes of uncultivated microorganisms (blue arrows). **c**, A synthase-selective sorting approach involves labelling a microbial sample with a fluorescent probe that is incorporated by carrier protein domains of polyketide synthases and non-ribosomal peptide synthetases. Fluorescence-activated cell sorting is leveraged to isolate single cells with these active synthases. After whole-genome amplification of sorted cells and sequencing, single amplified genomes (SAGs) are assembled and then subject to BGC prediction (red arrows).



generally involve eliciting production in native producers[75] or heterologous expression in genetically tractable chassis strains[23]. Meanwhile, cell-free systems are being increasingly considered as platforms for BGC expression[29].

Eliciting production in native producers

To promote silent BGC expression, native producers can be grown under various physiological conditions[76], with libraries of various compounds[77] (Fig. 5a), or with other organisms[78]. Microorganisms can

Fig. 5 | High-throughput approaches for activating BGCs in native producers. **a**, High-throughput elicitor screening involves growing a strain with a library of compounds that may elicit the expression of inactive biosynthetic gene clusters (BGCs). Cultures can be screened for active BGC expression and secondary metabolite production through bioactivity-guided assays (top and bottom) or mass spectrometry (MS; middle). The assay depicted on top involves inoculating microbial cultures with supernatants from the elicited cultures and monitoring growth curves to screen for BGCs that produce antibiotics. The approach drawn on the bottom is informed by the known DNA-cleaving activity of enediynes and thus leverages a stem-looped oligonucleotide probe with a fluorophore and a quencher on either end. Supernatants containing enediynes cleave the probe and the resulting fluorescence resonance energy transfer (FRET) can be assayed. MS may also be used as an untargeted approach (middle) to identify profiles of secondary metabolites produced in elicited cultures. **b**, Mutant library screening involves transposon mutagenesis to randomly inactivate genes, some of which may negatively control the expression or activity of a target BGC. A fluorescence-based screening approach (top) involves introducing a *gfp* gene downstream of the promoter of the target BGC into the microorganism of interest followed by generating a transposon mutant library from this strain. Mutations resulting in expression of the target BGC will also lead to the expression of GFP, which enables the quick identification of mutants with activated BGCs. Alternatively, transposon libraries may be prepared from strains without prior genomic integration of a reporter gene (bottom). MS-based approaches enable the detection of secondary metabolites produced in positive mutants. **c**, Utilizing fluorescence-activated cell sorting (FACS) for mutant library screening enables further enrichment of strains with highly activated BGCs. The promoter of a target BGC is first identified in a production strain and then cloned into a construct, where it controls the expression of *gfp* and an antibiotic resistance gene such as *kanR*, which encodes aminoglycoside 3′-phosphotransferase, which manifests kanamycin resistance. The construct is transformed into the production strain, from which a transposon library is generated. Mutants with inactive repressing pathways will thus express *gfp* and *kanR* and can be initially selected for with kanamycin. Those with strong expression from the target promoter are further enriched by their fluorescent properties through FACS, which enables the distinction of high-production strains. MALDI, matrix-assisted laser desorption ionization.

also be genetically manipulated to hijack the regulatory systems that govern BGC expression[79,80] or to delete or mutate genes that inactivate clusters[75] (Fig. 5b,c). In addition to these, recent in-depth investigations into BGC activation and up-regulation are empowering new engineering paradigms for endogenous secondary metabolite synthesis[81,82].

Culture-based approaches. Culturing microorganisms under different conditions, including co-cultivation and adding small molecules to the media, can elicit changes in their secondary metabolomes as initially demonstrated by the overproduction of certain pigments in *Streptomyces coelicolor*[83,84]. These techniques are recognized as One Strain Many Compounds (OSMAC) approaches because they elicit the production of several secondary metabolites from a single strain[76]. Co-cultivating different microbial species[85] or testing different growth conditions for well-established co-cultures[78] offers significant potential for secondary metabolite discovery (Fig. 4a). However, establishing cross-species cultures is a tedious task, complicated by several factors, including possible competition among the microbes and incompatible growth dynamics, and the discovery rate may be further limited by the throughput of the detection method[77]. Meanwhile, high-throughput elicitor screening (HiTES) workflows can screen larger (>500-member) compound libraries and detect more global effects of these elicitors on the secondary metabolite production of a microorganism (Fig. 5a). Recent examples include bioactivity-guided HiTES, which involves monitoring the optical density of *Escherichia coli* cultures that are robotically inoculated with culture supernatants of induced *Streptomyces* sp. to find elicitors of antibacterial clusters[86]. Matrix-assisted laser desorption ionization (MALDI)-MS HiTES was introduced to detect peptidic secondary metabolites specifically[87]. Around 500 metabolomes were analysed by MALDI-MS after cultivating *Streptomyces ghanaensis* with potential elicitors from a ~500-member library. This led to the identification of the antibiotic cinnapeptin, which was connected to an NRPS–PKS hybrid cluster. The latest iteration of HiTES incorporates fluorescence resonance energy transfer (FRET) to identify small-molecule activators of uncharacterized enediyne BGCs in *Streptomyces clavuligerus*[88]. Enediynes function as antibiotics by cleaving DNA, so their activity can be screened in induced cultures using an oligonucleotide probe with a stem-loop structure that carries a donor fluorophore and acceptor quencher on opposite ends. Mostly steroid elicitors were found to result in fluorescent cultures, which led to the characterization of new enediyne-derived compounds called clavulynes.

Genetic manipulation. Random mutagenesis studies have previously demonstrated increased activation of silent BGCs, with mutations being linked to the inactivation of certain background pathways and repressors[75]. Transposon mutagenesis coupled with various screening methods can enable high-throughput active BGC detection (Fig. 5b). A reporter-guided screening technique involves the integration of a GFP reporter downstream of a BGC promoter in *Burkholderia thailandensis*[89]. Followed by transposon mutagenesis, a library of ~500 mutants was screened for higher fluorescence and three BGCs were successfully activated. This method is extended by an MS-guided approach that does not rely on a fluorescence reporter, instead providing untargeted readouts of secondary metabolism in transposon mutants[90]. Chromatography-based MS detection was compared to a rapid imaging-based MS technique since the latter provides higher throughput but lower resolution. These two methods enabled the identification of three new cryptic metabolites in *Burkholderia* strains and their previously unelucidated background regulators.

These reporter-based or MS-based screening methods require cultivating hundreds to thousands of mutants individually, limiting screening throughput. A more scalable technology leverages a FACS to enrich mutant cells based on fluorescence intensity[91] (Fig. 5c). Specifically, a construct carrying the promoter of a target operon upstream of an antibiotic selection marker and the *sfGFP* gene is used to select for and sort mutants with active expression from the target promoter. This workflow was demonstrated with the promoter for an inactive BGC encoding a polyketide pigment, mutaxanthene, in an *Amycolaptosis orientalis* strain, followed by random mutagenesis and FACS-based screening from 50,000 mutants. Although this technique is limited to activating BGCs in transformable microorganisms, the double selection technique enabled the rapid identification of mutants with active pigment production. This method, nevertheless, will further benefit from developing a reporter-independent detection method as the current method requires creating a reporter strain for every BGC and then creating a mutant library in each reporter strain.

CRISPR–Cas systems afford the manipulation of genes and their transcriptional regulation in parallel[92,93]. Recently developed

CRISPR-based technologies that activate silent clusters in native producers are thus promising steps towards high-throughput secondary metabolite discovery in the future. In Streptomycetes, these include a CRISPR–Cas9 approach for knocking in strong promoters upstream of BGCs to activate and upregulate secondary metabolite synthesis in various *Streptomyces* hosts[80] as well as CRISPR interference and activation systems that enable the systematic repression and activation of gene expression, respectively[79]. Further developing CRISPR-based tools already utilized in other filamentous bacteria (for example, *Corynebacterium*, *Mycobacterium* and *Saccharopolyspora*)[94] and in filamentous fungi (for example, *Aspergillus*, *Cordyceps* and *Myceliophthora*)[95] towards multiplexed editing will expand discovery efforts.

Deepening insights into the evolution and regulation of BGCs are facilitating the emergence of new engineering paradigms for improving secondary metabolite synthesis in native producers. High-throughput chromosome conformation capture has been used to dissect 3D chromosomal rearrangements of the model bacterium *S. coelicolor* A3(2) in different growth phases[81]. Evidently, BGC transcription increases with the frequency of chromosomal interactions within its vicinity. This was validated by the insertion of a reporter gene and a natural tetronate BGC into chromosomal regions with varying spatiotemporal contexts. Meanwhile, another group reports on silent cluster activation when their co-evolved clusters are also expressed[82]. They demonstrated this by conducting a pan-genomic analysis of *Streptomyces* species that encode several PKSs to identify a conserved pyrroloquinoline quinone cluster. Engineering this cluster into 11 *Streptomyces* species resulted in the improved production of 34 known metabolites, including polyketides, non-ribosomal peptides and compounds synthesized via hybrid pathways. Indeed, augmenting our understanding of regulatory networks within these microbes will be key to developing new activation strategies.

Recombinant cell-based and cell-free expression

Expressing BGCs directly from recombinant DNA templates is a common way to characterize small sets of candidate BGCs or to overproduce natural products, irrespective of the culturability of the source organism[25]. To use this for high-throughput natural product discovery, biofoundries have been established to express BGCs from a range of source organisms or biosynthetic classes in heterologous hosts[96–99]. FAST-NPS, a fully automated platform for the high-throughput discovery of bioactive natural products, was recently introduced[97]. The workflow takes advantage of ARTS to identify 10–100-kb BGCs that specifically contain resistance genes to circumvent cytotoxicity issues. These sequences were then heterologously expressed in *Streptomyces lividans* TK24, a model *Streptomyces* strain, through a fully automated workflow with a runtime of 20–26 days. This work impressively enabled large BGC expression with reduced time and labour. Of 105 successfully cloned BGCs, 23 natural products could be identified from 12 BGCs, representing successful secondary metabolite synthesis from ~10% of cloned BGCs.

Leveraging several diverse heterologous expression hosts is one way to improve the success rate of secondary metabolite production[23]. This was demonstrated with chassis-independent recombinase-assisted genome engineering (CRAGE), which domesticates genetically recalcitrant bacteria by inserting a landing pad containing a *Cre* gene flanked by mutually exclusive *lox* sites into the genome. The landing pad is then replaced by the *Cre* recombinase with a plasmid carrying a whole BGC flanked by the same *lox* sites. CRAGE was used to successfully integrate 9 BGCs into 25 different gammaproteobacteria. Through this approach, a total of 22 products were synthesized from 6 out of the 9 BGCs, with unique metabolites being identified from 4 of these[100]. Remarkably, hosts that were closely related to the source organism of the BGC had higher yields and more diverse secondary metabolite product profiles. CRAGE can thus be used to expand the panel of bacterial hosts for heterologous BGC expression, which thus far has been limited to a handful of bacteria (that is, mostly *Streptomyces* spp. and *E. coli*)[23] and fungi (mostly *Aspergillus nidulans*, *Aspergillus oryzae* and *Saccharomyces cerevisiae*)[101,102].

Heterologous host environments must provide suitable conditions for protein folding and production, fluxes of metabolic precursors and/or cofactors, and auxiliary enzymes or pathways for post-translational modification when expressing BGCs[103,104]. Gene refactoring is also necessary to circumvent native regulatory mechanisms[105]. Optimizing these parameters to achieve sufficient metabolite synthesis in selected chassis oftentimes requires several iterations of designing, building and testing[106]. This tedious process is a barrier to mobilizing a broad range of species as producer strains, especially those that grow slowly[107,108]. For example, transforming cells with DNA templates is a time-consuming step in cell-based expression[109]. Transformation efficiencies are high in model chassis like *E. coli* and *S. lividans* TK24, which enables their smooth integration into automated workflows[97,110], but protocols are unique to different strain groups and may require optimization for non-model species[100] (Fig. 6). Additionally, although some yeast species are genetically tractable, transformation remains especially tedious for filamentous fungi, which are preferred hosts for natural product discovery given that they harbour a more native environment for fungal BGC expression[102].

Cell-free expression (CFE) systems have emerged to circumvent some of the issues that burden cell-based expression[109,111–113]. These are open, in vitro transcription-translation reactions minimally comprised of three main modules. The first is the transcription-translation (TX-TL) machinery, prepared either as purified components (that is, PURExpress)[114] or from crude cell extracts (that is, lysate-based systems)[111]. The other components are a buffered mixture of salts, cofactors, dNTPs, amino acids and supplements that energize TX-TL as well as a DNA template[115]. CFE offers several unique advantages[113]. All modules are combined by simple pipetting, alleviating the need for transformation, and can be distributed into microlitre volumes to enable parallel expression. Because CFE systems lack survival objectives, resources are directed towards TX-TL, which enables saturated protein production within hours and milligrams per millilitre yields in optimized systems. Optimizing these open and directly manipulable systems also requires significantly less and shorter design–build–test cycles. For instance, necessary precursor metabolites and ancillary enzymes can be added right into the reaction, bypassing extra strain engineering steps to promote catalysis. CFE systems have already been utilized to produce RiPPs[116], non-ribosomal peptides[117–119], terpenoids[120] and a polyketide[121], among others. These BGC-focused efforts are summarized in recent reviews[29,122,123].

Lysate-based cell-free systems are particularly positioned to enable high-throughput BGC expression[29] (Fig. 6). They are inexpensive and their complex reaction environments may already provide most necessary elements for secondary metabolite synthesis. Most demonstrations with BGCs in these platforms used well-established *E. coli* extract systems because of their high yields (that is, 1–4 mg/ml protein in batch reactions)[124]. *Streptomyces*-based systems have been more recently forwarded as alternatives given that their proteomes (that is,

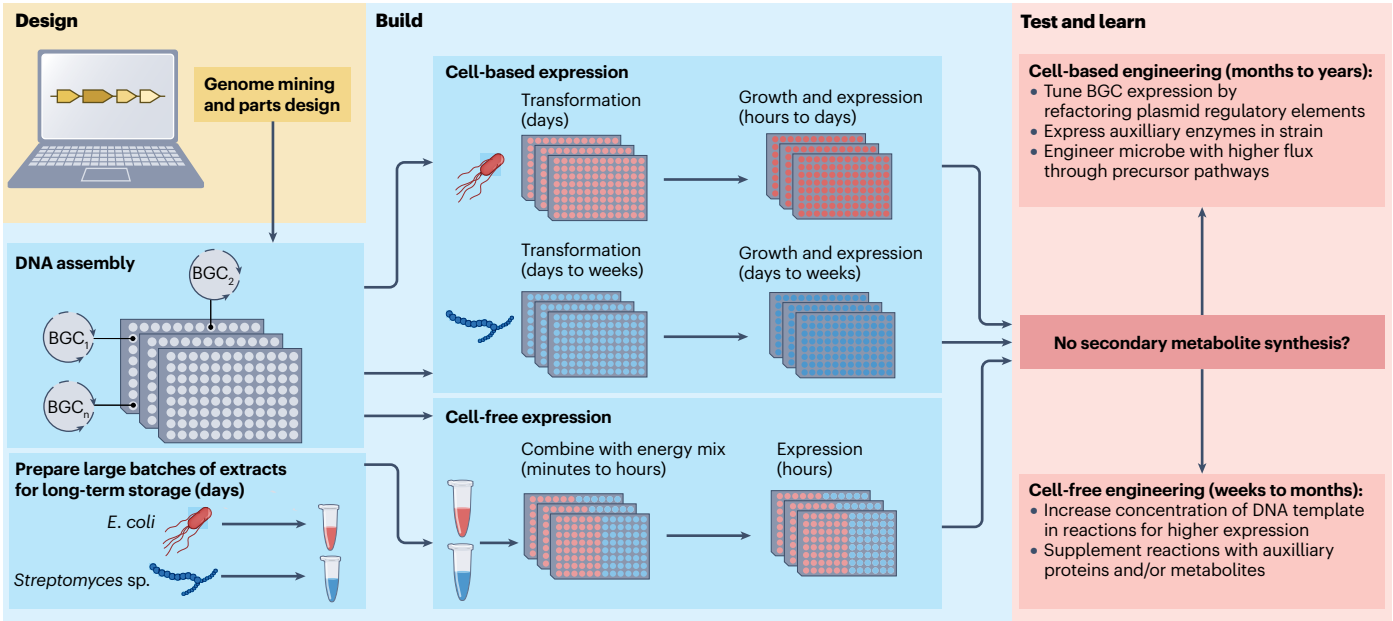


Fig. 6 | Comparison of current cell-based and proposed cell-free workflows for testing and optimizing heterologous BGC expression. The parallel heterologous expression of multiple biosynthetic gene clusters (BGCs) entails iterating through design, build, test and learn stages. In the design stage, the sequences of target BGCs are mined from genomic databases and modified for downstream expression. Biosynthetic gene clusters are assembled from DNA parts in a high-throughput multiplex DNA assembly approach (BGC_1, BGC_2…BGC_n denoting the multiplicity). For conventional cell-based production, the BGCs are transformed into a live heterologous host. To achieve higher success rates of expression, a panel of heterologous hosts can be leveraged. Alternatively, BGCs may be combined with cell lysates of diverse organisms for cell-free expression and biosynthesis. In the test and learn phases, BGC expression in their respective platform is evaluated. If no secondary metabolite synthesis is detected from a particular BGC, cellular or cell-free engineering considerations must be made for the next design–build–test–learn cycle. Optimizing the expression of complex BGCs in cells remains a tedious and time-consuming process constrained by the varying requirements of living systems (for example, growth) and long and variable time frames, whereas cell-free engineering bypasses or minimizes these constraints and is further simplified by, for example, the open environment of these in vitro reactions. *E. coli*, *Escherichia coli*.

auxiliary pathways or post-translational modification enzymes) and metabolomes (that is, high G+C (%) tRNA pools or precursor loads) are likely to be more conducive to secondary metabolite synthesis[122]. *S. lividans*[125] and *Streptomyces venezuelae*[126] systems can already synthesize between 0.2 and 0.5 mg/ml protein and can express high G+C (%) genes robustly – sometimes with higher solubilities than an *E. coli*-based platform[127]. In line with this, an important advantage of lysate-based systems is that they can be optimized from extracts of taxonomically diverse source organisms that have less extensive genetic toolkits[128,129]. Following findings from the CRAGE study[100], a selection of diverse CFE environments for testing BGC expression may be key to more rapid and successful sequence-to-function pipelines in the future.

Importantly, the application of CFE for secondary metabolite synthesis or BGC expression is sparse compared to well-established cell-based heterologous expression, and the limitations of its application are still being understood. For example, a proteomic investigation of *E. coli*-based CFE reactions showed that the synthesis of a large (394 kDa) modified PKS, DEBS1-TE, is mostly truncated[130]. Due to the relatively static nature of the metabolism within the lysate compared to the genome-regulated metabolism of cells, resource limitation can interfere with complete protein synthesis in these extracts, especially when producing synthases, given that their active domains can compete with the TX-TL machinery for cofactors, salts and energy substrates[119]. Future optimizations of current *E. coli*-based systems or testing of non-E. *coli* lysates for large-scale BGC expression will be necessary to better evaluate this platform for accelerated secondary metabolite discovery. Further, while fungi-based CFE systems from *Saccharomyces*[131], *Aspergillus*[132] and *Neurospora*[132] have been reported, these all underperform in terms of protein expression compared to their cell-based counterparts. The expression of fungal BGCs via CFE will not be feasible without significantly improving current roadmaps for developing fungal lysate-based platforms.

Secondary metabolite detection

Determining chemical structures from expressed BGCs remains a bottleneck when linking genes to their products but advancements in the field of analytical chemistry are affording the high-throughput exploration of a broader natural product space.

Bioassays commonly used to detect natural products in the 1900s primarily screened for antibiotic activity[3]. The advent of less targeted, MS-based technologies has supported the detection of new bioactivities from secondary metabolomes by providing outputs that are rich in information[28]. Molecular networking, which groups related compounds based on similarities in their MS/MS fragmentation spectra, enables the annotation of known metabolites in the sample and therefore the identification of unknown groups of metabolites. However, it is important to note that alterations in fragmentation patterns may also arise from chemical variants generated by a single BGC[133]. The IsoAnalyst workflow distinguishes these families of compounds to identify novel chemical variants produced by characterized and uncharacterized BGCs[133].

This is done by fermenting a microorganism in minimal media with different stable isotope labelled (SIL) precursors that are likely to be incorporated into a secondary metabolite (for example, [methyl-^{13}C] methionine to label methylated structures). The IsoAnalyst algorithm then determines SIL precursor incorporation patterns and matches these to theoretical precursor incorporation rates predicted using antiSMASH and MIBiG information. Using this approach, lobosamide D, a previously uncharacterized lobosamide, was detected in the metabolome of a *Micromonospora* sp. isolate. This approach thus simplifies the connection of new metabolites to their source BGCs and will benefit from greater availability of BGC annotation tools and SIL precursors for different BGC classes. A significant disadvantage of this method is that it would be difficult to implement for secondary metabolites that are only synthesized under complex media conditions.

In addition to MS-based techniques, microcrystal electron diffraction (microED) continues to cement itself as an analytical tool in natural product discovery[31,134]. It was introduced in 2013 as a cryo-electron microscopy technique that elucidates structures of crystallized proteins and small molecules from electron diffraction patterns. MicroED advantageously needs significantly less sample (that is, nanogram quantities) compared to X-ray crystallography and, in some cases, sample crystallization is unnecessary. Further, sample purities can be lower compared to those prepared for NMR spectroscopy. However, it is important to note that crystallization remains a challenging and unpredictable process. MicroED has been leveraged to elucidate structures of natural products, like macrocyclic drugs[135] and fischerin[136], that had previously eluded characterization.

MicroED applications are currently sparse but, if the technique can be implemented reproducibly, it would be an extremely powerful tool for accelerating natural product discovery, especially as its throughput increases. A recently introduced technique called ArrayED fractionates crude extracts via HPLC that are then deposited into picolitre-sized droplets on TEM grids, supporting the simultaneous analysis of 96 fractions by microED. Compared to analysing fractions individually on a single TEM grid, the multiplexed approach reportedly reduces the workflow time from 200+ h to 4–8 h. The application of ArrayED led to the structural characterization of 14 natural products, including 4 novel structures[137].

Conclusion

The number of natural products isolated from bacteria and fungi per year has progressively increased from ~680 in 2000 to ~2,700 in 2019 (ref. 14). We expect these numbers to increase with continued innovations in computational and empirical BGC research. Optimizing and integrating these technologies into discovery workflows would ideally enable the prioritization of truly novel BGCs and high-throughput synthesis and detection of their cognate metabolites in the future. However, to realize this, certain limitations of these techniques need to be overcome.

Predicting truly novel BGC types and architectures in silico remains a challenging task. Non-canonical BGCs can evade prediction by rule-based algorithms as was the case for fungal ICS BGCs. Pattern-based approaches show promise but these are still trained on genomic data sets that are biased towards well-studied taxa. Advancements in metagenomics may remedy the latter issue, especially with long-read sequencing allowing the recovery of full-length BGCs from uncultivated lineages. Novel complex BGC architectures may be captured across different genomes in these data, facilitating detection by pattern-based tools.

BGC expression remains the most significant bottleneck in modern discovery workflows. Activating BGCs in native producers can be conducted with high throughput by screening elicitors or creating mutant libraries but these approaches are limited to microbes that are culturable and genetically tractable. Furthermore, these screenings require rapid detection methods of expression or production, which often require BGC-type or product-specific readouts and limit the throughput. Heterologous expression is an excellent strategy for expressing sequences that are mined in silico but host selection is constraining. Technologies like CRAGE now enable the domestication of non-model chassis for paralleled BGC expression but this will be challenging to automate. CFE may complement in vivo approaches but it has had relatively limited applications in secondary metabolite discovery.

Advances in untargeted detection techniques allow better resolution and quicker structural characterization. IsoAnalyst can connect a chemical signature to its source BGC but it can only be applied to cells expressing BGCs in minimal media. Alternatively, SIL-based approaches can be coupled with CFE, where metabolic complexity is less of an issue. MicroED has shown promise but its utility for secondary metabolite discovery requires further demonstration given the infancy of this strategy.

Nascent approaches are attempting to tie BGC prediction to activity, especially for health and agricultural applications. By doing so, more targeted approaches for the development of secondary metabolites for commercial applications can be realized especially as the wealth of BGC data increases from genomic and post-genomic investigations.

Addressing these and building accelerated discovery pipelines will help uncover more natural products that can be assayed for unique or enhanced bioactivities. These molecules will have a crucial role in the transition to a bioeconomy by further advancing medicine and agriculture, and through newer applications across other economic sectors.

Published online: 17 January 2025

References

1. Global Bioeconomy Summit, Communiqué of the Global Bioeconomy Summit 2015. *Making Bioeconomy Work for Sustainable Development* (Global Biosecurity Summit, 2015).
2. Baltz, R. H. Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J. Ind. Microbiol. Biotechnol.* **46**, 281–299 (2019).
3. Katz, L. & Baltz, R. H. Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol.* **43**, 155–176 (2016).
   The article provides a comprehensive review of the history of microbial natural product research from the 1940s to the first half of the 2010s.
4. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
5. Sparks, T. C., Sparks, J. M. & Duke, S. O. Natural product-based crop protection compounds horizontal line origins and future prospects. *J. Agric. Food Chem.* **71**, 2259–2269 (2023).
6. Tewari, D., Atanasov, A. G., Semwal, P. & Wang, D. Natural products and their applications. *Curr. Res. Biotechnol.* **3**, 82–83 (2021).
7. George, K. W., Alonso-Gutierrez, J., Keasling, J. D. & Lee, T. S. In: *Biotechnology of Isoprenoids* (eds Schrader, J. & Bohlmann, J.) 355–389 (Springer International Publishing, 2015).
8. Hill, P. et al. Clean manufacturing powered by biology: how Amyris has deployed technology and aims to do it better. *J. Ind. Microbiol. Biotechnol.* **47**, 965–975 (2020).
9. Yuzawa, S., Keasling, J. D. & Katz, L. Bio-based production of fuels and industrial chemicals by repurposing antibiotic-producing type I modular polyketide synthases: opportunities and challenges. *J. Antibiot.* **70**, 378–385 (2017).
10. Wang, Z. et al. A microbial platform for recyclable plastics with customizable properties. Preprint at *ResSq.* https://doi.org/10.21203/rs.3.rs-3171588/v1 (2023).
11. Bérdy, J. Thoughts and facts about antibiotics: where we are now and where we are heading. *J. Antibiot.* **65**, 385–395 (2012).

12. Schneider, Y. K. Bacterial natural product drug discovery for new antibiotics: strategies for tackling the problem of antibiotic resistance by efficient bioprospecting. *Antibiotics* **10**, 842 (2021).
13. Mouncey, N. J., Otani, H., Udwary, D. & Yoshikuni, Y. New voyages to explore the natural product galaxy. *J. Ind. Microbiol. Biotechnol.* **46**, 273–279 (2019).
14. Jeffrey, A. et al. The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.* **50**, D1317–D1323 (2022).
15. Blair, P. M. et al. Exploration of the biosynthetic potential of the *Populus* microbiome. *mSystems* **3**, e00045-18 (2018).
16. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
The paper describes the analysis of a large, taxonomically and biogeographically diverse metagenomic data set derived from Earth's microbiomes and reports on 104,000 predicted BGCs from these data, of which approximately 84% were not identified in NCBI's sequence databases.
17. Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440–444 (2018).
18. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
19. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
20. Krause, D. J. et al. Functional and evolutionary characterization of a secondary metabolite gene cluster in budding yeasts. *Proc. Natl Acad. Sci. USA* **115**, 11030–11035 (2018).
21. Demain, A. L. Importance of microbial natural products and the need to revitalize their discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 185–201 (2014).
22. Seshadri, R. et al. Expanding the genomic encyclopedia of Actinobacteria with 824 isolate reference genomes. *Cell Genomics* **2**, 100213 (2022).
23. Kadjo, A. E. & Eustáquio, A. S. Bacterial natural product discovery by heterologous expression. *J. Ind. Microbiol. Biotechnol.* **50**, kuad044 (2023).
24. van Bergeijk, D. A., Terlouw, B. R., Medema, M. H. & van Wezel, G. P. Ecology and genomics of actinobacteria: new concepts for natural product discovery. *Nat. Rev. Microbiol.* **18**, 546–558 (2020).
25. Scherlach, K. & Hertweck, C. Mining and unearthing hidden biosynthetic potential. *Nat. Commun.* **12**, 3864 (2021).
26. Caesar, L. K., Montaser, R., Keller, N. P. & Kelleher, N. L. Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities. *Nat. Product. Rep.* **38**, 2041–2065 (2021).
27. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes — a review. *Nat. Product. Rep.* **33**, 988–1005 (2016).
This paper discusses how genome-mining approaches have evolved from classical approaches to evolutionary genome-mining strategies along with the advancement of genome-sequencing technologies.
28. Bouslimani, A., Sanchez, L. M., Garg, N. & Dorrestein, P. C. Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Product. Rep.* **31**, 718 (2014).
29. Bogart, J. W. et al. Cell-free exploration of the natural product chemical space. *ChemBioChem* **22**, 84–91 (2021).
This article reviews the use of cell-free systems for BGC expression in recent years and offers perspectives on how cell-free expression may be leveraged for high-throughput BGC characterization and engineering.
30. Katz, M., Hover, B. M. & Brady, S. F. Culture-independent discovery of natural products from soil metagenomes. *J. Ind. Microbiol. Biotechnol.* **43**, 129–141 (2016).
31. Danelius, E., Halaby, S., van der Donk, W. A. & Gonen, T. MicroED in natural product and small molecule research. *Nat. Prod. Rep.* **38**, 423–431 (2021).
32. Sayers, E. W. et al. GenBank 2024 update. *Nucleic Acids Res.* **52**, D134–D137 (2024).
33. Chen, I. M. A. et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* **51**, D723–D732 (2023).
34. Yuan, D. et al. The European Nucleotide Archive in 2023. *Nucleic Acids Res.* **52**, D92–D97 (2024).
35. Petras, D. et al. GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. *Nat. Methods* **19**, 134–136 (2022).
36. Navarro-Munoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
This introduces BiG-SCAPE and CORASON and uses them within a single framework, wherein BiG-SCAPE is used to generate sequence similarity networks of BGCs and CORASON elucidates phylogenetic analyses within and across these networks.
37. Blin, K., Shaw, S., Medema, M. H. & Weber, T. The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more. *Nucleic Acids Res.* **52**, D586–D589 (2024).
38. Zdouc, M. M. et al. MIBiG 4.0: advancing biosynthetic cluster curation through global collaboration. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkae1115 (2024).
39. Udwary, D. W. et al. The secondary metabolism collaboratory: a database and web discussion portal for secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* gkae1060 https://doi.org/10.1093/nar/gkae1060 (2024).
40. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–D497 (2021).
41. Louwen, J. J. R. & Van Der Hooft, J. J. J. Comprehensive large-scale integrative analysis of omics data to accelerate specialized metabolite discovery. *mSystems* **6**, e0072621 (2021).
42. Ferrinho, S., Connaris, H., Mouncey, N. J. & Goss, R. J. M. Compendium of metabolomic and genomic datasets for cyanobacteria: mined the gap. *Water Res.* **256**, 121492 (2024).
43. Schorn, M. A. et al. A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).
44. Louwen, J. J. R., Medema, M. H. & Van Der Hooft, J. J. J. Enhanced correlation-based linking of biosynthetic gene clusters to their metabolic products through chemical class matching. *Microbiome* **11**, 13 (2023).
45. Blin, K. et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
This article highlights the newest features and updates to antiSMASH, which include new detection rules to identify new cluster types, improvements to NRPS, PKS and RiPP predictions, the ability to find transcription factor binding sites, and features that support a more user-friendly interface.
46. Klau, L. J. et al. The Natural Product Domain Seeker version 2 (NaPDoS2) webtool relates ketosynthase phylogeny to biosynthetic function. *J. Biol. Chem.* **298**, 102480 (2022).
47. Kautsar, S. A., van der Hooft, J. J. J., Ridder, D. & Medema, M. H. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience* **10**, giaa154 (2021).
48. Chevrette, M. G. et al. In *Engineering Natural Product Biosynthesis* (ed. E. Skellam) 129–155 (Springer, 2022).
49. Mungan, M. D. et al. ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.* **48**, W546–W552 (2020).
50. Alanjary, M. et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.* **45**, W42–W48 (2017).
51. Yılmaz, T. M., Mungan, M. D., Berasategui, A. & Ziemert, N. FunARTS, the Fungal bioActive compound Resistant Target Seeker, an exploration engine for target-directed genome mining in fungi. *Nucleic Acids Res.* **51**, W191–W197 (2023).
52. Nickles, G. R., Oestereicher, B., Keller, N. P. & Drott, M. T. Mining for a new class of fungal natural products: the evolution, diversity, and distribution of isocyanide synthase biosynthetic gene clusters. *Nucleic Acids Res.* **51**, 7220-7235 (2023).
53. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
54. Sélem-Mojica, N., Aguilar, C., Gutiérrez-García, K., Martínez-Guerrero, C. E. & Barona-Gómez, F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genom.* **5**, e000260 (2019).
55. Mullowney, M. W. et al. Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
56. Sanchez, S. et al. Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.23.540769 (2023).
57. Almeida, H., Palys, S., Tsang, A. & Diallo, A. B. TOUCAN: a framework for fungal biosynthetic gene cluster discovery. *NAR Genom. Bioinform.* **2**, lqaa098 (2020).
58. Carroll, L. M. et al. Accurate de novo identification of biosynthetic gene clusters with GECCO. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.03.442509 (2021).
59. Lee, Y.-Y. et al. HypoRiPPAtlas as an Atlas of hypothetical natural products for mass spectrometry database search. *Nat. Commun.* **14**, 4219 (2023).
60. Donia, M. S., Ruffner, D. E., Cao, S. & Schmidt, E. W. Accessing the hidden majority of marine natural products through metagenomics. *Chembiochem* **12**, 1230–1236 (2011).
61. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
62. Klapper, M. et al. Natural products from reconstructed bacterial genomes of the Middle and Upper Paleolithic. *Science* **380**, 619–624 (2023).
This study reports on the use of untargeted metagenomics to assemble ancient MAGs from the microbiomes of humans and Neanderthals, leading to the discovery of ancient secondary metabolites termed paleofurans.
63. Meziti, A. et al. The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* **87**, e02593-20 (2021).
64. Van Goethem, M. W. et al. Long-read metagenomics of soil communities reveals phylum-specific secondary metabolite dynamics. *Commun. Biol.* **4**, 1302 (2021).
65. Waschulin, V. et al. Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing. *ISME J.* **16**, 101–111 (2022).
66. Borsetto, C. et al. Microbial community drivers of PK/NRP gene diversity in selected global soils. *Microbiome* **7**, 78 (2019).
67. Gehrig, J. L. et al. Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microb. Genom.* **8**, 000794 (2022).
68. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
69. Libis, V. et al. Multiplexed mobilization and expression of biosynthetic gene clusters. *Nat. Commun.* **13**, 5256 (2022).
70. Libis, V. et al. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).
71. Dzunkova, M. et al. Synthase-selected sorting approach identifies a beta-lactone synthase in a nudibranch symbiotic bacterium. *Microbiome* **11**, 130 (2023).
The authors improve upon their previously developed synthase-selective approach to isolate biosynthetic microbes from microbiomes, resulting in the identification of a new species that encodes a 27.9-kb beta-lactone gene cluster.

72. Kim, W. E. et al. Synthase-selective exploration of a tunicate microbiome by activity-guided single-cell genomics. *ACS Chem. Biol.* **16**, 813–819 (2021).
73. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc. Natl Acad. Sci. USA* **111**, 9259–9264 (2014).
74. Craney, A., Ahmed, S. & Nodwell, J. Towards a new science of secondary metabolism. *J. Antibiot.* **66**, 387–400 (2013).
75. Mao, D., Okada, B. K., Wu, Y., Xu, F. & Seyedsayamdost, M. R. Recent advances in activating silent biosynthetic gene clusters in bacteria. *Curr. Opin. Microbiol.* **45**, 156–163 (2018).
This review article covers principles behind HiTES and random mutagenesis tactics for high-throughput BGC activation and detection in native host organisms.
76. Pan, R., Bai, X., Chen, J., Zhang, H. & Wang, H. Exploring structural diversity of microbe secondary metabolites using OSMAC strategy: a literature review. *Front. Microbiol.* **10**, 294 (2019).
77. Okada, B. K. & Seyedsayamdost, M. R. Antibiotic dialogues: induction of silent biosynthetic gene clusters by exogenous small molecules. *FEMS Microbiol. Rev.* **41**, 19–33 (2017).
78. Hoshino, S., Onaka, H. & Abe, I. Activation of silent biosynthetic pathways and discovery of novel secondary metabolites in actinomycetes by co-culture with mycolic acid-containing bacteria. *J. Ind. Microbiol. Biotechnol.* **46**, 363–374 (2019).
79. Ameruoso, A., Villegas Kcam, M. C., Cohen, K. P. & Chappell, J. Activating natural product synthesis using CRISPR interference and activation systems in *Streptomyces*. *Nucleic Acids Res.* **50**, 7751–7760 (2022).
80. Zhang, M. M. et al. CRISPR–Cas9 strategy for activation of silent *Streptomyces* biosynthetic gene clusters. *Nat. Chem. Biol.* **13**, 607–609 (2017).
81. Deng, L. et al. Dissection of 3D chromosome organization in *Streptomyces coelicolor* A3(2) leads to biosynthetic gene cluster overexpression. *Proc. Natl Acad. Sci. USA* **120**, e2222045120 (2023).
82. Wang, X. et al. Elucidation of genes enhancing natural product biosynthesis through co-evolution analysis. *Nat. Metab.* **6**, 933–946 (2024).
83. Bode, H. B., Bethe, B., Höfs, R. & Zeeck, A. Big effects from small changes: possible ways to explore nature's chemical diversity. *ChemBioChem* **3**, 619 (2002).
84. Craney, A., Ozimok, C., Pimentel-Elardo, S. M., Capretta, A. & Nodwell, J. R. Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* **19**, 1020–1027 (2012).
85. Zhuang, L. & Zhang, H. Utilizing cross-species co-cultures for discovery of novel natural products. *Curr. Opin. Biotechnol.* **69**, 252–262 (2021).
86. Moon, K., Xu, F., Zhang, C. & Seyedsayamdost, M. R. Bioactivity-HiTES unveils cryptic antibiotics encoded in actinomycete bacteria. *ACS Chem. Biol.* **14**, 767–774 (2019).
87. Zhang, C. & Seyedsayamdost, M. R. Discovery of a cryptic depsipeptide from *Streptomyces ghanaensis* via MALDI-MS-guided high-throughput elicitor screening. *Angew. Chem. Int. Ed. Engl.* **59**, 23005–23009 (2020).
88. Han, E. J., Lee, S. R., Townsend, C. A. & Seyedsayamdost, M. R. Targeted discovery of cryptic enediyne natural products via FRET-coupled high-throughput elicitor screening. *ACS Chem. Biol.* **18**, 1854–1862 (2023).
89. Mao, D., Yoshimura, A., Wang, R. & Seyedsayamdost, M. R. Reporter-guided transposon mutant selection for activation of silent gene clusters in *Burkholderia thailandensis*. *Chembiochem* **21**, 1826–1831 (2020).
90. Yoshimura, A. et al. Unlocking cryptic metabolites with mass spectrometry-guided transposon mutant selection. *ACS Chem. Biol.* **15**, 2766–2774 (2020).
91. Akhgari, A. et al. Single cell mutant selection for metabolic engineering of actinomycetes. *Metab. Eng.* **73**, 124–133 (2022).
92. Zhao, Y. et al. CRISPR/dCas9-mediated multiplex gene repression in *Streptomyces*. *Biotechnol. J.* **13**, 1800121 (2018).
93. McCarty, N. S., Graham, A. E., Studená, L. & Ledesma-Amaro, R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nat. Commun.* **11**, 1281 (2020).
94. Heng, E., Tan, L. L., Zhang, M. M. & Wong, F. T. CRISPR-Cas strategies for natural product discovery and engineering in actinomycetes. *Process. Biochem.* **102**, 261–268 (2021).
95. Shen, Q. et al. Utilization of CRISPR-Cas genome editing technology in filamentous fungi: function and advancement potentiality. *Front. Microbiol.* **15**, 1375120 (2024).
96. Yuan, Y. et al. Efficient exploration of terpenoid biosynthetic gene clusters in filamentous fungi. *Nat. Catal.* **5**, 277–287 (2022).
97. Yuan, Y., Huang, C., Singh, N., Xun, G. & Zhao, H. Automated, self-resistance gene-guided, and high-throughput genome mining of bioactive natural products from *Streptomyces*. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.26.564101 (2023).
98. Ayikpoe, R. S. et al. A scalable platform to discover antimicrobials of ribosomal origin. *Nat. Commun.* **13**, 6135 (2022).
99. Tellechea-Luzardo, J., Otero-Muras, I., Goñi-Moreno, A. & Carbonell, P. Fast biofoundries: coping with the challenges of biomanufacturing. *Trends Biotechnol.* **40**, 831–842 (2022).
100. Wang, G. et al. CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nat. Microbiol.* **4**, 2498–2510 (2019).
The article describes high successful BGC expression rates by domesticating and utilizing a diverse set of non-model organisms as chassis strains.
101. Caesar, L. K., Kelleher, N. L. & Keller, N. P. In the fungus where it happens: history and future propelling Aspergillus nidulans as the archetype of natural products research. *Fungal Genet. Biol.* **144**, 103477 (2020).
102. Meng, X. et al. Developing fungal heterologous expression platforms to explore and improve the production of natural products from fungal biodiversity. *Biotechnol. Adv.* **54**, 107866 (2022).
103. Lawson, C. E. et al. Common principles and best practices for engineering microbiomes. *Nat. Rev. Microbiol.* **17**, 725–741 (2019).
104. Hug, J. J., Krug, D. & Muller, R. Bacteria as genetically programmable producers of bioactive natural products. *Nat. Rev. Chem.* **4**, 172–193 (2020).
105. Schmidt, M. et al. Maximizing heterologous expression of engineered type I polyketide synthases: investigating codon optimization strategies. *ACS Synth. Biol.* **12**, 3366–3380 (2023).
106. Dudley, Q. M., Karim, A. S. & Jewett, M. C. Cell-free metabolic engineering: biomanufacturing beyond the cell. *Biotechnol. J.* **10**, 69–82 (2015).
107. Karim, A. S. et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat. Chem. Biol.* **16**, 912–919 (2020).
108. Karim, A. S. & Jewett, M. C. in *Methods in Enzymology* Vol. 608 (ed. Scrutton, N.) 31–57 (Academic Press Inc., 2018).
109. Garenne, D. et al. Cell-free gene expression. *Nat. Rev. Methods Prim.* **1**, 149 (2021).
110. Konczal, J. & Gray, C. H. Streamlining workflow and automation to accelerate laboratory scale protein production. *Protein Expr. Purif.* **133**, 160–169 (2017).
111. Gregorio, N. E., Levine, M. Z. & Oza, J. P. A user's guide to cell-free protein synthesis. *Methods Protoc.* **2**, 24–24 (2019).
112. Ji, X., Liu, W. Q. & Li, J. Recent advances in applying cell-free systems for high-value and complex natural product biosynthesis. *Curr. Opin. Microbiol.* **67**, 102142 (2022).
113. Silverman, A. D., Karim, A. S. & Jewett, M. C. Cell-free gene expression: an expanded repertoire of applications. *Nat. Rev. Genet.* **21**, 151–170 (2020).
114. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
115. Dopp, B. J. L., Tamiev, D. D. & Reuel, N. F. Cell-free supplement mixtures: elucidating the history and biochemical utility of additives used to support in vitro protein synthesis in *E. coli* extract. *Biotechnol. Adv.* **37**, 246–258 (2019).
116. Si, Y., Kretsch, A. M., Daigh, L. M., Burk, M. J. & Mitchell, D. A. Cell-free biosynthesis to evaluate lasso peptide formation and enzyme-substrate tolerance. *J. Am. Chem. Soc.* **143**, 5917–5927 (2021).
This study investigates cell-free expression as a platform for synthesizing lasso peptides and demonstrates its utility for generating thousands of sequence-diverse lassopeptides from precursor peptide variants.
117. Siebels, I. et al. Cell-free synthesis of natural compounds from genomic DNA of biosynthetic gene clusters. *ACS Synth. Biol.* **9**, 2418–2426 (2020).
118. Zhuang, L. et al. Total in vitro biosynthesis of the nonribosomal macrolactone peptide valinomycin. *Metab. Eng.* **60**, 37–44 (2020).
119. Dinglasan, J. L. N., Sword, T. T., Barker, J. W., Doktycz, M. J. & Bailey, C. B. Investigating and optimizing the lysate-based expression of nonribosomal peptide synthetases using a reporter system. *ACS Synth. Biol.* **12**, 1447–1460 (2023).
120. Dudley, Q. M., Karim, A. S., Nash, C. J. & Jewett, M. C. In vitro prototyping of limonene biosynthesis using cell-free protein synthesis. *Metab. Eng.* **61**, 251–260 (2020).
121. Sword, T. T. et al. Profiling expression strategies for a type iii polyketide synthase in a lysate-based, cell-free system. *Sci. Rep.* **14**, 12983 (2024).
122. Moore, S. J., Lai, H. E., Li, J. & Freemont, P. S. Streptomyces cell-free systems for natural product discovery and engineering. *Nat. Prod. Rep.* **40**, 228–236 (2023).
123. Sword, T. T., Abbas, G. S. K. & Bailey, C. B. Cell-free protein synthesis for nonribosomal peptide synthetic biology. *Front. Nat. Products* **3**, https://doi.org/10.3389/fntpr.2024.1353362 (2024).
124. Garenne, D., Thompson, S., Brisson, A., Khakimzhan, A. & Noireaux, V. The all-E. coliTXTL toolbox 3.0: new capabilities of a cell-free synthetic biology platform. *Synth. Biol.* **6**, ysab017 (2021).
125. Xu, H. et al. Regulatory part engineering for high-yield protein synthesis in an all-*Streptomyces*-based cell-free expression system. *ACS Synth. Biol.* **11**, 570–578 (2022).
126. Moore, S. J. et al. A *Streptomyces venezuelae* cell-free toolkit for synthetic biology. *ACS Synth. Biol.* **10**, 402–411 (2021).
127. Li, J., Wang, H., Kwon, Y.-C. & Jewett, M. C. Establishing a high yielding *Streptomyces*-based cell-free protein synthesis system. *Biotechnol. Bioeng.* **114**, 1343–1353 (2017).
128. Yim, S. S. Multiplex transcriptional characterizations across diverse bacterial species using cell-free systems. *Mol. Syst. Biol.* **15**, e8875 (2019).
129. Moore, S. J. et al. Rapid acquisition and model-based analysis of cell-free transcription–translation reactions from nonmodel bacteria. *Proc. Natl Acad. Sci. USA* **115**, 4340–4349 (2018).
130. Hurst, G. B. et al. Proteomics-based tools for evaluation of cell-free protein synthesis. *Anal. Chem.* **89**, 11443–11451 (2017).
131. Gan, R. & Jewett, M. C. A combined cell-free transcription-translation system from *Saccharomyces cerevisiae* for rapid and robust protein synthesis. *Biotechnol. J.* **9**, 641–651 (2014).
132. Schramm, M. et al. Cell-free protein synthesis with fungal lysates for the rapid production of unspecific peroxygenases. *Antioxidants* **11**, 284 (2022).
133. McCaughey, C. S., van Santen, J. A., van der Hooft, J. J. J., Medema, M. H. & Linington, R. G. An isotopic labeling approach linking natural products with biosynthetic gene clusters. *Nat. Chem. Biol.* **18**, 295–304 (2022).
The authors describe a new metabolomics method for quickly connecting compounds in MS data sets to their source BGCs by leveraging stable isotope-labelled precursors.
134. Clark, L. J., Bu, G., Nannenga, B. L. & Gonen, T. MicroED for the study of protein-ligand interactions and the potential for drug discovery. *Nat. Rev. Chem.* **5**, 853–858 (2021).

135. Danelius, E., Bu, G., Wieske, L. H. E. & Gonen, T. MicroED as a powerful tool for structure determination of macrocyclic drug compounds directly from their powder formulations. *ACS Chem. Biol.* **18**, 2582–2589 (2023).
136. Kim, L. J. et al. Prospecting for natural products by genome mining and microcrystal electron diffraction. *Nat. Chem. Biol.* **17**, 872–877 (2021).
137. Delgadillo, D. A. et al. High-throughput identification of crystalline natural products from crude extracts enabled by microarray technology and microED. *ACS Cent. Sci.* **10**, 176–183 (2024).
This reports on a high-throughput workflow leveraging microED for characterizing natural products from crude cell extracts.
138. Haslinger, K., Peschke, M., Brieke, C., Maximowitsch, E. & Cryle, M. J. X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. *Nature* **521**, 105–109 (2015).
139. Gallo, M. E. *The Bioeconomy: A Primer* (Congressional Research Service, 2022).
140. Tan, E. C. D. & Lamers, P. Circular bioeconomy concepts — a perspective. *Front. Sustain.* **2**, 53–53 (2021).
141. Nowruzi, B., Sarvari, G. & Blanco, S. The cosmetic application of cyanobacterial secondary metabolites. *Algal Res.* **49**, 101959 (2020).
142. Gupta, P. L., Rajput, M., Oza, T., Trivedi, U. & Sanghvi, G. Eminence of microbial products in cosmetic industry. *Nat. Prod. Bioprospect.* **9**, 267–278 (2019).
143. Fouillaud, M. & Dufossé, L. Microbial secondary metabolism and biotechnology. *Microorganisms* **10**, 123 (2022).
144. Süssmuth, R. D. & Mainz, A. Nonribosomal peptide synthesis — principles and prospects. *Angew. Chem. Int. Ed.* **56**, 3770–3821 (2017).
145. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and diversity of assembly-line polyketide synthases. *Chem. Rev.* **119**, 12524–12547 (2019).
146. Wang, J., Deng, Z., Liang, J. & Wang, Z. Structural enzymology of iterative type I polyketide synthases: various routes to catalytic programming. *Nat. Product. Rep.* **40**, 1498–1520 (2023).
147. Palmer, C. M. & Alper, H. S. Expanding the chemical palette of industrial microbes: metabolic engineering for type III PKS-derived polyketides. *Biotechnol. J.* **14**, 1700463 (2019).
148. Wang, J. Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. *Microb. Cell Factories* **19**, 110 (2020).
149. Danby, P. M. & Withers, S. G. Advances in enzymatic glycoside synthesis. *ACS Chem. Biol.* **11**, 1784–1794 (2016).
150. Avalos, M. et al. Biosynthesis, evolution and ecology of microbial terpenoids. *Nat. Product. Rep.* **39**, 249–272 (2022).
151. Montalbán-López, M. et al. New developments in RiPP discovery, enzymology and engineering. *Nat. Product. Rep.* **38**, 130–239 (2021).
152. Keatinge-Clay, A. T. Polyketide synthase modules redefined. *Angew. Chem. Int. Ed.* **56**, 4658–4660 (2017).
153. Felnagle, E. A. et al. Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Mol. Pharmaceutics* **5**, 191–191 (2008).
154. Martínez-Núñez, M. A. & López, V. E. L. Nonribosomal peptides synthetases and their applications in industry. *Sustain. Chem. Process.* **4**, 13 (2016).
155. Okano, A., Isley, N. A. & Boger, D. L. Total syntheses of vancomycin-related glycopeptide antibiotics and key analogues. *Chem. Rev.* **117**, 11952–11993 (2017).
156. Bozhüyük, K. A. J. et al. Evolution-inspired engineering of nonribosomal peptide synthetases. *Science* **383**, eadg4320 (2024).
157. Mabesoone, M. F. J. et al. Evolution-guided engineering of *trans*-acyltransferase polyketide synthases. *Science* **383**, 1312–1317 (2024).
158. Van Staden, A. D. P., Van Zyl, W. F., Trindade, M., Dicks, L. M. T. & Smith, C. Therapeutic application of lantibiotics and other lanthipeptides: old and new findings. *Appl. Environ. Microbiol.* **87**, e0018621 (2021).
159. Negash, A. W. & Tsehai, B. A. Current applications of bacteriocin. *Int. J. Microbiol.* **2020**, 4374891 (2020).
160. Cheng, C. & Hua, Z. C. Lasso peptides: production and potential medical application. *Front. Bioeng. Biotechnol.* **8**, 571165 (2020).
161. Hussain, H. et al. Fungal glycosides: structure and biological function. *Trends Food Sci. Technol.* **110**, 611–651 (2021).
162. Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R–37R (2006).
163. Krause, K. M., Serio, A. W., Kane, T. R. & Connolly, L. E. Aminoglycosides: an overview. *Cold Spring Harb. Perspect. Med.* **6**, a027029 (2016).
164. Sundin, G. W. & Wang, N. Antibiotic resistance in plant-pathogenic bacteria. *Annu. Rev. Phytopathol.* **56**, 161–180 (2018).
165. Moutinho, L. F., Moura, F. R., Silvestre, R. C. & Romão-Dumaresq, A. S. Microbial biosurfactants: a broad analysis of properties, applications, biosynthesis, and techno-economical assessment of rhamnolipid production. *Biotechnol. Prog.* **37**, e3093 (2021).
166. Rudolf, J. D., Alsup, T. A., Xu, B. & Li, Z. Bacterial terpenome. *Nat. Prod. Rep.* **38**, 905–980 (2021).

Acknowledgements

The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Author contributions

J.L.N.D. wrote the initial draft and made most of the revisions. The remaining authors contributed equally to all other aspects of the article.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Microbiology* thanks Nancy Keller, who co-reviewed with Grant Nickles, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

Antibiotics and Secondary Metabolite Analysis Shell Database: https://antismash-db.secondarymetabolites.org
Secondary Metabolites Collaboratory: https://smc.jgi.doe.gov