Mining genomes to illuminate the specialized chemistry of life

Nature Reviews Genetics

Medema, Marnix H.; Rond, Tristan; Moore, Bradley S.

https://doi.org/10.1038/s41576-021-00363-7

Mining genomes to illuminate the specialized chemistry of life

Marnix H. Medema[1], *Tristan de Rond*[2] *and Bradley S. Moore*[2,3]✉

Abstract | All organisms produce specialized organic molecules, ranging from small volatile chemicals to large gene-encoded peptides, that have evolved to provide them with diverse cellular and ecological functions. As natural products, they are broadly applied in medicine, agriculture and nutrition. The rapid accumulation of genomic information has revealed that the metabolic capacity of virtually all organisms is vastly underappreciated. Pioneered mainly in bacteria and fungi, genome mining technologies are accelerating metabolite discovery. Recent efforts are now being expanded to all life forms, including protists, plants and animals, and new integrative omics technologies are enabling the increasingly effective mining of this molecular diversity.

Natural products
Organic compounds originating from living organisms or natural sources, often prized for their medicinal properties or other biological activities of utility to humanity. The term is typically used to refer to products of secondary metabolism, but also includes primary metabolites.

Specialized metabolites
Natural compounds of limited clade-specific or niche-specific distribution, known or presumed to have a specialized role in ecology or physiology.

[1]*Bioinformatics Group, Wageningen University, Wageningen, The Netherlands.*

[2]*Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA.*

[3]*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA.*

✉*e-mail: bsmoore@ucsd.edu*

https://doi.org/10.1038/s41576-021-00363-7

Genetically encoded organic molecules are the common chemical language that unites all life, from single cells to communities of organisms. Whereas many biochemical compounds are shared among large swaths of the tree of life, some molecules are biosynthesized only by a select subset of organisms and/or are specific to certain ecological niches. The terms natural products, specialized metabolites and secondary metabolites are often used interchangeably for these molecules (although see REFS[1–3] for in-depth discussions of the definitions of these terms and their differences). They range in size, shape and complexity, from small terpenes and phosphonates to large and heavily post-translationally modified gene-encoded peptides; other prominent classes include polyketides, non-ribosomally synthesized peptides, alkaloids, glycosides and phenylpropanoids.

Specialized metabolites have evolved to impart diverse cellular intraspecies and interspecies functions that perform key roles in physiology and in simple to complex ecosystems. These metabolites provide organisms — from single-cell microorganisms to multicellular plants and animals — with some of their most distinguishing chemical features of colour, smell, taste or toxicity. In other words, the blend of specialized metabolites endowed to an organism makes it unique. Production of a siderophore or antioxidant can enable an organism to thrive in an environment hostile to others; hormones allow different tissues of a complex organism to communicate while carrying out specialized tasks; and toxins, venoms, scents and pigments shape the role an organism plays in its ecosystem. Besides their natural functions, these molecules are widely applied in human society, as medicines, crop protection agents, food additives, colourants and fragrances. Molecules such as penicillin, oestradiol and caffeine are just a small selection of nature's chemical bounty that has had profound societal impact (FIG. 1a).

Most specialized metabolites have been identified through experimental discovery approaches that take advantage of a chemical or biological feature of the expressed molecule to guide its isolation. The rapid accumulation of genomic and transcriptomic information in recent years has revealed that the metabolic capacity of virtually all organisms is vastly underappreciated, with millions of additional molecules awaiting discovery[4–6]. Genome mining seeks to harness gene-based big data methods to expedite the concomitant discovery of specialized metabolites and their biosynthetic genes. With increasing technological improvements in genome sequencing, early mining experiments of relatively simple microbial genomes have been followed in recent years by much more complex genomes and metagenomes of plants, animals and other eukaryotic organisms that differ in the organization of their biosynthesis genes (FIG. 1). Additionally, to truly arrive at a deeper understanding of life's chemistry, genome mining approaches are being developed that provide insight into the functions that these molecules perform in physiology and ecology. Here, we address the why, what, where and how of genome mining and discuss key challenges in deciphering what nature is 'verbalizing'.

Why we mine and what to mine

Natural chemicals have been identified dating back to 1803, with the isolation of morphine from opium poppy[7]. Historically, specialized metabolites have been isolated and characterized from biological samples collected from the environment or from laboratory-grown organisms, whereby organic extracts of tissues or cells are chemically and biologically analysed. Whereas

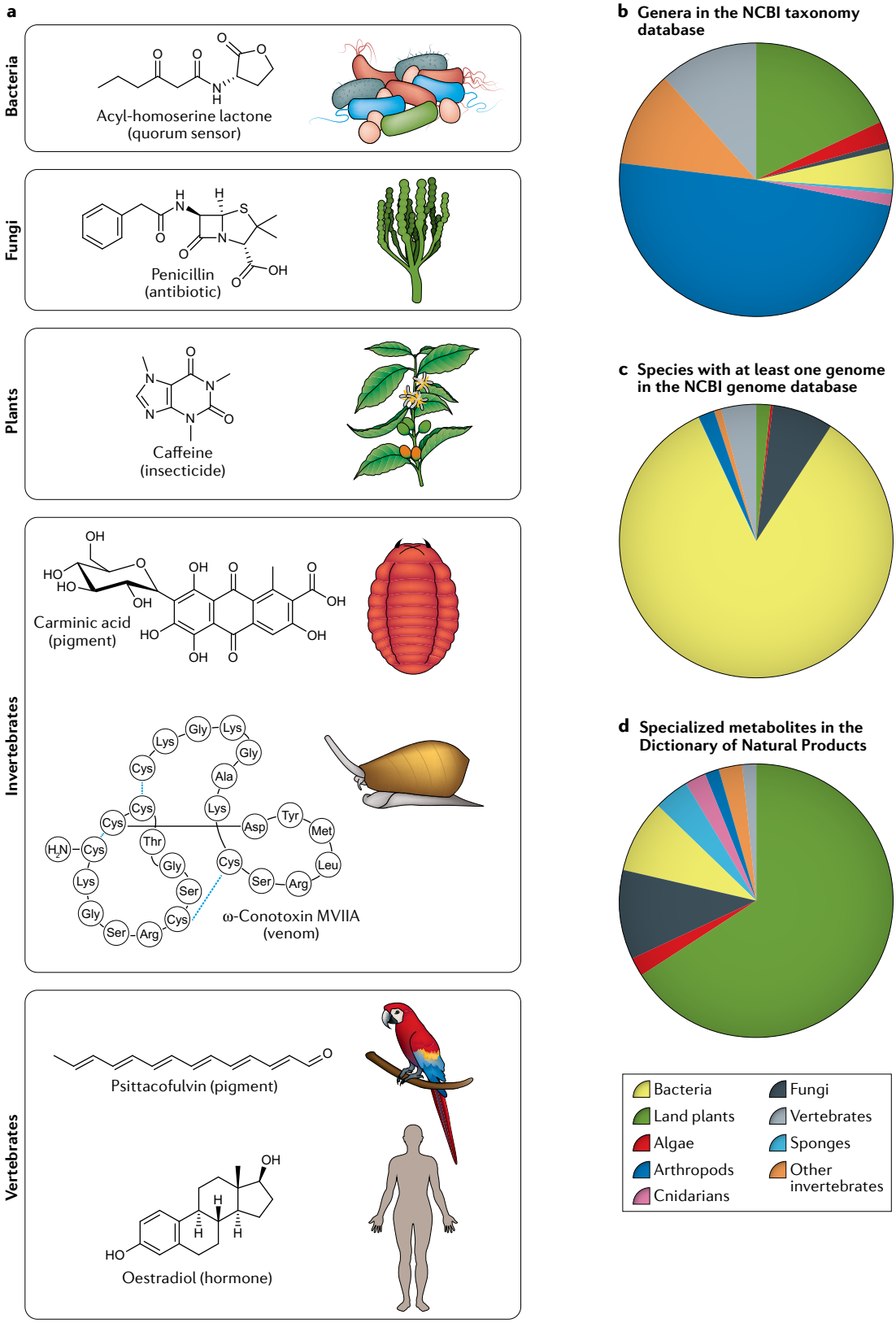
Secondary metabolites
Metabolites that are not strictly required for growth and development, as opposed to primary metabolites, but are often important for survival of an organism in its environment. In the classical meaning, secondary metabolites do not include proteins or large gene-derived peptides that are not post-translationally modified by enzymes.

Siderophore
A metabolite that binds (chelates) iron ions from the environment and is re-imported back into a cell for iron acquisition. Other 'metallophores' bind trace metals such as zinc and copper.

Biosynthetic genes
Genes encoding enzymes that catalyse transformations in a biosynthetic pathway.

analytical chemistry tools continue to improve in sensitivity and speed[8], the past decades have shown a clear deceleration of the discovery of novel structure chemotypes versus the rediscovery of well-known molecular families with subtle chemical modifications[9].

Genome mining has the potential to increase the discovery rate and facilitate the characterization of novel molecules and biosynthetic pathways. For example, the model organism *Streptomyces coelicolor* A3(2) (REF.[10]) was heavily studied for about half a century



Ribosomally synthesized and post-translationally modified peptide (RiPP). A peptide biosynthesized through the action of tailoring enzymes on a ribosomally translated precursor peptide.

Heterologous expression Expression of one or more genes originating from one organism in another organism; often used to obtain higher production titres or to independently verify their chemical structure or biological function.

before the publication of its genome sequence, with around a dozen (types of) specialized metabolites discovered. Genome mining has since led to the discovery of seven additional metabolites from diverse classes: the non-ribosomal peptides coelibactin[10] and coelichelin[11], the sesquiterpene (+)-epi-isozizaene[12], 2-alkyl-4-hydroxymethylfuran-3-carboxylic acids[13], the ribosomally synthesized and post-translationally modified peptide (RiPP) SCO-2138 (REF.[14]), the polyketide coelimycin P1 (REF.[15]) and a new set of partially characterized arsenopolyketides[16].

Mining genomes has key advantages over the use of analytical chemistry techniques alone. First, mining can access specialized metabolites that may not be produced under the growth conditions studied. Second, the approach inherently connects any discovered molecules to their biosynthetic genes, allowing for heterologous expression and bulk production. This factor is particularly significant because many medicinally valuable molecules are isolated from dwindling natural resources or organisms that are difficult to cultivate, and genome sequencing typically requires much less biomass than structural analytics.

The motivations for genome mining have largely tracked those of the natural products community at large: historically, this has primarily been the exploration of life's biochemical prowess, the understanding of physiology and the pursuit of therapeutics. In the past century, the first specialized metabolites were linked to their biosynthetic genes usually from cloned DNA fragments that could be used to experimentally complement random mutations carried in those genes[17–19]. By the 2000s, genome sequencing had started to mature, and the biosynthetic logic of some major classes of medicinal natural products, including polyketides, non-ribosomal peptides and terpenoids, had been deciphered to some extent. Newly sequenced genomes often harboured homologues of genes encoding biosynthetic machinery for these classes of compounds, but had not been associated with a metabolic product. Heterologous expression of these 'orphan' biosynthetic genes resulted in the discovery of several novel natural products, including triterpenes from the *Arabidopsis* genome[20] and the hybrid peptide–polyketide aspyridones from the genome of the model filamentous fungus *Aspergillus flavus*[21]. Since these proofs of concept, countless new members of established major compound classes have been discovered through genome mining.

Genome mining is also contributing to the ongoing fundamental search for chemical and biosynthetic novelty in nature. Several specialized metabolites harbouring chemical moieties unprecedented for their class, such as furanone[22,23] and benzo[a]tetraphene[24] polyketides, and aminovinylcysteine-based RiPPs[25], were discovered through genome mining. Even among known specialized metabolites, there are numerous structures for which the biosynthetic machinery was recently elucidated through genome mining, such as the piperazate[26], thiotetronate[27], oxazolone[28,29], isoxazole[30], alkyne[31,32], *N*-nitroso[33], and diazo[34] moieties, polybrominated phenolics from marine bacteria[35], plant-like isoquinoline alkaloids in diverse fungi[36] and vinca alkaloids from medicinal plants[37]. As new biosynthetic reactions and structural classes are discovered, our ability to reliably predict orphan genes coding for the biosynthesis of novel structural features will continue to improve. Still, there are many classes of specialized metabolites for which the genetic basis is still completely or mostly a mystery, such as the polycyclic ethers found in dinoflagellates[38] or the ladderanes produced by anammox bacteria[39,40]. Undoubtedly, numerous chemical features not represented among known specialized metabolites remain to be discovered through genome mining.

Our understanding of ribosomally synthesized peptides has particularly benefited from the rise of genome mining, as their structures can often be fairly easily predicted from genomic data. Among these peptides, RiPPs[41] are particularly noteworthy for their broad distribution across all three domains of life and our growing knowledge of their diversity of peptidic modifications[42]. Even before the genomic era, RiPPs already played important societal roles[43]; for example, the bacterial lanthipeptide nisin is a widely used food preservative[44], and the marine cone snail-derived ω-conotoxin MVIIA (the synthetic analogue of which is known as ziconotide) has been developed into a drug (Prialt) for the amelioration of chronic pain[45]. New structural families of RiPPs continue to be discovered, such as the spliceotides[46] and epipeptides[47] from bacteria, dikaritins[48,49] from fungi and the lyciumins[50] from plants. Ribosomally derived specialized metabolites are not always RiPPs and range remarkably in size, from small molecules, such as the pyrroloquinoline alkaloid ammosamide[51,52], to small proteins, such as the three-finger toxins from spitting cobras[53], venom proteins from spiders[54,55] and antimicrobial proteins in humans[56]. Similar discovery trends can be seen in the other major biosynthetic lineages, where the mining of genomes has resulted in the growth of chemical and biochemical knowledge.

In recent years, new motivations for genome mining have emerged from two new areas of research: microbiomes and synthetic biology. In microbiome research, the mining of specialized metabolites and the genes encoding their biosynthetic machinery provides a window into the mechanisms responsible for key phenotypes mediated by the microbiome, such as pathogen suppression[57,58] or host immunomodulation[59]. Moreover, it potentially enables the design of synthetic microbial consortia that can be used as live therapies or biopharmaceuticals[60–62], based on genome-based prediction of the chemical capabilities of individual strains. In synthetic biology, pathways being mined from

◀ Fig. 1 | **Life's chemical diversity. a** | Bacteria, fungi, plants and animals produce a wide range of specialized metabolites that help them thrive in their respective environments. **b–d** | There is a large disconnect between the number of taxonomic genera in the biosphere (based on the National Center for Biotechnology Information (NCBI) taxonomy database) (part **b**), the number of genomes available for these species (based on the number of species represented in the NCBI genome database) (part **c**) and the number of specialized metabolites isolated (based on the number of molecules ascribed to these classes of organisms in the Dictionary of Natural Products) (part **d**). There is likely great potential for discovering new metabolites from animals and protists, and identifying new biosynthetic pathways from plants, animals and protists. Algae includes green, red and brown algae, diatoms and dinoflagellates. Heterotrophic protists and archaea were not included due to the low number of specialized metabolites isolated from these organisms.

genomes, mainly as a source of enzymological diversity, are starting to be used as 'parts' for metabolic engineering of novel molecules with desirable properties[63]. In the future, this approach may enable combinatorialization of enzymes[64] or even computer-aided design[65] to create 'new to nature' molecules.

Where to mine

Bacteria

Genome mining is predicated on the availability of omics data; thus, growth in the field has relied on improvements in sequencing technologies. To date, the majority of genome mining has been conducted on bacterial genomes, which, given their comparatively small size and low repeat content, dominate publicly available genomic databases (FIG. 1c). Further simplifying the mining process within bacteria is their propensity to physically cluster genes in operons and biosynthetic gene clusters (BGCs; BOX 1) for cooperative biosynthesis of specialized metabolites. This has allowed researchers to readily formulate hypotheses regarding the biosynthesis of molecules of interest, even in cases where substrates and enzymes have no precedent. Genes that cluster with another gene known or are suspected to be involved in the biosynthesis of a specialized metabolite are often promising candidates for the identification of other genes involved in their biosynthetic pathway.

Soil microorganisms, and in particular the actinomycetes, were already a popular source of specialized metabolites in the pre-genomic era and were thus obvious targets for early sequencing and mining efforts. The first genomes of *Streptomyces*, *Salinispora* and *Saccharopolyspora* species pre 2008 revealed that the actinomycetes were metabolically richer than originally thought, with many species dedicating over 10% of their genomic space to the production of dozens of specialized metabolites[10,66–68]. This trend has now been observed in many other environmental bacteria, especially those with large genomes in excess of 10 Mb. The filamentous marine cyanobacterium *Moorea producens*, for instance, devotes roughly one-fifth of its genome in this manner[69]. Due to decreasing costs of bacterial genome sequencing, recent efforts have ballooned in scale to mining 10,000–100,000+ genomes at a time for novel molecules[70,71].

The specialized chemistry of uncultivated bacteria that dominate the microbiota of animals, plants and other host organisms has also been examined through genome mining, highlighting the importance of microbial metabolites in modulating health and disease within their hosts. Be it human gut bacteria[72], plant rhizosphere microbial communities[73] or marine sponge microbiota[74], metagenomic mining of the microbial 'dark matter' of life is quickly revealing that microorganisms are indispensable for host chemical fitness. Even without a living host, such as in soils, seawater and the air, environmental DNA has further revealed the exquisite metabolic capacity of the Earth's microbiota through the diversity of associated biosynthetic genes[75,76]. Although attempts to exploit environmental DNA as a genetic resource for natural product discovery were already initiated two decades ago[77], better computational infrastructure such as reference databases[78] and profiling software[79], as well as massively increased sequencing volumes, have now turned this into a promising technology. Indeed, innovative efforts have now led to the engineered production of drug leads directly from the mining of soil environmental DNA samples[80,81].

Fungi

Filamentous fungi, such as *Aspergillus nidulans* and *Penicillium chrysogenum*, have long been known to cluster their genes for the biosynthesis of, for example, the carcinogenic toxin aflatoxin or the antibiotic penicillin[18,82]. Although fungi and bacteria share many of the same hallmark secondary metabolic pathways, fungi also feature distinctive enzymatic reactions such as the reducing iterative polyketide synthases, which produce the cholesterol-reducing agent lovastatin[83]. With their larger genomes, fungi also encode many more biosynthetic pathways than the most prolific bacteria. The genome of the fungus *Aspergillus tanneri* NIH1004 has 95 BGCs[84], setting it up as the strain with the largest specialized metabolic capacity amongst fungi discovered thus far.

Plants

Long thought to be a uniquely microbial phenomenon, it is now becoming increasingly clear that BGCs are found throughout the tree of life (BOX 1). Land plants dwarf all other organisms for known specialized metabolites (FIG. 1d). Plant molecules, such as the anticancer drug taxol, the plant hormone gibberellin or caffeine (which functions as an insecticide yet is best known as a constituent of coffee and other caffeinated drinks), dominate the literature on specialized metabolism, with more than 145,000 described molecules. Early experiments connecting plant chemistry and genes relied upon sequencing expressed sequence tag libraries and transcriptomes[85–87]. In recent years, plant genomics has gained traction, revealing the genomic context of specialized metabolism. The triterpene thalianol in *Arabidopsis* was one of the first plant compounds whose encoding genes were found to be chromosomally clustered[88], albeit in a manner much unlike bacterial BGCs. Genes within plant BGCs are typically not organized in tight operons but, rather, with large intergenic regions that can span up to a few hundred kilobases in stretches; as such, genes are typically transcribed separately[89]. Recent plant omic studies have connected genes to the production of iconic opioid, cannabinoid and vinca alkaloid plant molecules, leading to renewable fermentation opportunities for their robust production[37,90,91].

Algae

The success of the plant community in connecting genes to specialized chemistry has led to the investigation of other eukaryotic systems that each harbour distinctive chemistry. For instance, some of the most notorious environmental toxins are produced by diverse marine microalgae. Recently, a BGC was established in the diatom *Pseudo-nitschia multiseries* for the global production of the amnesic shellfish toxin domoic acid[92]. By contrast, dinoflagellates produce arguably the largest and most complex chemicals known from nature,

Biosynthetic gene clusters (BGCs). Sets of genes that are physically co-located on a chromosome and together encode the production, regulation and transport of one or more specific metabolites.

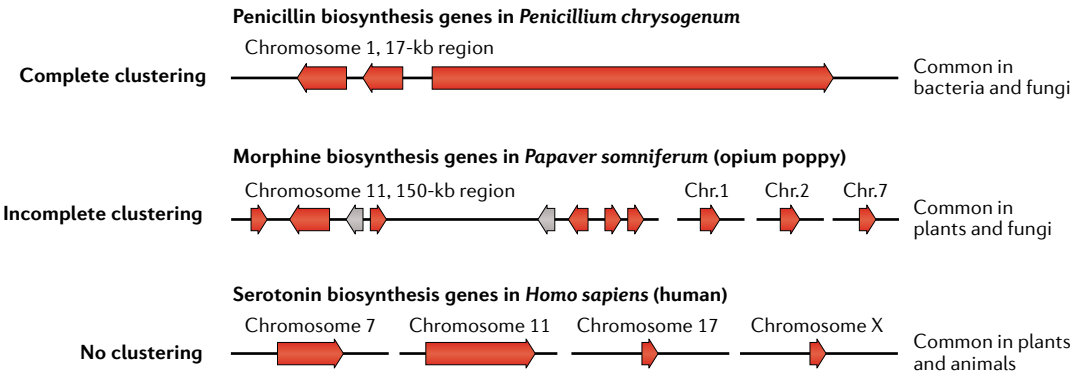
Polyketide synthases
Enzymes involved in the biosynthesis of polyketide metabolites; some form modular assembly lines of multidomain proteins, whereas others act as stand-alone enzymes.

Box 1 | **Gene clustering in specialized metabolism**

In most organisms, genes involved in specialized metabolic pathways are encoded contiguously on the chromosome in so-called biosynthetic gene clusters (BGCs). The extent to which biosynthetic genes are clustered differs between different taxonomic groups, and specifically between the plant, fungal and bacterial kingdoms, which show increasing degrees of gene clustering (see the figure). As an illustration, in the model actinomycete bacterium *Streptomyces coelicolor*, 22 BGCs have been experimentally characterized and linked to products (including 2 single enzyme-coding genes), and for none of the corresponding pathways is there evidence of encoding in multiple genomic loci. On the other hand, out of the 23 BGCs experimentally characterized in the model fungus *Aspergillus nidulans*, at least 3 pathways have been shown to be split over multiple loci: those for the biosynthesis of austinol/dehydroaustinol[221], emericellin[222] and nidulanin A[223]. In the model plant *Arabidopsis thaliana*, only four pathways have been experimentally shown to be encoded by BGCs: those for the biosynthesis of thalianol, marneral, arabidiol and tirucalladienol. Although several other pathways seem to show partial clustering[164,224], the pathways for the biosynthesis of glucosinolates, flavonoids, strigolactones, arabidopyrones, camalexin and 4-hydroxyindole-3-carbonyl nitrile seem to be (almost) devoid of clustering. Still, even in plants, BGCs are an attractive target for pathway discovery, as they provide 'low-hanging fruits' that can be straightforwardly identified in genome sequences[5]. In protists, several examples of BGCs have been reported[92,225], whereas not much is known about gene clustering in animals. Yet a recent global synteny network analysis shows that the gene order in mammals is clearly non-random and may have large functional repercussions[226].

There are several hypotheses for why the genes of specialized metabolic pathways are clustered on the genome. The four main ones are the following:

1. **Coordinated gene expression**. In bacteria, given that transcription and translation occur in the same cellular location, the biophysics of transcriptional regulation favours co-regulation of operons located near the gene encoding a pathway-specific regulator[102]. In fungi and plants, there is evidence that clustered genes are co-regulated through epigenetic modification of chromosomal regions[227,228].
2. **The selfish operon hypothesis**. Given that horizontal gene transfer of BGCs, but also their deletion, occurs frequently in bacteria and fungi, the 'survival' of BGCs in the biosphere may depend on their ability to spread to other strains and species; clustering may increase the chances of genes being jointly transferred[229]. This can be supplemented by a 'persistence hypothesis', stating that clustered genes are less likely to be interrupted by a segmental duplication and, therefore, are more likely to survive as a unit[230].
3. **Avoiding toxic intermediates**. According to this hypothesis, clustering of genes is an adaptation against the accumulation of toxic pathway intermediates. Clustering promotes co-inheritance of the entire pathway, so that (sub) lethal genotypes carrying only part of the pathway are avoided[231].
4. **Co-adaptation through co-inheritance**. Many clusters in plants and fungi have formed in dynamic chromosomal regions as part of evolutionary arms races with competing species[232]. Especially in sexual organisms, rapid adaptation of pathways may only be possible when co-adapted alleles of the underlying genes are not constantly separated by recombination events. This has recently been proposed to drive repeated and independent evolution of gene clusters encoding phenylpropanoid degradation pathways in fungi[233].



polyether toxins such as brevetoxin and maitotoxin[93]. Although biosynthesis genes have yet to be identified for these dinoflagellate compounds — perhaps owing to their large genome sizes that regularly exceed that of the human genome and assemble into liquid crystalline chromosomes[94] — the recent assembly of the ~6.4-Gb draft genome of the toxic *Amphidinium gibbosum* revealed an abundance of suspected polyketide synthase and non-ribosomal peptide synthetase (NRPS) genes[95]. The recent reconstruction of hundreds of genomes of plankton species from metagenomic data provides an additional rich set of unexplored genomic data to mine for specialized metabolic diversity[96].

Non-ribosomal peptide synthetase
(NRPS). An enzyme involved in the polymerization of amino acids or other organic acids into peptide metabolites without involvement of the ribosome.

Metazoa

The anthropocentric bias of biomedical research has led scientists to qualify compounds isolated from many animals as distinct from bacterial, fungal and plant specialized metabolites. However, a more impartial perspective should recognize that many animal-specialized molecules are chemically related to and perform functions similar to their non-animal counterparts. Although, in some cases, animal-derived specialized metabolites are biosynthesized by specialized microbiome members[97,98], the biosynthetic capacities of the animal itself should not be underestimated. Humans, for instance, produce numerous steroid hormones such as oestradiol, cortisol

and aldosterone, the thyroid hormone triiodothyronine and even the antiviral ribonucleotide 3′-deoxy-3′,4′-didehydro-CTP (REF.[99]). The recently discovered routes from bird[100–102] and mollusc[102,103] genomes to produce complex polyketides, as well as a novel sesquiterpene biosynthetic pathway from flea beetles[104], exemplify the chemical ingenuity of animals in making important molecules key to their fitness and survival. Ecologically, venoms such as the conotoxin RiPPs produced by cone snails play major roles in predation and defence[105]. In some cases, animal pathways have been acquired through horizontal gene transfer from bacteria, as is evident for the β-lactam antibiotic biosynthetic genes found in the genome of the springtail *Folsomia candida*[106,107], but in most of the documented cases mentioned above, their biosynthesis seems to have evolved independently[100–102,104], indicating that considerable quantities of distinct chemistry may be discovered through mining animal genomes.

Now that eukaryotic genome sequencing is becoming more routine, we anticipate that genome mining projects will soon extend to all organisms (BOX 2). Although there have been sporadic reports of specialized biosynthetic genes and gene clusters being functionally elucidated from, for example, the nematode *Caenorhabditis elegans*[108], the fruit fly *Drosophila melanogaster*[109] and the seaweed *Digenea simplex*[110], large swaths of organisms such as arthropods, cnidarians and other invertebrates are understudied for their biosynthetic capacities yet well known for their specialized chemistry.

How to mine

Identifying candidate biosynthetic genes

A range of computational approaches has been developed to automatically identify the sets of genes that encode specialized metabolic enzymes across genome sequences (TABLE 1). Many of these approaches have originally been developed for bacteria (and sometimes for fungi and plants), but the principles employed have the potential to be extended to other life forms. Below, we review these methodologies and the taxa they support, and what would be required to extend them into new taxonomic spaces.

The physical clustering of enzyme-coding genes in BGCs greatly facilitates the identification of biosynthetic pathways. Although BGCs are highly variable in terms of gene content and are often strain-specific due to their rapid evolution and frequent horizontal gene transfer[111], they often do possess common properties in the form of enzyme families that are responsible for the catalysis of biochemical reactions central to the biosynthesis of entire specialized metabolite compound classes. This feature has made it possible to largely automate the identification of BGCs in genomes. Widely used software tools such as antiSMASH[112] and PRISM[113] employ profile hidden Markov models (pHMMs[114]) of protein domains to identify gene combinations encoding enzyme families that are signatures for specific pathway types. Although both of these tools generally provide very similar results, development of antiSMASH has focused more on functional and comparative analyses, whereas PRISM has specialized in combinatorial predictions of chemical structures that can be used for automated matching with mass-spectral data. The use of pHMMs is very reliable for identifying BGCs encoding many well-established types of biosynthetic machinery such as polyketide synthases, NRPSs and known classes of RiPPs, but risks overlooking less studied and wholly novel classes of BGCs. Probabilistic BGC prediction methods such as ClusterFinder[115] (which is also integrated into antiSMASH) and DeepBGC[116], or comparative genomics approaches that identify metabolism-associated non-syntenic blocks of genes between genomes, are more likely to detect non-standard BGCs, but have higher false-positive rates. In addition, for RiPPs, specialized tools have emerged for the identification of BGCs encoding the production of distant members of known classes or members of altogether novel classes. Some of these, such as BAGEL[117], use pHMM-based detection techniques similar to those seen in antiSMASH and PRISM. Others make use of either bait-based approaches (using specific query enzyme-encoding genes to identify loci that contain homologues of them)[118,119] or machine learning approaches to identify potential precursor peptide-encoding genes[120–122], the hits of which can be prioritized using metabolomics-based matching[121] or comparative genomics to identify operons that are taxon-specific and thus deemed to encode a specialized metabolic function[122]. For publicly available genomes, BGCs identified using antiSMASH can be interactively browsed in online databases such as IMG-ABC[123] and antiSMASH-DB[124].

Recently, it has become clear that, in plants, specialized metabolic pathways are sometimes encoded by BGCs[89] (BOX 1), and specific algorithms have been devised for their detection[125,126]. However, there are also many examples of pathways in plants that are encoded by sets of genes distributed across multiple chromosomes instead of being located in a single gene cluster. When extending genome mining approaches to unexplored parts of the tree of life, it remains to be seen to what extent genes in these taxa will be clustered. Some recent evidence suggests that the phenomenon of gene clustering also occurs in protists; for example, the domoic acid biosynthetic pathway in the diatom *P. multiseries* was shown to be encoded by a four-gene cluster[92]. However, gene cluster detection algorithms originally devised for bacteria may require considerable optimization to make them effective for studying protist or animal genomes. Efforts to adapt antiSMASH for detecting BGCs in plants in the form of a new tool called plantiSMASH[126] showed that, for this to be effective, new libraries of pHMMs focused on plant enzymology needed to be constructed, and the algorithm had to be adjusted to account for the considerably larger (and more variable) intergenic regions found in plant genomes[113].

Prioritizing candidate biosynthetic genes

Computational predictions often lead to an overabundance of candidate biosynthetic genes that could be investigated, necessitating prioritization. Given that the chemical structures of hundreds of thousands of specialized metabolites have been elucidated, a considerable number of these will be responsible for the biosynthesis

Horizontal gene transfer
Acquisition of genetic material by one organism, originating from another. This is often facilitated by plasmids, viruses or mobile elements.

Profile hidden Markov models
(pHMMs). Computational models, trained on a multiple-sequence alignment of a protein family, used to assess whether proteins are part of (or related to) a family.

Box 2 | **How much is there to mine?**

Both the large diversity of molecules found in nature and the even larger diversity of biosynthetic genes found in genome sequences make it clear that the chemical and enzymological space available to genome mining is vast. Yet it is difficult to gauge just how vast it is.

Focusing on possibly the most chemically diverse clade of microorganisms, the actinomycetes, Doroghazi et al. posited that sequencing a well-chosen set of only ~15,000 actinomycete genomes would reveal virtually all naturally occurring gene cluster families in this class of bacteria[132]. This statement was based on extrapolating a rarefaction curve of gene cluster families, in which sampling had been corrected for phylogeny within the limits of the data set used. However, a subsequent study on the diversity of non-ribosomal peptide synthetase (NRPS) gene clusters, which included a larger number of genomes and used chemical structure predictions to support family assignments, indicated no signs of saturation around 15,000 genomes[165], suggesting that genome-encoded biosynthetic diversity may be larger than previously estimated, at least for this class of pathways. Similarly, Schorn et al. revisited estimates of biosynthetic diversity based on a study of rare marine actinomycete genomes, which suggested that rarefaction analyses may be too conservative to estimate diversity across the biosphere, as they inherently do not take into account genomes from unsampled ecological niches and taxonomic subgroups[234].

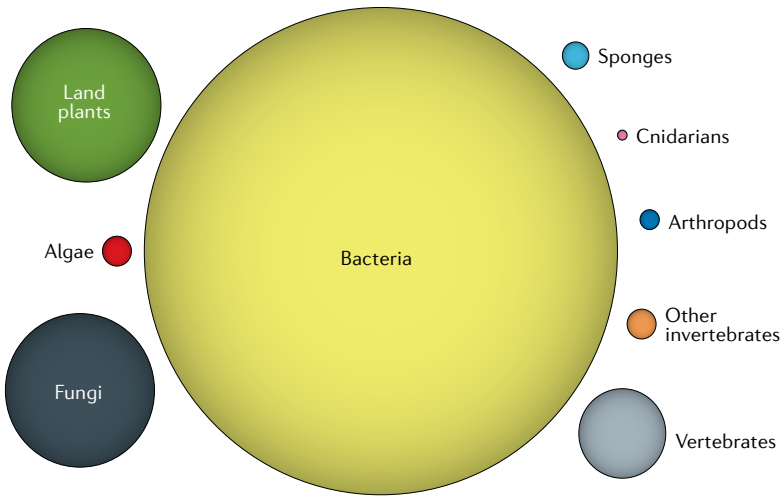
A rough estimate of the total number of specialized metabolites employed by life can be made based on known biodiversity (FIG. 1b) and metabolic diversity (see the figure, panel **a** and FIG. 1d), by multiplying the number of specialized metabolites reported for a relatively well-studied genus — sourced from Natural Product Atlas[235] for *Pseudomonas* and *Aspergillus*, and from the Dictionary of Natural Products for all other genera, and assumed to be representative for the genus — by the number of genera for the type of organism: this results in a total in the order of tens of millions. These could be overestimates because genera may share specialized metabolites, or underestimates because more specialized metabolites may be discovered for the chosen genus or more genera may still be discovered. Contrasting this to the number of elucidated specialized metabolites (in the order of half a million) suggests we have merely scratched the surface of the biochemical diversity present in the biosphere. Studies on bacteria and fungi support this notion, showing that, regardless of the rapid accumulation of known specialized metabolites and associated risks of rediscovery, the absolute numbers of structurally novel specialized metabolites discovered over the past 20 years has remained remarkably steady, at around 150–250 per year[9,236].

Although estimates (see the figure, panel **a**) suggest there is great potential for the discovery of specialized metabolites throughout the whole tree of life, our understanding of their biosynthesis is heavily skewed towards bacteria (see the figure, panel **b**, in which areas indicate relative numbers of specialized metabolites whose biosynthetic genes have been identified, based on estimates made by the authors). This is likely due to the greater availability of genomic data for bacteria (FIG. 1c). Even for the relatively well-studied specialized metabolism of bacteria, our understanding of culturable species dwarfs uncultured bacteria. This could be remedied by bringing more bacterial species into culture through new sampling or cultivation strategies[237,238], or by expanding metagenomic studies of diverse environments globally, and in turn mining the resulting genomics data. Nevertheless, to spur our understanding of specialized metabolism throughout the whole tree of life, it will similarly be imperative to collect thorough genomic data for a wide variety of eukaryotic organisms. NCBI, National Center for Biotechnology Information.

a

	Genera in NCBI	Example of well-studied genus	Specialized metabolites isolated from this genus	Extrapolated number of specialized metabolites
Land plants	15,573	*Brassica*	349	~5,400,000
Algae	2,206	*Laurencia*	902	~2,000,000
Fungi	716	*Aspergillus*	2,034	~1,500,000
Bacteria	3,980	*Pseudomonas*	318	~1,300,000
Sponges	499	*Dysidea*	515	~250,000
Cnidarians	1,152	*Sinularia*	807	~900,000
Arthropods	41,922	*Drosophila*	104	~4,400,000
Other invertebrates	9,706	*Caenorhabditis*	52	~500,000
Vertebrates	9,838	*Dendrobates*	142	~1,400,000
			Total:	~18,000,000

b **Specialized metabolites ascribed to genes**



Gene cluster families
Families comprising a set of similar biosynthetic gene clusters across strains or species, the members of which are responsible for the production of the same or very similar metabolites.

of known molecules or their closely related variants. Hence, it is beneficial to assess whether biosynthetic genes and their likely products are novel or whether they have been discovered and characterized previously.

The simplest way of prioritizing is based on sequence information: if a BGC of interest is highly similar in sequence to a gene cluster that has been experimentally linked to a known specialized metabolite, it likely codes for the production of the same molecule. In 2015, a community effort established the Minimum Information about a Biosynthetic Gene cluster (MIBiG)[78], a data standard and online repository for depositing annotations and metadata on BGCs for which a product has been identified. The antiSMASH pipeline for BGC identification automatically compares each identified BGC against this repository of ~2,000 BGCs of known function. When studying large numbers of genomes at once, BGC sequence similarity networks[115] can be utilized to identify gene cluster families that cluster together with MIBiG reference clusters. The BiG-SCAPE software

framework automates the process of generating these networks and facilitates their interactive exploration, which makes it possible to quickly explore the biosynthetic diversity within hundreds or even thousands of prokaryotic genomes at once[127]. It remains to be seen to what extent this technology is universally applicable across the tree of life. For example, it was recently shown that plant triterpene biosynthetic loci may be

Table 1 | **Genome mining technologies that combine genome sequence with other data**

Hypothesis-generating method	Input data	Helps generate hypotheses about	Select examples	Select software pipelines
Classic genome mining	Genome or transcriptome sequences	Gene–chemistry relationships	Coelichelin[11,202] Triterpenes[20] (Reviewed in[4,203,204])	antiSMASH[112] PRISM[113] DeepBGC[116] CO-ED[29] Focused on RiPPs: RODEO[118] DecRiPPter[122]
Function-directed genome mining	Genome or transcriptome sequences	Gene–chemistry–function relationships	Aspterric acid[156] Siderophores[153] Thiolactomycin[27]	Target-directed genome mining: ARTS[157], FRIGG[205] Genome neighbourhood analysis: EFI-GNT[206]
Co-expression analysis	Gene expression levels Genome or transcriptome sequences (optional)	Gene–chemistry relationships	4-Hydroxyindole-3-carbonyl nitrile (REF.[146]) Steroidal glycoalkaloids[144] (Reviewed in[5,147,207,208])	WGCNA[209] CoExpNetViz[210] plantiSMASH[126] mr2mods (REF.[148])
Gene expression–metabolite correlation analysis	Gene expression levels Analytical chemistry features (e.g. peaks)	Gene–chemistry relationships	Falcarindiol[143] Proteomining[211]	NA
Pattern-based genome mining(metabologenomics)	Genome or transcriptome sequences Gene expression levels Analytical chemistry features (e.g. peaks)	Gene–chemistry relationships	Tambromycin[131] Tyrobetaines[134] Several *Salinispora* BGCs[130] Zealexin biosynthesis through association mapping in maize[212] Chemically guided functional profiling[213] (Reviewed in[133,214])	NPLinker[135] EFI-CGFP[206] MAGI[215]
Gene–phenotype correlation analysis	Genome or transcriptome sequences Gene expression levels Bioactivities or phenotypes	Gene–function relationships	Bitterness in cucumber[161] Flavobacterial NRPS–PKS in disease suppression[58]	BiG-MAP[216]
Activity-guided genome mining	Genome or transcriptome sequences Bioactivities or phenotypes Analytical chemistry features (e.g. peaks)	Gene–chemistry–function relationships	Iomaiviticin[217]	NA
Gene–metabolite substructure matching	Genome or transcriptome sequence Fragmentation spectra	Gene–chemistry relationships	Peptidogenomics Glycogenomics (Reviewed in[133,218])	RiPPquest[138] Pep2Path (REF.[137]) MetaMiner[139]
Retro-biosynthetic matching	Genome or transcriptome sequences In silico library of structural formulae	Gene–chemistry relationships	Several PKs, NRPs	GRAPE/GARLIC[141] rBAN[219]
Spectral dereplication	Fragmentation spectra In silico library of structural formulae	Novel chemistry	Reviewed in[220]	VarQuest[193] MS2LDA (REF.[169]) CSI:FingerID[168]

Each combination has its own strengths and may allow generating hypotheses focused on finding an unknown biosynthetic pathway for an important known molecule, discovering new metabolites with desired biological activities or identifying potential links between metabolites and the genes and gene clusters that likely encode their biosynthesis. ARTS, Antibiotic Resistant Target Seeker; BGC, biosynthetic gene cluster; NA, not available; NRP, non-ribosomal peptide: NRPS, non-ribosomal peptide synthetase; PK, polyketide; PKS, polyketide synthase; RiPP, ribosomally synthesized and post-translationally modified peptide.

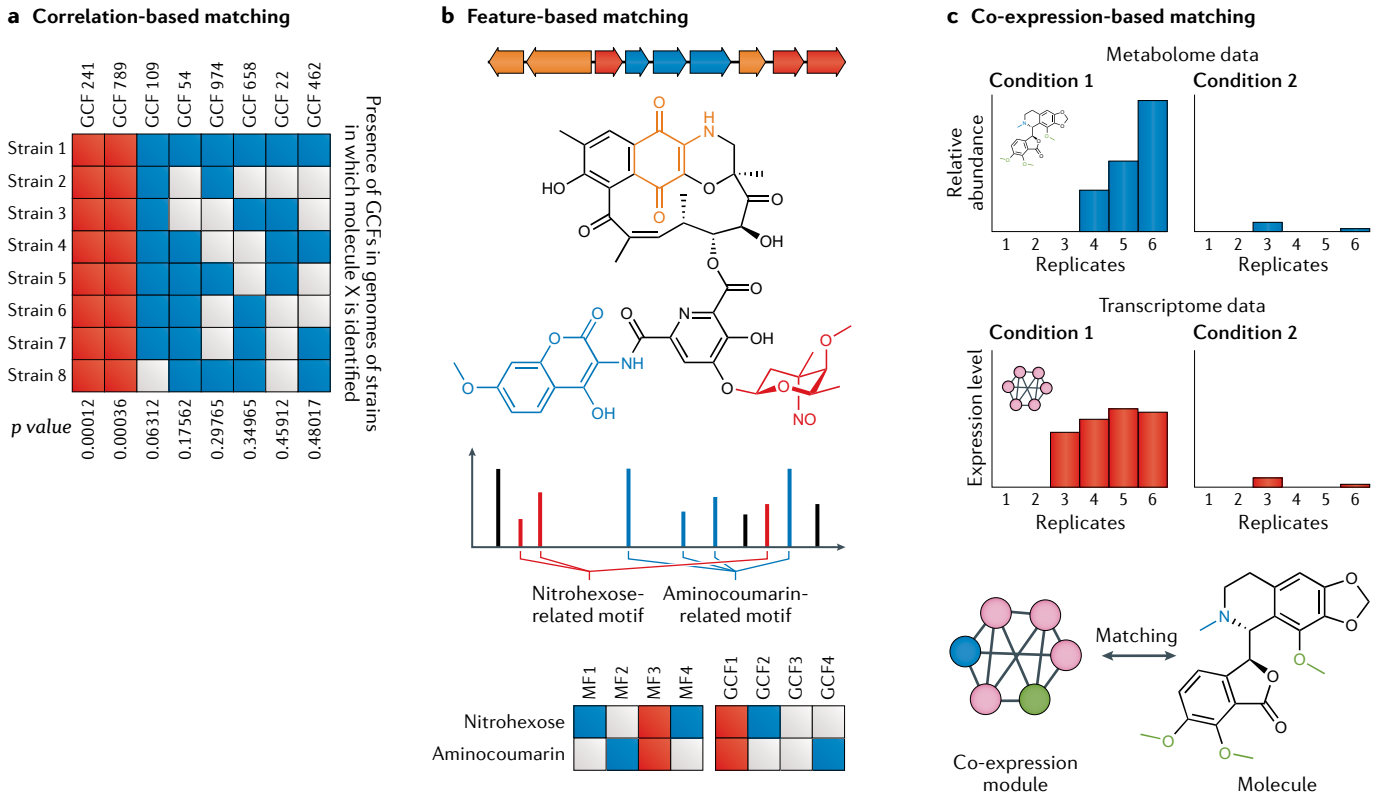


Fig. 2 | **Linking genes to molecules using metabolomics and transcriptomics.** Several approaches have been developed to link metabolites to genes and gene clusters encoding their biosynthesis. **a** | In bacteria, pattern-based genome mining approaches have been developed that match families of molecules (related by spectral similarity) to gene cluster families (GCFs; related by sequence similarity) through metabologenomic correlation[131], identifying which GCFs co-occur strongly in the same strains where a given metabolite is observed. **b** | Molecules can also be connected to genes and gene clusters through feature-based matching, in which chemical features (substructures and modifications that are either manually annotated or identified using algorithms that identify motifs in tandem mass spectrometry data) are linked to genes and gene modules that are known to be responsible for the biosynthesis of such features. **c** | Transcriptomic data can also be used to identify potential biosynthetic pathways for a molecule of interest by, for example, identifying modules of co-expressed genes whose expression correlates with the presence of a given metabolite across a range of divergent conditions (for example, different biological stresses[143]). MF, molecular family.

highly similar in terms of domain composition, although having evolved independently and leading to divergent chemical outcomes[128]. These analyses suggest that at least certain categories of biosynthetic pathways in plants evolve through combinatorialization of a limited set of enzyme families, of which the members can have different catalytic activities or act upon different sites within their target molecules. Hence, for pathway types and organisms in which gene evolution is largely decoupled from gene cluster evolution, more automated phylogenetic methods need to be developed to perform comparative analysis at the gene level as well as the gene cluster level. Beyond plants, it should not be excluded that this is the case for other eukaryotic branches of the tree of life as well.

Improving genome mining with other data

All genome mining techniques discussed thus far rely on analysing genomic or transcriptomic sequence data on their own. However, the predictive power of these approaches can be further enhanced by combining them with different types of information, such as gene expression levels (as measured by, for example, RNA sequencing or quantitative proteomics) or known phenotypes or bioactivities exhibited by the organisms, or extracts thereof. Analytical chemistry data, such as the presence and intensity of chromatographic peaks and fragmentation patterns observed in tandem mass spectra, can be particularly valuable for the discovery of metabolites and their biosynthetic genes[129]. For instance, if the same or similar molecules are produced by different organisms, they can be expected to harbour the same or similar biosynthetic genes. Pattern-based genome mining[130] (also known as metabologenomic correlation analysis[131,132]; FIG. 2a) correlates the presence of metabolites to homologous biosynthetic genes across strains. This approach (reviewed in detail in REF.[133]) has mostly been pioneered in bacteria, for which sufficiently large numbers of genomes and metabolomes can be obtained. In one metabologenomic correlation study, gene cluster families were linked to a molecular network based on mass-spectral fragmentation patterns, leading to the discovery of the tyrobetaine metabolites[134]. Recently, the mathematics behind the association scoring were improved and formalized in a software tool called NPLinker[135]. The advantage of this technology is

that no prior knowledge on biosynthetic mechanisms is required to link molecules to gene clusters, as it is purely based on correlations. A strategy that establishes genomic–metabolomic co-occurrence patterns has great potential to mine the genomes of understudied organisms, even when virtually nothing is known about a taxon's enzymology.

Another approach that also harnesses analytical chemistry to improve genome mining predictions is the correlation of mass shifts in tandem mass spectrometry fragmentation patterns to a BGC's bioinformatically predicted building blocks (FIG. 2b). At first, semi-manual approaches were developed that allowed matching of peptides (peptidogenomics[14]) and glycosylated specialized metabolites (glycogenomics[136]) to BGCs. More recently, this matching has been automated for peptides in algorithms such as Pep2Path (REF.[137]), RiPPquest[138] and MetaMiner[139]. These algorithms, which have a major focus on RiPPs, could also be very relevant for finding novel peptidic metabolites in uncharted taxa, as recent evidence is emerging that RiPPs are produced by not only bacteria but also fungi[140], plants[50] and animals[105]. Going forward, the bigger challenge will be to extend these approaches beyond peptides to specialized metabolites in general[133].

Instead of partial structural information from mass spectra, previously elucidated chemical structures can also be used to connect biosynthetic genes to 'orphan' metabolites and, conversely, identify those coding for novel molecules. There are many specialized metabolites for which the chemical structure is known but the biosynthetic genes are not. For drug discovery purposes, not having the opportunity to check for novelty by sequence may pose a major problem, given the considerable effort wasted elucidating the chemical structure of a known molecule. Recently, the innovative method GRAPE/GARLIC was established[141] to connect genes to molecules for polyketides and non-ribosomal peptides in an automated fashion: by breaking down known specialized metabolite structures into their biochemical building blocks and retro-biosynthetically matching these with building blocks predicted to be incorporated into molecules based on BGC sequence information, the authors were able to identify thousands of putative matches between gene clusters and molecules. Of ~16,831 BGCs, approximately 2,500 had best-matching scores to reference molecules that were so low they very likely encode the biosynthetic machinery for novel products. Although this number may seem fairly small, one should consider that the remaining set of ~14,000 BGCs is enriched with many near-copies of BGCs from highly studied taxa for which large numbers of genomes have been sequenced. The retro-biosynthetic principle, although useful, seems largely limited to bacterial polyketides and non-ribosomal peptides, and expanding retro-biosynthetic algorithms to other life forms will require considerable expansions of our knowledge of their biosynthetic routes. Training more generic models for enzymatic mechanisms based on large-scale experimental data is needed here, as well as high-throughput assays on 'enzymatic dark matter' from unexplored taxa to provide the required training data for such models.

The presence of specialized metabolites can also be correlated to biosynthetic genes' transcriptional levels in different conditions or across different tissues (FIG. 2c). For example, the biosynthetic pathway for ingenol mebutate from *Euphorbia* plants was unravelled by identifying members of relevant enzyme families that were highly expressed in seeds[142]. Similarly, another recent study analysed the production of the defence metabolite falcarindiol by tomato across seven different biotic stress treatments, to identify relevant enzyme-coding genes that were upregulated in conditions when increased amounts of the molecule were observed[143]. This principle seems universally applicable and is widely useful for accelerating genome mining efforts.

Indeed, in plants, co-expression analysis has already been frequently used with success to identify genes that show similar expression patterns across a large number of samples, within the same species or even cross-species[144]. Often, this is done using one or more 'bait' genes, which are predicted or even known to belong to a pathway of interest, to recruit additional members of that pathway[145,146]. However, unsupervised approaches are also being developed, which can be used to predict candidate pathways without prior knowledge. These methods rely on detecting co-expressed modules of genes given a set of transcriptomic samples, a procedure for which a range of algorithms is available[147]. Recently, the identification of co-expression modules was shown to effectively and comprehensively retrieve genes implicated in methionine-derived aliphatic glucosinolate biosynthesis in *Arabidopsis thaliana* and *Brassica rapa*[148]. A key factor in the success of this study was the use of a graph clustering method that allows modules to overlap in their gene content, which makes sense given that specialized metabolic enzymes from plants are often promiscuous and may have dual functions in multiple pathways. In general, the advantage of co-expression approaches seems to be that they are generally applicable, also when the genes encoding a pathway of interest are only partially clustered or not clustered at all. Moreover, for eukaryotes with complex genomes that are hard to assemble contiguously, co-expression-based approaches could also be performed on the basis of fragmented genome assemblies or transcriptome assemblies. A challenge for these approaches is how to find the right combination of conditions that distinguishes expression patterns of a pathway of interest most effectively from those of other pathways, without requiring massive amounts of expensive transcriptome sequencing. One possible strategy to do this would be to first generate (targeted or untargeted) metabolome data for various samples, before choosing which samples are prioritized for RNA sequencing. Alternatively, integrative approaches could be developed that leverage structural information from metabolome data (for example, mass shifts and predicted substructures) to help prioritize which sets of co-expressed enzyme-coding genes are most likely responsible for the production of a given metabolite.

Function-first approaches

No matter how powerful modern genome mining approaches are to identify the genomic basis for chemical diversity, these methods are fairly blind and

untargeted — usually, a molecule's physiological and ecological importance is only considered at the very end, after structural characterization and elucidation of its biosynthetic pathway. Function has traditionally been investigated only in a very narrow sense, that is, by considering hits in activity assays relevant to human health and prosperity, to the neglect of physiological and subtler ecological functions. Functions such as the arthropod-attracting capabilities of geosmin and 2-methylisoborneol terpenoids from streptomycete bacteria[149] or the conferring of heat stress resilience by flavonols by regulating levels of reactive oxygen species[150] were only identified decades after these metabolites were structurally characterized. To truly deepen our understanding of the fundamental roles of these molecules in biology and to allow for more targeted approaches to leverage them in, for example, drug discovery, it will be crucial to devise methods to help prioritize biosynthetic pathway candidates based on the specialized metabolite's predicted function.

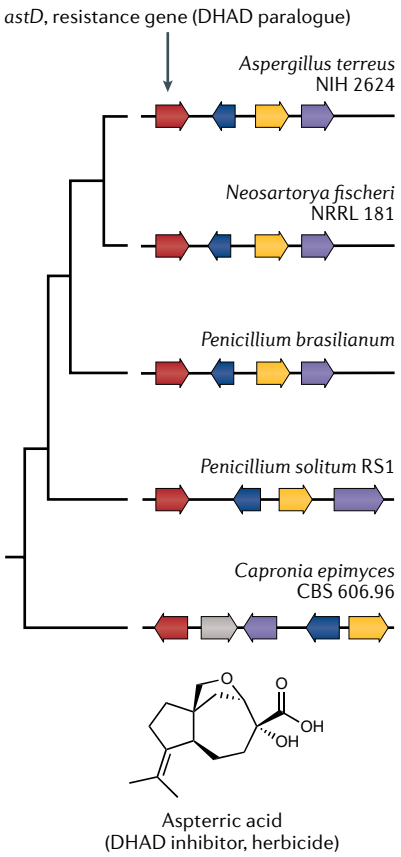
Target-directed genome mining. A good example of such a 'function-first' method, which has already gained traction, is based on the co-localization of genes within the same BGC that are indicative of function. For example, the co-localization of iron transport genes with biosynthetic genes led to the discovery of siderophore molecules, such as coelichelin and salinichelins in bacteria[151] and sideretin from plants[152] (and this principle has recently been generalized[153]). The co-localization of resistance genes or duplicated genes resembling antimicrobial targets within BGCs offers a more generalizable approach to the discovery of bioactive molecules with specific cellular targets (FIG. 3a). This approach, called target-directed genome mining, was first validated with the rediscovery of the thiolactomycin antibiotics as fatty acid synthase inhibitors from orphan bacterial BGCs that contain an open reading frame predicted to be a resistance gene[27], associated with target modification of the FabF fatty acid ketosynthase. Newer studies co-localizing putative target-modifying resistance genes with BGCs to identify compounds with activities against the resistance gene target include the proteasome inhibitor fellutamide B from the fungus *A. nidulans*[154] and the topoisomerase inhibitors pyxidicylines from the myxobacterium *Pyxidicoccus fallax* An d48 (REF.[155]). A clever twist on this resistance gene-guided approach led to the discovery of the fungal sesquiterpenoid aspterric acid as a potent herbicide, by deploying the fungal dihydroxy acid dehydratase self-resistance gene as a transgene to render plants resistant to aspterric acid[156]. To automate the resistance-based genome mining procedure, a web service called the Antibiotic Resistant Target Seeker (ARTS) was developed to identify BGCs containing likely self-resistance genes, suggesting they code for the production of specialized metabolites with specific biological targets[157]. Intuitively, the approach is probably applicable to any organisms in which biosynthetic pathways are genomically clustered, so long as there is sufficient selective pressure for the resistance genes to co-cluster (through facilitating co-expression and co-inheritance with the pathway).

Although resistance-based genome mining is a breakthrough as a key function-first strategy, the vast majority of BGCs do not contain self-resistance genes or other genes that unambiguously indicate a specific function. Hence, there is a great need for the development of additional strategies to generate hypotheses about the function of the molecules produced by the remaining majority of pathways. We believe that, again, the essence of these approaches will be in combining genomics with other types of data. Below, we outline three possible ways in which this could be achieved.

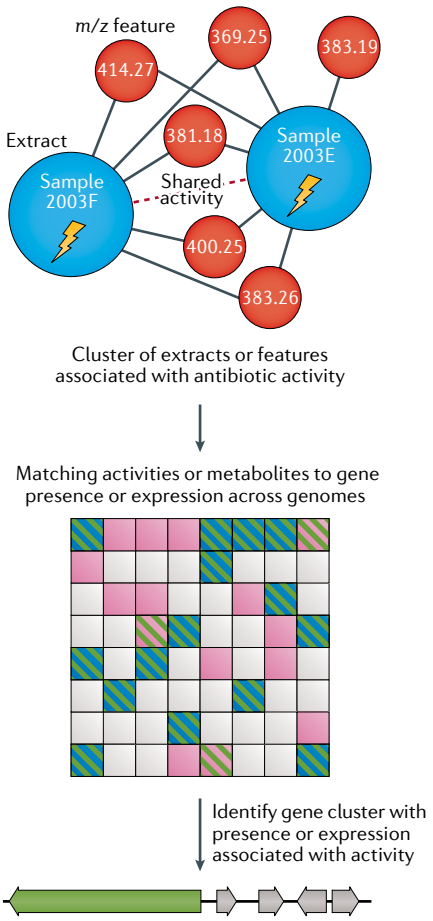
Cytological profiling and compound activity mapping. A first possibility would entail correlating genomic information to bioactivities displayed by extracts (FIG. 3b). There has already been some success in correlating bioactivities of extracts as determined by cytological profiling[158] to untargeted metabolomics of the same extracts using a technique called Compound Activity Mapping[159], facilitating the discovery of the quinocinnolinomycins, a new family of specialized metabolites that cause endoplasmic reticulum stress. The obvious next step will be to combine this with genomic and/or transcriptomic data to immediately identify the genes responsible for an activity of interest. Also, when cytological profiling does not give immediate insights into the mode of action of a molecule, it could be complemented with transcriptome analysis of the target cells during exposure. Indeed, machine learning methods have recently been devised that predict pharmacological properties of drug molecules, directly related to the mechanism of action, based on large-scale transcriptional response data[160]. In principle, this approach would be applicable to any life forms for which extracts can be made, including many protists, plants and invertebrates. This could also be done through genome-wide association studies that map phenotypes to genetic variation within a species, as has been successfully practised to discover the cucurbitacin gene cluster responsible for the bitter taste in cucumber[161].

Co-expression-based function prediction. A second way to perform function-first genome mining would be to study the effects of the expression of BGCs on other community members within their native ecosystem, and, optionally, how they relate to emergent properties of such an ecosystem (FIG. 3c). This applies primarily to microbial ecosystems and microbiota associated with plant or animal hosts. For example, metatranscriptome data from soil microbial communities were recently used to investigate the ecological roles of BGCs from novel bacterial clades identified through metagenomic binning; co-expression of BGCs with iron starvation response genes or antimicrobial resistance genes thus indicated roles for their products as siderophores or antimicrobials[162]. This concept could be extended by also looking at co-expression across species, that is, correlating the expression of putative antibiotic biosynthesis BGCs with stress responses in other organisms in the community to identify the likely target organisms. The expression of specific BGCs could also be correlated to microbiome-associated phenotypes[163] that a community confers to its host, such as disease suppression or

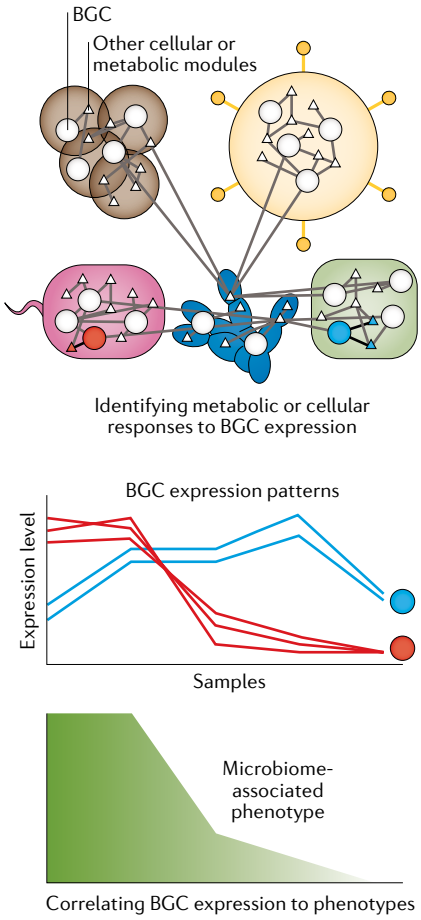
a Target-based genome mining using resistance genes



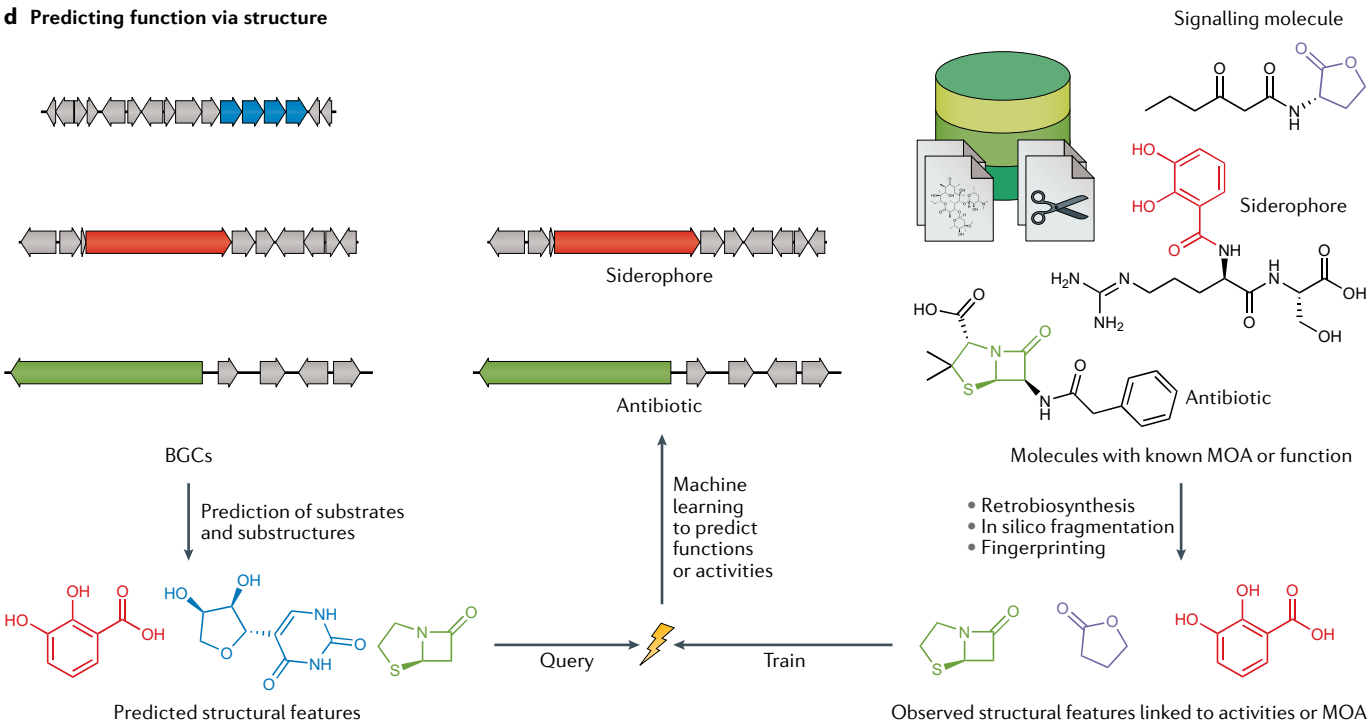
b Cytological profiling and compound activity-mapping



c Predicting function based on co-expression analysis



d Predicting function via structure



◀ Fig. 3 | **Function-first genome mining approaches.** In order to more effectively identify molecules with desired activities, function-first genome mining approaches have been and are being developed. **a** | In target-based genome mining approaches, self-resistance genes are identified that genomically cluster with the biosynthetic genes. Such self-resistance genes are often resistant copies of a housekeeping gene whose protein product is targeted by the metabolite biosynthesized from the pathway. This provides a way to directly predict the mechanism of action for metabolic products of a subset of gene clusters. **b** | Cytological profiling can be used to identify the effects that metabolic extracts have on certain cell lines, and compound activity mapping can identify which underlying mass-spectral features are likely responsible for activities that are shared between extracts[159]. The activities and/or metabolites can then be matched to the presence or expression of genes and gene clusters to identify a candidate biosynthetic route towards the underlying molecule. **c** | Functions of products of biosynthetic genes and gene clusters can be predicted by looking for co-expression with other genes in the same organism (predicting function based on the guilt by association principle) or across organisms (identifying the potential effect that a pathway has on other organisms or on a microbiome-associated phenotype). **d** | Structural features and substructures that are likely part of the metabolic product of a gene cluster can be predicted in silico; sometimes, these substructures are diagnostic for a certain mechanism of action or biological activity, and machine learning algorithms can be trained to predict these activities based on sets of structural features. BGC, biosynthetic gene cluster; MOA, mechanism of action.

stress resilience, to identify which molecules are likely to be responsible for mediating these phenotypes. In host organisms, such as plants and animals, expression of particular biosynthetic pathways can also be linked to functions by studying the effects on the microbiome composition; for example, a recent study linked specific triterpene pathways to either the promotion or inhibition of specific rhizosphere microbiome community members, which highlighted their specific roles in microbiome modulation[164].

Predicting function via structure. A third strategy for function-first genome mining would be combining (sub) structure prediction from sequences with structure-based prediction of biological activities and macromolecular targets (FIG. 3d). Both of these prediction tasks are currently highly prone to error, but significant progress is being made on both fronts, so a robust platform may become a reality in the not too distant future. Several tools are currently emerging that can predict the core scaffolds of key classes of specialized metabolites from sequence information with increasing accuracy and detail[112,113,165,166], and several efforts are underway to complement these with additional predictions of tailoring and cyclization reactions[113,167]. Also, genome-based structure predictions could be integrated with metabolomics-based (sub)structure predictions[168,169], which could confirm or guide routes through biochemical reaction space. Based on these developments, considerable improvements in specialized metabolite structure prediction from genome and metabolome data can be expected in the near future.

Meanwhile, within the field of computational drug discovery, methods are emerging that allow predicting macromolecular targets of drug molecules based on their chemical structures[170]. For example, the algorithm SPIDER dissects specialized metabolites into pharmacophore-sized fragments and predicts which proteins a compound targets by comparison with a library of 13,695 chemical structures of molecules of known function from the Collection of Bioactive Reference Analogues (COBRA)[171]. This method successfully predicted polypharmacological features of the macrolide archazolid A. Similarly, in another recent study, a deep learning model was trained that could successfully predict antibiotic activities of molecules with only limited chemical similarity to those used for training[172]. When, in the future, both sequence-based metabolite structure prediction and structure-based macromolecular target prediction become increasingly accurate, they could be coupled to predict biological targets directly from gene cluster sequences. The recently published PRISM4 pipeline provides a first step in this direction, using support-vector machines to predict the activities of the molecular products of gene clusters based on their predicted structures[173]. For the moment, this strategy is likely to be relevant mostly for bacteria and fungi, and to some extent for plants; however, when synthetic biology approaches[63] and in vitro expression systems[174] increasingly allow experimental characterization of large sets of enzymes from animals and protists, opportunities will likely emerge to apply this strategy in these taxa as well.

Testing candidate biosynthetic genes

Fundamentally, there are three types of approaches to identify the metabolic product(s) of a BGC: heterologous expression in a model organism, either in the BGC's original form or after refactoring; genetic manipulation of the native host; and in vitro reconstitution.

Heterologous expression. Heterologous expression involves the cloning (also known as 'capture') of a BGC or non-clustered biosynthetic genes into one or more plasmids, cosmids or artificial chromosomes, possible manipulation of the BGC, transfer into a genetically tractable heterologous host and testing for the presence of metabolic products compared with the unmodified heterologous host[175–177]. If possible, heterologous expression is a highly desired approach, because it enables both facile scale-up of metabolite production for structural elucidation and biological testing, and manipulation of the captured BGC for biosynthetic investigations and analogue production. The large size of many BGCs has spurred the development of cloning methods that can capture BGCs directly from genomic DNA, such as transformation-associated recombination in yeast[178,179], linear–linear homologous recombination in *Escherichia coli*[180] or programmable nucleases in vitro[181,182]. One benefit of these PCR-free techniques is that they avoid mutation of the BGC, making sequence verification unnecessary. BGCs can also be cloned and assembled using PCR-based techniques, but as sequence verification of large BCGs by Sanger sequencing can be a bottleneck, doing so using next-generation sequencing technologies[183] will likely gain popularity.

However, heterologous expression has some notable potential challenges: promoters and ribosome-binding sites may not be recognized; genes may require RNA splicing; proteins may require chaperones, post-translational modification or transport to organelles; required metabolic precursors or cofactors may not be present; or the heterologous pathway could encounter

Heterologous host
An organism different from the source organism of a gene under investigation, usually a model organism with a well-developed genetic toolkit. A heterologous host optimized for a specific biotechnological application such as small-molecule production is sometimes called a 'chassis'.

metabolic bottlenecks due to non-optimal enzyme stoichiometry. If the pathway's reactions are impeded to different extents, heterologous production could result in the production of metabolic intermediates or shunt products instead of the 'true' specialized metabolite. Conventional wisdom states that employing heterologous hosts that are phylogenetically close relatives to the organism from which the BGC originates improves the chances of success, but exceptions to this dogma are known, caused, for instance, by unexpected interactions with a host's gene regulatory machinery[184]. Techniques such as CRAGE[185] aim to streamline testing a BGC in a multitude of heterologous hosts, increasing the chances of at least one succeeding. Research dedicated to developing genetic toolkits for various organisms will be crucial to streamline the heterologous expression of BGCs from organisms not closely related to classic model organisms.

Synthetic biology and refactoring. Synthetic biology approaches aim to circumvent the aforementioned challenges associated with heterologous expression by 'refactoring' the candidate biosynthetic genes and/or engineering heterologous hosts ('chassis') optimized for heterologous expression of biosynthetic pathways. Chassis have been developed that provide metabolic precursors and post-translational modifications required for specific classes of specialized metabolism or to inactivate competing metabolic pathways. Refactoring usually entails bringing candidate biosynthetic genes under the control of well-characterized promoters and ribosome-binding sites, elimination of introns and organellar targeting signals, and codon optimization[63]. However, gaps in our understanding of these cellular processes — for instance, how codon optimization affects gene expression and protein folding — still limit the rationality of refactoring efforts. Several streamlined workflows for refactoring candidate biosynthetic genes have been described[186,187]. The use of combinatorial libraries[188] and independently tunable promoters[189] can help optimize the stoichiometry of biosynthetic genes in vivo. Although fully synthesizing refactored BGCs de novo, instead of refactoring captured BGCs, is currently still prohibitively expensive for all but the best-funded projects, we expect this practice to become widespread as gene synthesis costs continue to decline.

Genetic manipulation of the native host. Alternatively, the candidate gene(s) can be inactivated or repressed in their native host, followed by testing for the loss of, or decrease in the quantity of, a metabolite compared with the wild-type host. To more thoroughly establish the gene–metabolite link, ideally a genetic complementation experiment should also be carried out[190]. The biggest drawback to this approach is that it can be difficult or impossible to manipulate genes in non-model organisms, but thankfully this situation is improving thanks to the broad host range of CRISPR–Cas9 technologies. The emergence of CRISPR–Cas9-based 'microbiome editing' technologies[191,192] has even made it possible to knock out genes in specific members of a complex microbiome.

In vitro reconstitution. Reconstitution of the pathway in vitro provides some advantages orthogonal to the in vivo approaches above, such as allowing for easier identification of pathway intermediates, determination of enzyme kinetics and substrate specificities, and quick optimization of the pathway's enzyme stoichiometry[174]. However, in vitro reconstitution can be challenging if the metabolic precursor(s) or order of the enzymes in the metabolic pathway is unknown, or if any of the enzymes are insoluble, unstable or cannot be purified.

Structural elucidation of biosynthetic products. Once a metabolite has been identified as being the product of the candidate genes, its identity will need to be established. Depending on the method that was used to select the candidate genes, one may already have a hypothetical structure or chemical class. The act of 'dereplication' seeks to quickly identify whether the metabolite is, or is closely related to, any known molecule. Some currently popular approaches to dereplication are based on tandem mass spectrometry spectral networking (such as GNPS[8]), tandem mass spectrometry spectral-substructure matching (such as VarQuest[193], MS2LDA (REF.[169]) and CSI:FingerID[168]) and NMR spectral clustering (such as SMART[194]), but it is worth remembering that dereplication tools are only as effective as the databases/training data that underlie them. If the molecule is likely novel, structural elucidation will be necessary. Nowadays, this is most commonly achieved through 2D-NMR techniques, with a slow uptick in the application of computer-assisted structure elucidation[195] technologies. X-ray crystallography (occasionally aided by the crystalline sponge method[196]) and, more recently, microcrystal electron diffraction[197] can also provide important insights into challenging structural elucidation problems.

Chemical synthesis of predicted BGC products. Finally, some recent studies circumvent biological experimentation altogether by chemically synthesizing the predicted products of a BGC[198–201]. BGCs for RiPPs and non-ribosomally synthesized peptides are particularly amenable to this approach, as the structures of their products are highly predictable and their production can be streamlined through solid-phase peptide synthesis. Although doubt about the true identity of the BGC's product remains, this approach has yielded molecules with promising biological activities[198–200].

Conclusions and future perspectives

What else is there to mine, and what happens to genome mining after we have exhaustively identified all specialized metabolite scaffolds? Based on the inventory of known specialized metabolites and those that are already connected to biosynthetic genes, the future remains bright. Considering the efficiency and breadth of new strategies for genome mining and given the increased extent of resources available for mining, many new sources, enzymes and metabolites are expected to be discovered over the coming years. Even when mining of orphan genes leads to rediscovery of previously reported specialized metabolites, new biosynthetic

knowledge may have biotechnological utility for (enhanced) biological production of these and related molecules.

Biosynthetic gene identification and prioritization are moving towards the incorporation of an increasingly large number of different data types. Moving forward, pioneering approaches will likely harness an even larger number of data types simultaneously. Integrating multi-omics data, although computationally challenging, has great potential to identify true gene–metabolite relationships among thousands of potential ones, especially across larger sets of related organisms for which sequence-based predictions of metabolite structures can be combined with absence–presence patterns of candidate genes[133]. Improvements in documenting links between different types of omics data[129], statistical association techniques[135] and machine learning technologies for sequence-based prediction of enzyme activities and metabolite structures[173] will further accelerate such efforts. Moreover, these omics data will provide new ways to assess metabolite function at an early stage, by evaluating the triggers and consequences of the expression of their biosynthetic genes.

The study of the chemistry of life has been brought to a next level by genome mining technologies initially developed in microorganisms. Now that large-scale genome sequencing is expanding to all branches of the tree of life, there is a great opportunity to port and extend genome mining technologies to other life forms and engage in truly global studies of life's chemistry. At the same time, the microbial field has much to learn from scientists studying humans and mammals, who have been very effective at identifying physiological roles of mammalian specialized metabolites such as steroids, prostaglandins and peptide hormones. Additionally, plant biologists' extensive experience using gene expression analysis to link genes to molecules and identify their functions may become incredibly useful to the microbial field to acquire deeper perspectives into the physiological roles of many metabolites that have appeared 'inert' for so long. Finally, protists and invertebrates provide an immense uncharted biological diversity that is mostly untapped and likely to yield numerous new and surprising findings.

All in all, great potential presents itself in unifying these diverse scientific communities to find common ground between molecules and genes that may have seemed unrelated for so long. This will facilitate a deeper fundamental biological understanding of the ecological and physiological roles of life's chemistry, more effectively leveraging it for the common good in medicine, agriculture and nutrition.

Published online 3 June 2021

1. Davies, J. Specialized microbial metabolites: functions and origins. *J. Antibiot.* **66**, 361–364 (2013).
2. Chevrette, M. G. et al. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).
3. Erb, M. & Kliebenstein, D. J. Plant secondary metabolites as defenses, regulators, and primary metabolites: the blurred functional trichotomy. *Plant. Physiol.* **184**, 39–52 (2020).
This review provides a useful discussion on the categories of secondary metabolites, primary metabolites and hormones, and cases where these definitions overlap.
4. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes — a review. *Nat. Prod. Rep.* **33**, 988–1005 (2016).
5. Medema, M. H. & Osbourn, A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Nat. Prod. Rep.* **33**, 951–962 (2016).
6. Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nat. Rev. Microbiol.* **17**, 167–180 (2019).
7. Lockermann, G. Friedrich Wilhelm Serturner, the discoverer of morphine. *J. Chem. Educ.* **28**, 277–279 (1951).
8. Wang, M. et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
9. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl Acad. Sci. USA* **114**, 5601–5606 (2017).
This retrospective analysis quantifies bacterial and fungal natural products identified over the years and provides a perspective on the amount of structural novelty that is still being unearthed.
10. Bentley, S. D. et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
This paper is, in many ways, the foundation for the field of natural product genome mining; the genome sequence of *S. coelicolor* makes it clear that the coding capacity for specialized metabolite production is much greater than the number of metabolites that have been discovered from this model species.
11. Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* **1**, 265–269 (2005).
12. Lin, X., Hopson, R. & Cane, D. E. Genome mining in *Streptomyces coelicolor*: molecular cloning and characterization of a new sesquiterpene synthase. *J. Am. Chem. Soc.* **128**, 6022–6023 (2006).
13. Corre, C., Song, L., O'Rourke, S., Chater, K. F. & Challis, G. L. 2-Alkyl-4-hydroxymethylfuran-3-carboxylic acids, antibiotic production inducers discovered by *Streptomyces coelicolor* genome mining. *Proc. Natl Acad. Sci. USA* **105**, 17510–17515 (2008).
14. Kersten, R. D. et al. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).
This study pioneers the use of feature-based matching to link genes to molecules, focusing on ribosomally and non-ribosomally synthesized peptides in bacteria.
15. Gomez-Escribano, J. P. et al. Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the *cpk* gene cluster of *Streptomyces coelicolor* M145. *Chem. Sci.* **3**, 2716 (2012).
16. Cruz-Morales, P. et al. Phylogenomic analysis of natural products biosynthetic gene clusters allows discovery of arseno-organic metabolites in model streptomycetes. *Genome Biol. Evol.* **8**, 1906–1916 (2016).
17. Malpartida, F. & Hopwood, D. A. Molecular cloning of the whole biosynthetic pathway of a *Streptomyces* antibiotic and its expression in a heterologous host. *Nature* **309**, 462–464 (1984).
18. Smith, D. J., Burnham, M. K. R., Edwards, J., Earl, A. J. & Turner, G. Cloning and heterologous expression of the penicillin biosynthetic gene cluster from *Penicillium chrysogenum*. *Nat. Biotechnol.* **8**, 39–41 (1990).
19. Feitelson, J. S., Malpartida, F. & Hopwood, D. A. Genetic and biochemical characterization of the red gene cluster of *Streptomyces coelicolor* A3(2). *J. Gen. Microbiol.* **131**, 2431–2441 (1985).
20. Fazio, G. C., Xu, R. & Matsuda, S. P. T. Genome mining to identify new plant triterpenoids. *J. Am. Chem. Soc.* **126**, 5678–5679 (2004).
This paper constitutes the first demonstration of genome mining from a plant.
21. Bergmann, S. et al. Genomics-driven discovery of PKS–NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat. Chem. Biol.* **3**, 213–217 (2007).
22. Franke, J., Ishida, K. & Hertweck, C. Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic *Burkholderia* species. *Angew. Chem. Int. Ed. Engl.* **51**, 11611–11615 (2012).
23. Biggins, J. B., Ternei, M. A. & Brady, S. F. Malleilactone, a polyketide synthase-derived virulence factor encoded by the cryptic secondary metabolome of *Burkholderia pseudomallei* group pathogens. *J. Am. Chem. Soc.* **134**, 13192–13195 (2012).
24. Pidot, S., Ishida, K., Cyrulies, M. & Hertweck, C. Discovery of clostrubin, an exceptional polyphenolic polyketide antibiotic from a strictly anaerobic bacterium. *Angew. Chem. Int. Ed. Engl.* **53**, 7856–7859 (2014).
25. Claesen, J. & Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl Acad. Sci. USA* **107**, 16297–16302 (2010).
26. Du, Y.-L., He, H.-Y., Higgins, M. A. & Ryan, K. S. A heme-dependent enzyme forms the nitrogen–nitrogen bond in piperazate. *Nat. Chem. Biol.* **13**, 836–838 (2017).
27. Tang, X. et al. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.* **10**, 2841–2849 (2015).
28. Dassama, L. M. K., Kenney, G. E. & Rosenzweig, A. C. Methanobactins: from genome to function. *Metallomics* **9**, 7–20 (2017).
29. de Rond, T., Asay, J. E. & Moore, B. S. Co-occurrence of enzyme domains guides the discovery of an oxazolone synthetase. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-021-00808-4 (2021).
30. Obermaier, S. & Müller, M. Ibotenic acid biosynthesis in the fly agaric is initiated by glutamate hydroxylation. *Angew. Chem. Int. Ed. Engl.* **59**, 12432–12435 (2020).
31. Marchand, J. A. et al. Discovery of a pathway for terminal-alkyne amino acid biosynthesis. *Nature* **567**, 420–424 (2019).
32. Zhu, X., Liu, J. & Zhang, W. De novo biosynthesis of terminal alkyne-labeled natural products. *Nat. Chem. Biol.* **11**, 115–120 (2015).
33. Ng, T. L., Rohac, R., Mitchell, A. J., Boal, A. K. & Balskus, E. P. An N-nitrosating metalloenzyme

constructs the pharmacophore of streptozotocin. *Nature* **566**, 94–99 (2019).
34. Waldman, A. J. & Balskus, E. P. Discovery of a diazo-forming enzyme in cremeomycin biosynthesis. *J. Org. Chem.* **83**, 7539–7546 (2018).
35. Agarwal, V. et al. Metagenomic discovery of polybrominated diphenyl ether biosynthesis by marine sponges. *Nat. Chem. Biol.* **13**, 537–543 (2017).
36. Baccile, J. A. et al. Plant-like biosynthesis of isoquinoline alkaloids in *Aspergillus fumigatus*. *Nat. Chem. Biol.* **12**, 419–424 (2016).
37. Caputi, L. et al. Missing enzymes in the biosynthesis of the anticancer drug vinblastine in Madagascar periwinkle. *Science* **360**, 1235–1239 (2018).
38. Satake, M. et al. Brevisin: an aberrant polycyclic ether structure from the dinoflagellate *Karenia brevis* and its implications for polyether assembly. *J. Org. Chem.* **74**, 989–994 (2009).
39. Sinninghe Damsté, J. S. et al. Linearly concatenated cyclobutane lipids form a dense bacterial membrane. *Nature* **419**, 708–712 (2002).
40. Rattray, J. E. et al. A comparative genomics study of genetic products potentially encoding ladderane lipid biosynthesis. *Biol. Direct* **4**, 8 (2009).
41. Arnison, P. G. et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* **30**, 108–160 (2013).
42. Montalbán-López, M. et al. New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
43. Li, Y. & Rebuffat, S. The manifold roles of microbial ribosomal peptide-based natural products in physiology and ecology. *J. Biol. Chem.* **295**, 34–54 (2020).
44. Hansen, J. N., Norman Hansen, J. & Sandine, W. E. Nisin as a model food preservative. *Crit. Rev. Food Sci. Nutr.* **34**, 69–93 (1994).
45. Schmidtko, A., Lötsch, J., Freynhagen, R. & Geisslinger, G. Ziconotide for treatment of severe chronic pain. *Lancet* **375**, 1569–1577 (2010).
46. Morinaka, B. I. et al. Natural noncanonical protein splicing yields products with diverse β-amino acid residues. *Science* **359**, 779–782 (2018).
47. Freeman, M. F., Helf, M. J., Bhushan, A., Morinaka, B. I. & Piel, J. Seven enzymes create extraordinary molecular complexity in an uncultivated bacterium. *Nat. Chem.* **9**, 387–395 (2017).
48. Umemura, M. et al. Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in *Aspergillus flavus*. *Fungal Genet. Biol.* **68**, 23–30 (2014).
49. Nagano, N. et al. Class of cyclic ribosomal peptide synthetic genes in filamentous fungi. *Fungal Genet. Biol.* **86**, 58–70 (2016).
50. Kersten, R. D. & Weng, J.-K. Gene-guided discovery and engineering of branched cyclic peptides in plants. *Proc. Natl Acad. Sci. USA* **115**, E10961–E10969 (2018).
51. Jordan, P. A. & Moore, B. S. Biosynthetic pathway connects cryptic ribosomally synthesized posttranslationally modified peptide genes with pyrroloquinoline alkaloids. *Cell Chem. Biol.* **23**, 1504–1514 (2016).
52. Ting, C. P. et al. Use of a scaffold peptide in the biosynthesis of amino acid-derived natural products. *Science* **365**, 280–284 (2019).
53. Kazandjian, T. D. et al. Convergent evolution of pain-inducing defensive venom components in spitting cobras. *Science* **371**, 386–390 (2021).
54. Pineda, S. S. et al. Structural venomics reveals evolution of a complex venom by duplication and diversification of an ancient peptide-encoding gene. *Proc. Natl Acad. Sci. USA* **117**, 11399–11408 (2020).
55. Sanggaard, K. W. et al. Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Commun.* **5**, 3765 (2014).
56. Wang, G. Human antimicrobial peptides and proteins. *Pharmaceuticals* **7**, 545–594 (2014).
57. Gu, S. et al. Competition for iron drives phytopathogen control by natural rhizosphere microbiomes. *Nat. Microbiol.* **5**, 1002–1010 (2020).
58. Carrión, V. J. et al. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. *Science* **366**, 606–612 (2019).
This study illustrates how metagenome mining can be used to identify biosynthetic genes responsible for a microbiome-associated phenotype, fungal disease suppression in this case.
59. Guo, C.-J. et al. Discovery of reactive microbiota-derived metabolites that inhibit host proteases. *Cell* **168**, 517–526.e18 (2017).
60. Santhanam, R. et al. Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proc. Natl Acad. Sci. USA* **112**, E5013–E5020 (2015).
61. Durán, P. et al. Microbial interkingdom interactions in roots promote *Arabidopsis* survival. *Cell* **175**, 973–983.e14 (2018).
62. D'hoe, K. et al. Integrated culturing, modeling and transcriptomics uncovers complex interactions and emergent behavior in a three-species synthetic gut community. *eLife* **7**, e37090 (2018).
63. Smanski, M. J. et al. Synthetic biology to access and expand nature's chemical diversity. *Nat. Rev. Microbiol.* **14**, 135–149 (2016).
64. Reed, J. et al. A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules. *Metab. Eng.* **42**, 185–193 (2017).
65. Eng, C. H. et al. ClusterCAD: a computational platform for type I modular polyketide synthase design. *Nucleic Acids Res.* **46**, D509–D515 (2018).
66. Udwary, D. W. et al. Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc. Natl Acad. Sci. USA* **104**, 10376–10381 (2007).
67. Omura, S. et al. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc. Natl Acad. Sci. USA* **98**, 12215–12220 (2001).
68. Oliynyk, M. et al. Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat. Biotechnol.* **25**, 447–453 (2007).
69. Leao, T. et al. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus. *Proc. Natl Acad. Sci. USA* **114**, 3198–3203 (2017).
70. Ju, K.-S. et al. Discovery of phosphonic acid natural products by mining the genomes of 10,000 actinomycetes. *Proc. Natl Acad. Sci. USA* **112**, 12175–12180 (2015).
71. Shigdel, U. K. et al. Genomic discovery of an evolutionarily programmed modality for small-molecule targeting of an intractable protein surface. *Proc. Natl Acad. Sci. USA* **117**, 17195–17203 (2020).
This study describes the analysis of 135,000 actinobacterial genomes to identify new analogues of the immunosuppressant polyketide rapamycin.
72. Donia, M. S. et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
73. Mendes, R. et al. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332**, 1097–1100 (2011).
74. Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).
This paper reports the discovery of a new taxonomic group of thus far uncultivated bacteria, prominent among sponge microbiota, with a high biosynthetic capacity; it thus highlights the importance of mining the 'uncultivated majority'.
75. Owen, J. G. et al. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. *Proc. Natl Acad. Sci. USA* **110**, 11797–11802 (2013).
76. Charlop-Powers, Z. et al. Global biogeographic sampling of bacterial secondary metabolism. *eLife* **4**, e05048 (2015).
77. Brady, S. F., Chao, C. J., Handelsman, J. & Clardy, J. Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org. Lett.* **3**, 1981–1984 (2001).
78. Medema, M. H. et al. Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
79. Reddy, B. V. B., Milshteyn, A., Charlop-Powers, Z. & Brady, S. F. eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. *Chem. Biol.* **21**, 1023–1033 (2014).
80. Hover, B. M. et al. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nat. Microbiol.* **3**, 415–422 (2018).
81. Peek, J. et al. Rifamycin congeners kanglemycins are active against rifampicin-resistant bacteria via a distinct mechanism. *Nat. Commun.* **9**, 4147 (2018).
82. Trail, F. et al. Physical and transcriptional map of an aflatoxin gene cluster in *Aspergillus parasiticus* and functional disruption of a gene involved early in the aflatoxin pathway. *Appl. Environ. Microbiol.* **61**, 2665–2673 (1995).
83. Kennedy, J. et al. Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis. *Science* **284**, 1368–1372 (1999).
84. Mounaud, S. et al. Annotated genome sequence of *Aspergillus tanneri* NIH1004. *Microbiol. Resour. Announc.* **9**, e01374-19 (2020).
85. Lange, B. M. et al. Probing essential oil biosynthesis and secretion by functional evaluation of expressed sequence tags from mint glandular trichomes. *Proc. Natl Acad. Sci. USA* **97**, 2934–2939 (2000).
86. Jung, J. D. et al. Discovery of genes for ginsenoside biosynthesis by analysis of ginseng expressed sequence tags. *Plant. Cell Rep.* **22**, 224–230 (2003).
87. Teoh, K. H., Polichuk, D. R., Reed, D. W., Nowak, G. & Covello, P. S. *Artemisia annua* L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. *FEBS Lett.* **580**, 1411–1416 (2006).
88. Field, B. & Osbourn, A. E. Metabolic diversification-independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008).
This analysis provides a foundation for the study of biosynthetic gene clusters in plants, making it clear that these have evolved specifically in plants themselves.
89. Nützmann, H., Huang, A. & Osbourn, A. Plant metabolic clusters — from genetics to genomics. *N. Phytol.* **211**, 771–789 (2016).
90. Luo, X. et al. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **567**, 123–126 (2019).
91. Galanie, S., Thodey, K., Trenchard, I. J., Filsinger Interrante, M. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science* **349**, 1095–1100 (2015).
92. Brunson, J. K. et al. Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).
This article presents the first (major) genome mining effort in protists, revealing the biosynthetic pathway for domoic acid production in diatoms.
93. Kita, M. & Uemura, D. Marine huge molecules: the longest carbon chains in natural products. *Chem. Rec.* **10**, 48–52 (2010).
94. Chow, M. H., Yan, K. T. H., Bennett, M. J. & Wong, J. T. Y. Birefringence and DNA condensation of liquid crystalline chromosomes. *Eukaryot. Cell* **9**, 1577–1587 (2010).
95. Beedessee, G. et al. Integrated omics unveil the secondary metabolic landscape of a basal dinoflagellate. *BMC Biol.* **18**, 139 (2020).
96. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.15.341214 (2021).
97. Zan, J. et al. A microbial factory for defensive kahalalides in a tripartite marine symbiosis. *Science* **364**, eaaw6732 (2019).
98. Vaelli, P. M. et al. The skin microbiome facilitates adaptive tetrodotoxin production in poisonous newts. *eLife* **9**, e53898 (2020).
99. Gizzi, A. S. et al. A naturally occurring antiviral ribonucleotide encoded by the human genome. *Nature* **558**, 610–614 (2018).
This study describes the discovery of a novel host-produced antiviral specialized metabolite in humans guided by the knowledge that what turned out to be the biosynthetic gene conferred a viral-resistance phenotype.
100. Cooke, T. F. et al. Genetic mapping and biochemical basis of yellow feather pigmentation in budgerigars. *Cell* **171**, 427–439.e21 (2017).
101. Sabatini, M. et al. Biochemical characterization of the minimal domains of an iterative eukaryotic polyketide synthase. *FEBS J.* **285**, 4494–4511 (2018).
102. Torres, J. P., Lin, Z., Winter, J. M., Krug, P. J. & Schmidt, E. W. Animal biosynthesis of complex polyketides in a photosynthetic partnership. *Nat. Commun.* **11**, 2882 (2020).
This study shows that animals can produce complex polyketides, with the discovery of polypropionate compounds produced by sea slugs.
103. Cutignano, A. et al. Biosynthesis and cellular localization of functional polyketides in the gastropod mollusc *Scaphander lignarius*. *Chembiochem* **13**, 1759–1766 (2012). 1701.
104. Beran, F. et al. Novel family of terpene synthases evolved from *trans*-isoprenyl diphosphate synthases

in a flea beetle. *Proc. Natl Acad. Sci. USA* **113**, 2922–2927 (2016).
105. Safavi-Hemami, H. et al. Modulation of conotoxin structure and function is achieved through a multienzyme complex in the venom glands of cone snails. *J. Biol. Chem.* **287**, 34288–34303 (2012).
106. Roelofs, D. et al. A functional isopenicillin N synthase in an animal genome. *Mol. Biol. Evol.* **30**, 541–548 (2013).
107. Suring, W., Mariën, J., Broekman, R., van Straalen, N. M. & Roelofs, D. Biochemical pathways supporting β-lactam biosynthesis in the springtail *Folsomia candida*. *Biol. Open* **5**, 1784–1789 (2016).
108. Shou, Q. et al. A hybrid polyketide–nonribosomal peptide in nematodes that promotes larval survival. *Nat. Chem. Biol.* **12**, 770–772 (2016).
This article identifies a hybrid peptide–polyketide produced by nematodes, which promotes larval survival.
109. Izoré, T. et al. *Drosophila melanogaster* nonribosomal peptide synthetase Ebony encodes an atypical condensation domain. *Proc. Natl Acad. Sci. USA* **116**, 2913–2918 (2019).
110. Chekan, J. R. et al. Scalable biosynthesis of the seaweed neurochemical, kainic acid. *Angew. Chem. Int. Ed. Engl.* **58**, 8454–8457 (2019).
111. Medema, M. H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M. A. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. *PLoS Comput. Biol.* **10**, e1004016 (2014).
112. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
This article describes the antiSMASH pipeline, originally established in 2011, the first automated software tool to comprehensively identify BGCs in both bacterial and fungal genomes.
113. Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* **45**, W49–W54 (2017).
114. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
115. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
116. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
117. van Heel, A. J. et al. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.* **46**, W278–W281 (2018).
118. Tietz, J. I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
119. Santos-Aberturas, J. et al. Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool. *Nucleic Acids Res.* **47**, 4624–4637 (2019).
120. de Los Santos, E. L. C. NeuRiPP: neural network identification of RiPP precursor peptides. *Sci. Rep.* **9**, 13406 (2019).
121. Merwin, N. J. et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl Acad. Sci. USA* **117**, 371–380 (2020).
122. Kloosterman, A. M. et al. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS Biol.* **18**, e3001026 (2020).
123. Palaniappan, K. et al. IMG-ABC v.5.0: an update to the IMG/Atlas of biosynthetic gene clusters knowledgebase. *Nucleic Acids Res.* **48**, D422–D430 (2020).
124. Blin, K. et al. The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **47**, D625–D630 (2019).
125. Schläpfer, P. et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant. Physiol.* **173**, 2041–2059 (2017).
126. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63 (2017).
127. Navarro-Muñoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
This article presents a comprehensive pipeline for natural product genome mining across large numbers of genomes, including sequence similarity networking, gene cluster family assignment and multilocus phylogenetic analysis of related gene clusters.
128. Liu, Z. et al. Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *N. Phytol.* **227**, 1109–1123 (2020).
129. Schorn, M. A. et al. A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).
130. Duncan, K. R. et al. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from *Salinispora* species. *Chem. Biol.* **22**, 460–471 (2015).
131. Goering, A. W. et al. Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent. Sci.* **2**, 99–108 (2016).
132. Doroghazi, J. R. et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
This article is the first comprehensive example of metabologenomic pattern-based genome mining to correlate the presence/absence of metabolites to the presence/absence of BGCs across large numbers of strains.
133. van der Hooft, J. J. J. et al. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).
134. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).
135. Eldjárn, G. H. et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.12.148205 (2020).
136. Kersten, R. D. et al. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc. Natl Acad. Sci. USA* **110**, E4407–E4416 (2013).
137. Medema, M. H. et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* **10**, e1003822 (2014).
138. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
139. Cao, L. et al. MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. *Cell Syst.* **9**, 600–608.e4 (2019).
140. Vogt, E. & Künzler, M. Discovery of novel fungal RiPP biosynthetic pathways and their application for the development of peptide therapeutics. *Appl. Microbiol. Biotechnol.* **103**, 5567–5581 (2019).
141. Dejong, C. A. et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* **12**, 1007–1014 (2016).
This article uses a retrobiosynthetic approach to break down metabolites into their constituent building blocks and match these strings of building blocks to gene clusters using substrate specificity predictions of the encoded enzyme sequences.
142. Luo, D. et al. Oxidation and cyclization of casbene in the biosynthesis of *Euphorbia* factors from mature seeds of *Euphorbia lathyris* L. *Proc. Natl Acad. Sci. USA* **113**, E5082–E5089 (2016).
143. Jeon, J. E. et al. A pathogen-responsive gene cluster for highly modified fatty acids in tomato. *Cell* **180**, 176–187.e19 (2020).
This paper arguably represents the most comprehensive single co-expression data set used thus far for genome mining of a novel plant biosynthetic pathway.
144. Itkin, M. et al. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**, 175–179 (2013).
145. Lau, W. & Sattely, E. S. Six enzymes from mayapple that complete the biosynthetic pathway to the etoposide aglycone. *Science* **349**, 1224–1228 (2015).
146. Rajniak, J., Barco, B., Clay, N. K. & Sattely, E. S. A new cyanogenic metabolite in Arabidopsis required for inducible pathogen defence. *Nature* **525**, 376–379 (2015).
147. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, 1090 (2018).
148. Wisecaver, J. H. et al. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant. Cell* **29**, 944–959 (2017).
This study introduces a powerful co-expression-based pathway discovery method, using mutual ranks and clustering to identify co-expression modules.
149. Becher, P. G. et al. Developmentally regulated volatiles geosmin and 2-methylisoborneol attract a soil arthropod to *Streptomyces* bacteria promoting spore dispersal. *Nat. Microbiol.* **5**, 821–829 (2020).
150. Muhlemann, J. K., Younts, T. L. B. & Muday, G. K. Flavonols control pollen tube growth and integrity by regulating ROS homeostasis during high-temperature stress. *Proc. Natl Acad. Sci. USA* **115**, E11188–E11197 (2018).
151. Bruns, H. et al. Function-related replacement of bacterial siderophore pathways. *ISME J.* **12**, 320–329 (2018).
152. Rajniak, J. et al. Biosynthesis of redox-active metabolites in response to iron deficiency in plants. *Nat. Chem. Biol.* **14**, 442–450 (2018).
153. Crits-Christoph, A., Bhattacharya, N., Olm, M. R., Song, Y. S. & Banfield, J. F. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *Genome Res.* **31**, 239–250 (2021).
154. Yeh, H.-H. et al. Resistance gene-guided genome mining: serial promoter exchanges in *Aspergillus nidulans* reveal the biosynthetic pathway for fellutamide B, a proteasome inhibitor. *ACS Chem. Biol.* **11**, 2275–2284 (2016).
155. Panter, F., Krug, D., Baumann, S. & Müller, R. Self-resistance guided genome mining uncovers new topoisomerase inhibitors from myxobacteria. *Chem. Sci.* **9**, 4898–4908 (2018).
156. Yan, Y. et al. Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature* **559**, 415–418 (2018).
This article is a prime example of target-based genome mining, leading to the discovery of a novel herbicide from fungi.
157. 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).
158. Nonejuie, P. et al. Application of bacterial cytological profiling to crude natural product extracts reveals the antibacterial arsenal of *Bacillus subtilis*. *J. Antibiot.* **69**, 353–361 (2016).
159. Kurita, K. L., Glassey, E. & Linington, R. G. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. *Proc. Natl Acad. Sci. USA* **112**, 11999–12004 (2015).
160. Aliper, A. et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* **13**, 2524–2530 (2016).
161. Shang, Y. et al. Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084–1088 (2014).
This article uses genome-wide association studies to identify the BGC for cucurbitacin in cucumber, which is responsible for a characteristic bitter taste.
162. 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).
This paper identifies a new clade of uncultivated microbes as potent natural product producers, and introduces metatranscriptomics-based co-expression analysis to predict likely functions for some of their BGCs.
163. Oyserman, B. O., Medema, M. H. & Raaijmakers, J. M. Road MAPs to engineer host microbiomes. *Curr. Opin. Microbiol.* **43**, 46–54 (2018).
164. Huang, A. C. et al. A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. *Science* **364**, eaau6389 (2019).
165. Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R. & Medema, M. H. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **33**, 3202–3210 (2017).
166. Helfrich, E. J. N. et al. Automated structure prediction of *trans*-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).
167. Agrawal, P. & Mohanty, D. A machine-learning-based method for prediction of macrocyclization patterns of

polyketides and nonribosomal peptides. *Bioinformatics* **37**, 603–611 (2020).
168. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
169. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci. USA* **113**, 13738–13743 (2016).
170. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).
171. Reker, D. et al. Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **6**, 1072–1078 (2014).
This paper introduces computational methods to predict macromolecular targets for natural products, by comparing fragments of a query metabolite with those found in a training set of metabolites with known targets.
172. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **181**, 475–483 (2020).
173. Skinnider, M. A. et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
174. Karim, A. S. et al. In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat. Chem. Biol.* **16**, 912–919 (2020).
175. Zhang, J. J., Tang, X. & Moore, B. S. Genetic platforms for heterologous expression of microbial natural products. *Nat. Prod. Rep.* **36**, 1313–1332 (2019).
176. Huo, L. et al. Heterologous expression of bacterial natural product biosynthetic pathways. *Nat. Prod. Rep.* **36**, 1412–1436 (2019).
177. Lin, Z., Nielsen, J. & Liu, Z. Bioprospecting through cloning of whole natural product biosynthetic gene clusters. *Front. Bioeng. Biotechnol.* **8**, 526 (2020).
178. Lee, N. C. O., Larionov, V. & Kouprina, N. Highly efficient CRISPR/Cas9-mediated TAR cloning of genes and chromosomal loci from complex genomes in yeast. *Nucleic Acids Res.* **43**, e55 (2015).
179. Yamanaka, K. et al. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc. Natl Acad. Sci. USA* **111**, 1957–1962 (2014).
180. Fu, J. et al. Full-length RecE enhances linear–linear homologous recombination and facilitates direct cloning for bioprospecting. *Nat. Biotechnol.* **30**, 440–446 (2012).
181. Enghiad, B. & Zhao, H. Programmable DNA-guided artificial restriction enzymes. *ACS Synth. Biol.* **6**, 752–757 (2017).
182. Enghiad, B. et al. Cas12a-assisted precise targeted cloning using in vivo Cre-lox recombination. *Nat. Commun.* **12**, 1171 (2021).
183. Shapland, E. B. et al. Low-cost, high-throughput sequencing of DNA assemblies using a highly multiplexed Nextera process. *ACS Synth. Biol.* **4**, 860–866 (2015).
184. Zhang, J. J., Tang, X., Zhang, M., Nguyen, D. & Moore, B. S. Broad-host-range expression reveals native and host regulatory elements that influence heterologous antibiotic production in Gram-negative bacteria. *mBio* **8**, e01291-17 (2017).
185. Wang, G. et al. CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nat. Microbiol.* **4**, 2498–2510 (2019).
186. Harvey, C. J. B. et al. HEx: a heterologous expression platform for the discovery of fungal natural products. *Sci. Adv.* **4**, eaar5459 (2018).
This paper introduces a streamlined and largely automated workflow for genome mining and gene synthesis-based expression of fungal BGCs; the authors tested 41 different fungal BGCs and detected metabolites for 22 of them.
187. Casini, A. et al. A pressure test to make 10 molecules in 90 days: external evaluation of methods to engineer biology. *J. Am. Chem. Soc.* **140**, 4302–4316 (2018).
188. Smanski, M. J. et al. Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol.* **32**, 1241–1249 (2014).
189. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. *Escherichia coli* 'Marionette' strains with 12 highly optimized small-molecule sensors. *Nat. Chem. Biol.* **15**, 196–204 (2019).
Together with Smanski et al. (2014), this paper describes innovative methods to fine-tune the expression stoichiometry of synthetically refactored gene clusters (using either combinatorialization or sensor-based control of gene expression), in order to attain functional expression and production of the actual end compound of a pathway of interest.
190. Proctor, R. H., Hohn, T. M. & McCormick, S. P. Restoration of wild-type virulence to Tri5 disruption mutants of *Gibberella zeae* via gene reversion and mutant complementation. *Microbiology* **143**, 2583–2591 (1997).
191. Rubin, B. E. et al. Targeted genome editing of bacteria within microbial communities. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.17.209189 (2020).
192. Lam, K. N. et al. Phage-delivered CRISPR–Cas9 for strain-specific depletion and genomic deletions in the gut microbiota. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.09.193847 (2020).
193. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **3**, 319–327 (2018).
194. Reher, R. et al. A convolutional neural network-based approach for the rapid annotation of molecularly diverse natural products. *J. Am. Chem. Soc.* **142**, 4114–4120 (2020).
195. Burns, D. C., Mazzola, E. P. & Reynolds, W. F. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. *Nat. Prod. Rep.* **36**, 919–933 (2019).
196. Inokuma, Y. et al. X-ray analysis on the nanogram to microgram scale using porous complexes. *Nature* **495**, 461–466 (2013).
197. 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 (2020).
198. Chu, J. et al. Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat. Chem. Biol.* **12**, 1004–1006 (2016).
199. Chu, J., Vila-Farres, X. & Brady, S. F. Bioactive synthetic-bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome. *J. Am. Chem. Soc.* **141**, 15737–15741 (2019).
200. Chu, J. et al. Synthetic-nioinformatic natural product antibiotics with diverse modes of action. *J. Am. Chem. Soc.* **142**, 14158–14168 (2020).
201. Hudson, G. A., Hooper, A. R., DiCaprio, A. J., Sarlah, D. & Mitchell, D. A. Structure prediction and synthesis of pyridine-based macrocyclic peptide natural products. *Org. Lett.* **23**, 253–256 (2021).
202. Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol. Lett.* **187**, 111–114 (2000).
203. Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinform.* **20**, 1103–1113 (2019).
204. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
205. Kjærbølling, I., Vesth, T. & Andersen, M. R. Resistance gene-directed genome mining of 50 species. *mSystems* https://doi.org/10.1101/457903 (2019).
206. Zallot, R., Oberg, N. & Gerlt, J. A. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry* **58**, 4169–4182 (2019).
207. Usadel, B. et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant. Cell Env.* **32**, 1633–1651 (2009).
208. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink, W. Learning from co-expression networks: possibilities and challenges. *Front. Plant. Sci.* **7**, 444 (2016).
209. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
210. Tzfadia, O. et al. CoExpNetViz: comparative co-expression networks construction and visualization tool. *Front. Plant. Sci.* **6**, 1194 (2015).
211. Gubbens, J. et al. Natural product proteomining, a quantitative proteomics platform, allows rapid discovery of biosynthetic gene clusters for different classes of natural products. *Chem. Biol.* **21**, 707–718 (2014).
212. Ding, Y. et al. Genetic elucidation of interconnected antibiotic pathways mediating maize innate immunity. *Nat. Plants* **6**, 1375–1388 (2020).
213. Levin, B. J. et al. A prominent glycyl radical enzyme in human gut microbiomes metabolizes 4-hydroxy-l-proline. *Science* **355**, eaai8386 (2017).
214. Soldatou, S., Eldjarn, G. H., Huerta-Uribe, A., Rogers, S. & Duncan, K. R. Linking biosynthetic and chemical space to accelerate microbial secondary metabolite discovery. *FEMS Microbiol. Lett.* **366**, fnz142 (2019).
215. Erbilgin, O. et al. MAGI: a method for metabolite annotation and gene integration. *ACS Chem. Biol.* **14**, 704–714 (2019).
216. Pascal Andreu, V. et al. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.14.422671 (2020).
217. Kersten, R. D. et al. Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in *Salinispora tropica*. *Chembiochem* **14**, 955–962 (2013).
218. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. *Nat. Prod. Rep.* **33**, 73–86 (2016).
219. Ricart, E. et al. rBAN: retro-biosynthetic analysis of nonribosomal peptides. *J. Cheminform.* **11**, 13 (2019).
220. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software tools and approaches for compound identification of LC–MS/MS data in metabolomics. *Metabolites* **8**, 31 (2018).
221. Lo, H.-C. et al. Two separate gene clusters encode the biosynthetic pathway for the meroterpenoids austinol and dehydroaustinol in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **134**, 4709–4720 (2012).
222. Sanchez, J. F. et al. Genome-based deletion analysis reveals the prenyl xanthone biosynthesis pathway in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **133**, 4010–4017 (2011).
223. Andersen, M. R. et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc. Natl Acad. Sci. USA* **110**, E99–E107 (2013).
224. Huang, A. C. et al. Unearthing a sesterterpene biosynthetic repertoire in the Brassicaceae through genome mining reveals convergent evolution. *Proc. Natl Acad. Sci. USA* **114**, E6005–E6014 (2017).
225. Shoguchi, E. et al. A new dinoflagellate genome illuminates a conserved gene cluster involved in sunscreen biosynthesis. *Genome Biol. Evol.* **13**, evaa235 (2021).
226. Zhao, T. & Schranz, M. E. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proc. Natl Acad. Sci. USA* **116**, 2165–2174 (2019).
227. Bok, J. W. et al. Chromatin-level regulation of biosynthetic gene clusters. *Nat. Chem. Biol.* **5**, 462–464 (2009).
228. Yu, N. et al. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.* **44**, 2255–2265 (2016).
229. Lawrence, J. G. & Roth, J. R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860 (1996).
230. Ballouz, S., Francis, A. R., Lan, R. & Tanaka, M. M. Conditions for the evolution of gene clusters in bacterial genomes. *PLoS Comput. Biol.* **6**, e1000672 (2010).
231. McGary, K. L., Slot, J. C. & Rokas, A. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proc. Natl Acad. Sci. USA* **110**, 11481–11486 (2013).
232. Field, B. et al. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl Acad. Sci. USA* **108**, 16116–16121 (2011).
233. Gluck-Thaler, E. & Slot, J. C. Specialized plant biochemistry drives gene clustering in fungi. *ISME J.* **12**, 1694–1705 (2018).
234. Schorn, M. A. et al. Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology* **162**, 2075–2086 (2016).
235. van Santen, J. A. et al. The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
236. Skinnider, M. A. & Magarvey, N. A. Statistical reanalysis of natural products reveals increasing chemical diversity. *Proc. Natl Acad. Sci. USA* **114**, E6271–E6272 (2017).
237. Thrash, J. C. Culturing the uncultured: risk versus reward. *mSystems* **4**, e00130–19 (2019).

238. Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., International Natural Product Sciences Taskforce & Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* **20**, 200–216 (2021).

Acknowledgements
This work was supported by the US National Institutes of Health (NIH) (F32-GM129960 to T.dR. and R01-GM085770 to B.S.M.) and European Research Council Starting Grant 948770-DECIPHER (to M.H.M.). The authors thank members of the Moore and Medema laboratories for helpful discussions.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. The other authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks C. Gruber and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.