Natural Product Reports

The confluence of Big Data and evolutionary genome mining for the discovery of Natural Products

Journal:	*Natural Product Reports*
Manuscript ID	NP-REV-03-2021-000013.R2
Article Type:	Review Article
Date Submitted by the Author:	21-Jul-2021
Complete List of Authors:	Chevrette, Marc; University of Wisconsin-Madison, Wisconsin Institute for Discovery and Department of Plant Pathology Gavrillidou, Athina; Eberhard Karls Universitat Tubingen, Interfakultäres Institut für Mikrobiologie und Infektionsmedizin Mantri, Shrikant; Eberhard Karls Universitat Tubingen, Interfakultäres Institut für Mikrobiologie und Infektionsmedizin Sélem-Mojica, Nelly; Cinvestav-IPN, Evolution of Metabolic Diversity Laboratory, National Laboratory of Genomics for Biodiversity (Langebio) Ziemert, Nadine; Eberhard Karls Universitat Tubingen, Interfakultäres Institut für Mikrobiologie und Infektionsmedizin; Deutsches Zentrum für Infektionsforschung eV, partnersite Tuebingen Barona-Gomez, Francisco; Cinvestav-IPN, Evolution of Metabolic Diversity Laboratory National Laboratory of Genomics for Biodiversity (Langebio)

The confluence of Big Data and evolutionary genome mining for the discovery of Natural Products

Marc G Chevrette[1,*], Athina Gavrilidou[2,3,*], Shrikant Mantri[2,3,4,*], Nelly Selem-Mojica[5,6,†], Nadine Ziemert[2,3,†], Francisco Barona-Gomez[5,†]

[1] Wisconsin Institute for Discovery, Department of Plant Pathology, University of Wisconsin-Madison, Madison, WI, USA.

[2] Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Germany.

[3] German Centre for Infection Research (DZIF), Partner Site Tübingen, Germany.

[4] Computational Biology Laboratory, National Agri-Food Biotechnology Institute (NABI), Mohali, Punjab, India

[5] Laboratorio de Evolución de la Diversidad Metabólica, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, Guanajuato, México.

[*] authors contributed equally (MGC, AG, SM)

[†] corresponding authors (NSM, NZ, FBG)

[6] Current address: Centro de Ciencias Matemáticas (CCM), UNAM, Morelia, Mexico

This review covers literature between 2003-2021

Abstract

The development and application of genome mining tools has given rise to ever-growing genetic and chemical databases and propelled natural products research into the modern age of Big Data. Likewise, an explosion of evolutionary studies has unveiled genetic patterns of natural products biosynthesis and function that support Darwin's theory of natural selection and other theories of adaptation and diversification. In this review, we aim to highlight how Big Data and evolutionary thinking converge in the study of natural products, and how this has led to an emerging sub-discipline of evolutionary genome mining of natural products. First, we outline general principles to best utilize Big Data in natural products research, addressing key considerations needed to provide evolutionary context. We then highlight successful examples where Big Data and evolutionary analyses have been combined to provide bioinformatic resources and tools for the discovery of novel natural products and their biosynthetic enzymes. Rather than an exhaustive list of evolution-driven discoveries, we highlight examples where Big Data and evolutionary thinking have been embraced for the evolutionary genome mining of natural products. After reviewing the nascent history of this sub-discipline, we discuss the challenges and opportunities of genomic and metabolomic tools with evolutionary foundations and/or implications and provide a future outlook for this emerging and exciting field of natural product research.

1. Introduction

Evolution is a process; therefore, evolutionary theory seeks to describe the series of events that have allowed life to appear, develop, and diversify. Natural selection, postulated by Charles Darwin more than one hundred and fifty years ago, is perhaps the most recognized of these theories, linking the natural histories of all living forms to their reproductive fitness[1]. In the years since Darwin, we have come to appreciate that evolutionary processes display enormous complexity and act through both selective and neutral forces of varying physicochemical, ecological, temporal, and population-level constraints[2]. Neutral, non-adaptive evolution was once thought to be discordant with Darwinian evolution; now we appreciate that evolutionary histories provide evidence of both selective pressures and neutral events[3,4] . Founder effects, genetic drift, gene flow, and many other neutral mechanisms shape the genetic variation within populations upon which natural selection operates[5]. The enzymes of natural product (NP) biosynthesis are encoded in genomic information, and as such do not escape these forces of evolution [6,7]. This distinction is as important to recognize, as it is easy to neglect: NPs with antagonistic functions, like antibiotics or other biocides, are typically assumed to be under positive selection to maintain the interactions with their molecular target(s) necessary to retain function. Paradoxically, the historical use of the term ‘secondary metabolism’, synonymous with trivial or unimportant metabolism, at the same time, suggests neutral evolution, free to drift from one structure to the next. This conundrum highlights the importance of better defining evolutionary principles during chemical and biological investigation of natural products.

In this review, we aim at providing basic evolutionary principles as they have been embraced by genome miners interested in natural products-based drug discovery and the development of bioinformatics tools useful for this purpose. We discussed the origins of this sub-discipline (sub-section 1.1), as well as working definitions and core evolutionary and Big Data principles, both generally and specifically regarding evolution-driven genome mining approaches (sub-sections 2.1 and 2.2). We distinguish and highlight selected examples in which the confluence of Big Data and evolutionary genome mining for the discovery of natural products is more evident; and provide information to better understand and efficiently use these tools, but also to prompt newcomers and pave the way for the development of tools embracing the predictive power of the theory of evolution and the wealth of Big Data. Both databases and algorithms with relevant evolutionary features are presented in sub-sections 2.3 and 2.4. Selected examples of NPs research embracing evolutionary thinking - from enzymes to whole microbiomes - are provided in sub-sections 3.1 and 3.2. The selected cases highlight evolutionary thinking and include the few examples that involve tools of what we call evolutionary genome mining of natural products. The final sub-section 4 provides future directions for the development of this emerging sub-discipline as an important area of research to better understand NPs as whole and direct their biotechnological exploitation.

1.1 Origins of evolutionary genome mining of natural products

Advances in DNA sequencing have allowed for the study of allelic variation and how it relates to different phenotypes and evolutionary pressures[8]. These genetic investigations have developed into entire fields of molecular and genome evolution research, most notably advancing the areas of population genetics and phylogenetics. Population genetics investigates the frequencies and dynamics of genetic differences in and across populations, aiming to understand how some gene variants are more or less frequent than others[5]. In contrast, phylogenetics seeks to relate gene variants to each other by inferring an evolutionary history that explains differences between both genes and species[9]. Indeed, one might argue that phylogenetics was the first molecular biology Big Data method used broadly in biology, and remains so, as it aims to unveil hidden patterns otherwise ambiguous using empirical knowledge alone [10]. These inferences can be used to predict evolutionary histories through building networks of relatedness (e.g. phylogenetic trees) and reconstructing ancestral states, and therefore, in order to adopt evolutionary theory properly, these frameworks should be considered when approaching the evolution of NPs, especially when mining large datasets.

While evolutionary frameworks increasingly appear in the study of NPs, the extreme interdisciplinarity of NP research has led to adoption of evolutionary principles at different rates in different subdisciplines, depending on scientific goals and availability of data and the technologies used for their generation and analysis. For example, NP chemists often focus on empirical and mechanistic data to direct future investigations, and by doing so, they reinforce working models of biosynthetic logic in well-studied enzymes, for instance, nonribosomal peptide synthetases (NRPS)[11] and polyketide synthases (PKS)[12]. In contrast, phylogenetics, whether at the species, gene, or genome level, aims to unveil broader patterns and place them into evolutionary context. This is increasingly done for bacterial[13–15], fungal[16,17] and plant[18,19] NP biosynthetic enzymes, and even across different taxonomic lineages that produce similar NPs[20,21]. Phylogenetic insights may have limited mechanistic value, but they can assist in posing novel mechanistic hypotheses that can be experimentally tested. The combination of both approaches is embraced by Dean and Thornton's functional synthesis, which proposes that sequence analyses should be coupled with empirical, molecular experiments to retrace the evolutionary histories of biochemical processes and their phenotypes[22].

In recent years, these two apparently disparate schools of thoughts have converged, yielding new protein evolution theory[23,24] and NP genome-mining applications [25–27]. Indeed, the marriage of phylogenies and mechanistic insights, implicit in early protein evolution-rate studies[28], is the essence of evolutionary genome mining of NPs. The genes involved in NP biosynthesis and function, a subset of which have been validated through mechanistic studies, can be used to reconstruct large-scale phylogenies of multiple genes and their proteins. The genetic patterns uncovered by this Big Data approach can then feed back into more mechanistic predictions, providing hypotheses to further validate via new empirical, mechanistic studies. As these patterns can be affected by both evolutionary forces and genetic mechanisms underlying them (in bacteria [6,7], fungi[29–31] and plants[32,33] alike, yet each with their own intricacies) it is of utmost importance that these are clearly defined and appreciated by the natural products community when describing NP evolution.

2. Big Data and evolutionary genome mining of natura products: from key concepts to databases and algorithms

Genomic assemblies from DNA sequencing data and a strain's associated phenotypic and/or meta information are the source of Big Data needed for the development of NP evolutionary genome mining databases (training sets) and algorithms (tools). This stems from the fact that the interactions between the chemical products of natural product biosynthesis and their molecular targets are shaped by evolutionary processes that control chemical structure, regulation, and/or availability[6]. Thus, the enzymes that assemble natural products are subject to these evolutionary pressures as well[6,34]. Biosynthesis of natural products is typically a series of incorporating building blocks into a larger structure and adding stepwise chemical modifications. Precursors may be sourced from other parts of metabolism, the environment, or synthesized within the biosynthetic gene cluster itself [6,27]. Some biosynthesis belong to large macromolecular machinery, like NRPSs[11] or PKSs[12], while others are single domain enzymes[34]. BGCs can be as simple as a few genes or as complex as many dozens of genes whose encoded enzymes work in concert to produce the final product(s). The enzymes at work within natural product biosynthesis are as diverse and varied as the chemical structures they biosynthesize, the molecular targets with which they engage, and the interactions within and between species that they mediate. Taking this context into account, we next define evolutionary and Big Data key concepts as the foundations of evolutionary genome mining of natural products databases and algorithms.

2.1. Key Big Data concepts in Natural Products research

Big Data refers to datasets that fit four major criteria: volume, velocity, variety, and validation. First, volume: Big Data must be big [35]. This typically refers to having many different entries or examples or replicates, depending on your data type. The distinction between "normal" datasets and Big Data is an ever-changing definition: what is considered Big Data today will likely not be Big Data in the future. This is mainly due to scientific breakthroughs leading to technological improvements and data generation. Second, velocity: Big Data grows quickly, which is mainly prompted by technological advances. A useful example of volume and velocity is shown in **Figure 1**, highlighting the growth (volume) of genomes in NCBI over time (velocity). Third, variety: Big Data typically has several layers of information, which will be discussed below specifically for NP research. Finally, validation: a Big Data approach is only as good as its training data, so ensuring that training information is verified in some way is necessary for confidence in making forward predictions and identifying patterns. While validation is not strictly required for a dataset to be considered "Big", applications will have limited value if they are based on unverified information. This may sound fairly obvious yet is something that needs to be explicitly stated. Gene annotations are a common example where validation becomes very important: comparing your gene of interest to a validated dataset (e.g. UniProt, SwissProt) yields classifications that are much higher confidence than if you were to compare to unvalidated datasets (e.g. NCBI-NR) where the annotations of the dataset itself are unvalidated and errors can compound [36].

As datasets grow bigger (volume) at faster rates (velocity), an unvalidated dataset made up only of predictions may have misannotations. These errors can lead to many more subsequent misannotations, which themselves can further exacerbate these errors [37]. Thus, understanding the level of validation for your dataset is necessary to properly interpret your results. Together, these four Vs present analysis challenges, as Big Data is often too large or complex such that non-traditional or parallel computing tools are needed for analysis with ad hoc algorithms [38,39]. In general, for a natural products researcher in the early 2020s, data becomes 'Big Data' when it is too large or too complex to do simple statistics in spreadsheet-based software (e.g. Microsoft Excel). These data, moreover, are hard to process and visualize with available tools within tolerable computing times.

Standard genome mining approaches to uncover NP biosynthesis have been used to explore a wide range of taxa and environments, identifying "microbial dark matter" as a promising source of hidden chemical treasures. In evolutionary genome mining of NPs this becomes an essential consideration with potentially confounding factors. As shown in **Figure 1**, the first two 'Vs', volume and velocity, are currently covered by the sequence data in large databases. In NP research, however, data is not limited to genetics, but it has many other layers, including chemical, gene expression, ecological, and evolutionary data. For instance, the MIBiG[40] data repository is a good example of 'variety', in that it includes multifaceted chemical and genetic data. It also has a high standard of validation, as the level of validation is listed for each entry. These advantages come at the cost of volume and velocity: keeping the standards of variety and validation high mean that this repository grows at slower rates than for example the NCBI genome database. Important to evolutionary genome mining, MIBiG and other repositories tend to be biased towards a limited number of taxa that have been investigated in great detail, like species of the genus *Aspergillus* in fungi[17,29] or *Streptomyces*[41–45] within the Actinobacteria. While a bias towards this bacterial genus clearly exists, this issue is slowly decreasing with other genera such as *Nocardia*[46], *Amycolatopsis*[13], *Salinispora*[47], *Micromonospora*[48], *Pseudonocardia*[49], *Rhodococcus,*[50,51] etc. emerging as promising NP producers. Yet, bias in sampling remains a critical consideration in evolutionary studies as they can confound results and sometimes lead to erroneous conclusions, as argued recently in the case of *Aspergillus*[29].

In summary, Big Data available for evolutionary studies and genome mining of natural products come from several sources, including both broad and specialized chemical and genetic databases (see **Tables 1 and 2**). As an example, NCBI database contains over 1.4 million bacterial and over 38 thousand archaeal samples at the writing of this manuscript, with data existing as either genomes, transcriptomes, or metagenomes. These data however are far from being informative into NP research unless they are organized and/or translated into other forms or layers of information and analyzed with suitable tools. Based on our own experience, Big Data for natural products research today implies algorithms fast enough to conveniently analyze the genomes and/or metabolomes of over 30 thousand strains or samples. These numbers will rapidly multiply in the future, and thus it is critical to continually reassess "natural classifications" seen in evolutionary relationships, keeping in mind that sampling bias of training data remains a fundamental, yet often overlooked, issue. Scalability of tools is also a consideration. For example,

multiple sequence alignments and phylogenies of hundreds or thousands of genes was once considered Big Data, and remains so, yet now we can perform phylogenomic comparisons across entire kingdoms of life on an inexpensive laptop computer or free public web server [25,27]. This scalability of datasets and analysis tools can provide the genetic context necessary to perform evolutionary genome mining.

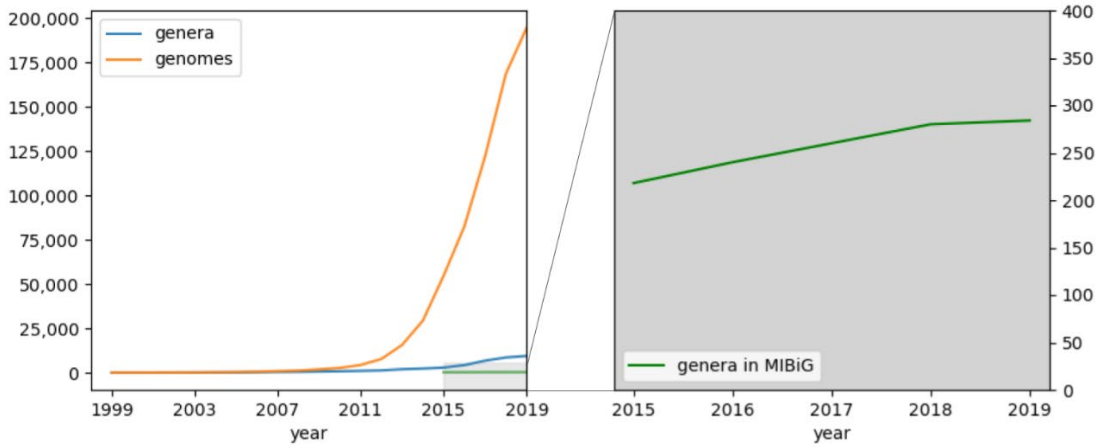

Figure 1. Growth of the number of NCBI genomes (bacteria and archaea) and genera per year from 1999 to 2019. Data from GTDB (release 95). Inset: number of genera represented by data in MIBiG.

2.2. Key evolutionary concepts in Natural Products research

Evolutionary pressures that drive the appearance and that overall shape the physicochemical and biomolecular features of natural products biosynthesis, can be incredibly dynamic and complex. Nevertheless, overarching principles of evolution of NP enzymes and/or pathways emerge. Just as biochemical principles (e.g. adenylation (A) domain specificity of NRPSs or chain elongation during PKS-catalyzed synthesis) are mechanistically fundamental for the understanding of NP biosynthesis, the following broad evolutionary principles, with a mechanistic bearing, can be considered:

(i) Enzyme promiscuity drives pathway evolution through genetic expansion-and-recruitment events, providing the building blocks to assemble, shuffle, and combine NP biosynthetic pathways[52–54] .

(ii) Once enzymes (or domains) are recruited into NP biosynthesis, they tend to cluster together as multidomain megasynthases and/or biosynthetic gene clusters (BGC)[6,7,31].

These two corollaries are valid across bacteria[6,27,45,55] fungi[17,29] and plants[33,56–58] within their unique physiological, morphological, and chromosomal peculiarities. They also hold across different taxonomic lineages that share homologous NP biosynthetic enzymes[59,60]. It is starting to be widely appreciated that the phenomena from which these corollaries

derive can occur under strong positive selection, but growing evidence and theory suggests a key role for negative selection and neutral forces on BGC dynamics[6]. Once recombination events cluster enzymes together, either as multidomain enzymes or BGCs, the resulting pathways can recruit other auxiliary elements, such as regulators, domain-domain interactors, transporters, and importantly, resistance genes[34]. As these principles were comprehensively demonstrated in the last decade or so, they were exploited by researchers for the development of the four main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining[26,27], (ii) ARTS[25,61](iii) BiG-SCAPE[45] and (iv) CORASON[45]. These tools are placed into the Big Data context and discussed in further detail in sub-section 2.4.

Using phylogenetics to unveil the evolutionary patterns of NPs follows two main approaches. On the one hand, gene trees can be used to infer a gene's evolutionary history and provide evidence for past events that have led to present-day data (i.e. branches or leaves of the tree). For evolutionary genome mining, gene trees can be useful in identifying expansions (e.g. duplications) and subsequent diversification of biosynthetic genes of interest. On the other hand, species trees describe the reconstructed evolutionary history of a set of species or individuals, and thus are useful for identifying larger-scale evolutionary events[62]. Critically assessing how the topologies of genes and species agree and disagree can shed light on important evolutionary events, such as horizontal transfers[63]. While NP research is focused on BGCs (a collection of genes), much can be learned from studying single-gene and species trees. Understanding the distribution and evolution of NPs within taxa, for example, is a prerequisite for effective sampling and bioprospecting strategies.

For those interested in evolutionary genome mining of NPs, it is important to note that the above-mentioned approaches are the result of properly embracing phylogenetics and evolutionary principles, often implementing concepts and principles not typically studied by NP chemists. **Figure 2** shows the main concepts that those interested in the use and development of these tools should take into account. As mentioned, the main two evolutionary mechanisms driving the appearance of novel NP biosynthetic pathways are diversification (enzyme promiscuity and BGC dynamics) and selection (positive, negative and neutral). However, it is only when these forces combine and impact the fitness of the NP-producing organism that pathways are assembled and reassembled during the course of evolution[34]. The main genetic mechanisms driving these evolutionary events have been identified and have been used in the development of NP evolutionary genome-mining tools (thicker arrows, **Figure 2**). However, much remains to be deciphered regarding the evolution of NPs, especially in terms of their expression and function in the real environmental settings of their producing organisms, where fitness operates. Study cases are available (see sub-section 3), but their scarcity makes them anecdotal and thus more data is needed to develop mining tools based on Big Data principles to investigate this layer of complexity (thinner and/or dashed arrows, **Figure 2**).

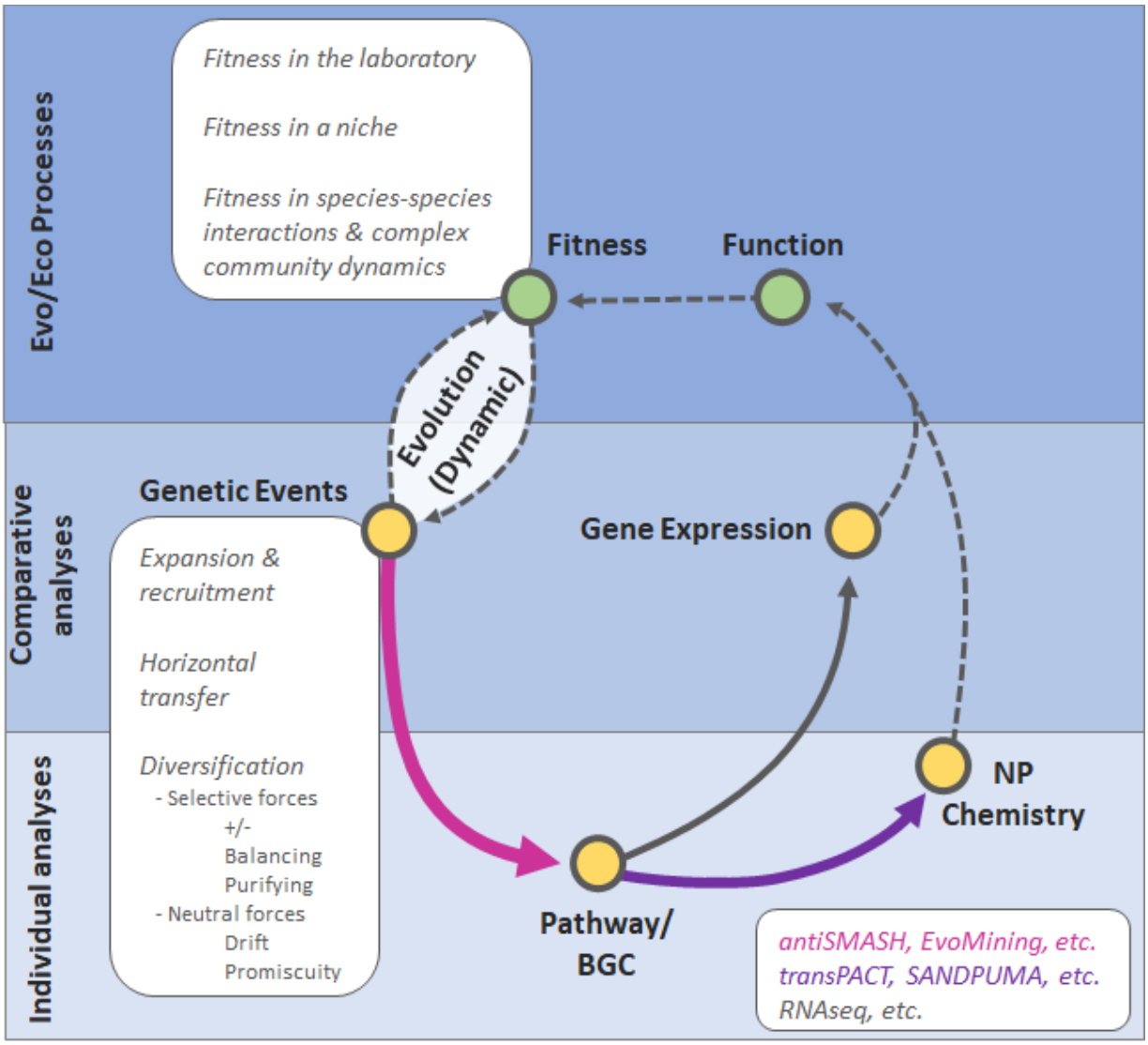

Figure 2. Evolutionary genome mining of natural products in a concept-driven framework. Studies on the evolutionary histories of NPs, their biosynthetic genes, and their producing organisms are driven by analyses at different levels of organization. Individual analyses (bottom) focus on a Pathway/BGC and their molecular product(s) or chemistry. Examples of tools that predict NP chemistry from BGCs are shown in purple. These individual data can then be contextualized with comparative analyses (middle) across many conditions or strains/species, with an emphasis in the genetic events underlying the evolution of NPs BGCs. One example is Gene Expression studies (gray, RNAseq) where comparisons of transcriptional patterns can place genes in a broader biological context. Analyses at the level of ecological and/or evolutionary processes (top) are the most challenging, and as a field we have only just begun to understand how Gene Expression, BGC, NP chemistry, and other "lower-level" data contribute to molecular function, and in turn how function contributes to an organism's fitness (linked by dotted lines to highlight that there are not yet standardized methods, but there is opportunity to develop them integrating Big Data). This remains a major challenge, as fitness is often a function of the environment. Evolution occurs as a dynamic process in which the fitness impact of a BGC's product influences the BGCs genetic components (e.g. diversification, selection, and other processes; see box). These in turn can feed back into fitness. Previously characterized genes and/or patterns of genetic events can then be used to identify and characterize BGCs de novo from genomic data (pink), either through rules-based or evolutionary methods.

2.3 Natural Products databases available for evolutionary genome mining

As mentioned, data available for investigating natural products in the Big Data era comes from several sources. However, this information only becomes useful when organized on databases (training sets) that can be coupled with metadata of the organisms themselves, but also with information about the technology and methods used to generate the data. Examples of well-executed databases include the GNPS mass spectra public database[64], the MIBiG repository with experimentally validated datasets[40,65], and the bioinformatically predicted BGCs of the antiSMASH DB [66,67] (**Tables 1 and 2**). Recently, the first evolutionary database, i.e. ActDES, which is specific for the Actinobacteria, has been reported[68]. All of these databases, despite complying with the four 'Vs' in one way or another, including variety, are useful in comparative or evolutionary studies, but not sufficient as none of them provide a comprehensive multi-layer database including or embracing evolution. In turn, at this stage, it is down to the evolutionary genome miner to select and integrate the most suitable and relevant DBs from those provided in **Tables 1 and 2**, within a phylogenomics framework. Selected DBs are highlighted throughout this review with the aim of emphasising their value in relation to the four 'Vs'.

Table 1. Genomic databases to explore natural products diversity and evolution.

Database Name *	Parameter Name	Parameter Value	Current Version (date)
MIBiG [65]	BGCs	1,923	2.0 (2019)
IMG-ABC [69]	BGCs	410,683	5.0
antiSMASH-db [67]	BGCs	147,517	3.0
BiG-FAM [70]	BGCs	1,225,071	1.0
NCBI Genome	Bacteria spp.	278,820	November 2020
	Archaea spp.	5,625	November 2020
	Eukaryote spp.	14,486	November 2020
MGnify[71]	Metagenomes	32,746	November 2020
IMG/M[72]	MAGs	52,515	November 2020
	BGCs	104,211	November 2020
CARD[73]	Alleles	213,809	February 2021
	Reference sequences	3,146	February 2021
SRA (Bacteria)	Datasets	1,466,494	November 2020
SRA (Archaea)	Datasets	38,592	November 2020
NCBI WGS (Bacteria)	Projects	941,266	December 2020
NCBI WGS (Archaea)	Projects	6,225	December 2020

Resfinder 4.0[74]	Resistance genes	2,690	December 2020
MG-RAST 4.0.3[75]	Metagenome	447,497	January 2021

* Most of the listed databases in Table 1 and Table 2 arguably satisfy the Big Data characteristics of volume and variety. Since there have been only few periodic releases for some of these databases, the velocity characteristics of Big Data can be appreciated for only a few of these. The month and year (date) of each database in Table 1 and Table 2, when last accessed, are provided. Exact dates for current versions are not provided as are not available.

Table 2. Chemical databases to explore natural products diversity and evolution.

Database Name *	Parameter Name	Parameter Value	Current Version (date)
MACADAM[76]	Metabolites	7,921	1
PubChem[77]	Compounds	111,456,896	November 2020
GNPS[64]	NP compounds Spectra	18,163 221,083	1
NP Atlas[78]	Compounds	24,594	v 2020_06
COCONUT[79]	Compounds	406,747	March 2021
StreptomeDB[80]	Compounds	4,000	2
PoDP[81]	Paired (meta)genomes and metabolomes	4,853	2021 GitHub v0.9.2
Siderophore DB	Compounds	262	June 2021
LOTUS[82]	NP compounds	276,518	February 2021

* Refer to table notes in Table 1.

2.4 Big Data and Natural Products evolutionary genome mining algorithms

Communication between evolutionary biologists, computer scientists and mathematicians has historically led to biological insight, including the developments of population genetics theory and the transition matrices that are key to common genomic search algorithms like BLAST[83]. These disciplines have successfully converged again in recent years for the development of sophisticated NP genome-mining algorithms and platforms (**Table 3**). In this subsection, we list and explain major evolutionary genome mining of NPs approaches available to date with a focus on those that directly or indirectly rely on the use of the theory of evolution in any of its forms, either within the algorithms themselves or in their visualizations. The availability of genomic data (e.g. MIBiG, CARD, antiSMASH DB, **Table 1**) is fundamental, but probably more often will also be inputs from purely chemical

DBs (**Table 2**), e.g. GNPS, Paired Omics Data Platform [PODP], which can also serve as training data in supervised algorithms. Notably, some of these genomic-based algorithms already include input from chemical databases [64,77,78]. Thus, the integration of data types, as in MIBiG or PODP, may provide training datasets with valuable links between genomic and chemical data, further embracing variety. This integration holds great promise and value to the field, but since it is only beginning to occur, it remains to be seen how regularly chemical data will be embraced by evolution-driven genome mining efforts.

Currently, evolutionary genome-mining for the discovery of novel NPs[84] aims to provide answers to two main questions, and by doing so, generate predictions: (i) which genes and/or BGCs produce metabolites not typically associated with central metabolism? and (ii) which genes or domains specific to a lineage represent innovation and diversification compared to ancestral states? As mentioned, several specialty databases (**Table 1 & 2**) are available and are used by the main evolutionary genome mining tools that the NP community has used to identify and investigate novel pathways: (i) EvoMining[26,27], (ii) ARTS[25,61] (iii) BiG-SCAPE[45] and (iv) CORASON[45]. Following a similar rationale, a conceptual framework for mining siderophore BGCs based on their transporters has recently been reported[85]. Importantly, available tools can be used independently or in combination, and go in hand with species-level phylogenetic analyses which directly integrate NP biosynthesis (e.g. AutoMLST[86]) or analyses that are part of more generalized phylogenetic pipelines[87]. The combination of the latter, i.e. a species tree, with large-scale BGC prediction and their taxonomic distribution, is BiG-SLiCE output [88]

Supervised algorithms make use of the DBs mentioned in the previous sub-section in the form of training sets with validated labels about what is an NP BGC and what is not[37]. Here, the “correct” classifications are known for training data and used to make predictions about new data. These methods typically require heavy (and often manual) curation of training sets, and thus the importance of the fourth V, validation. So far, most of NP research adopting genome mining approaches employs supervised algorithms, mainly used in classification problems that require prior knowledge[89]. Unsupervised algorithms, instead aim to extract patterns and trends from unlabeled data[90], similar to phylogenies. These can be helpful to identify data features (e.g. genes and domains) that are important for categorization, but since no “true” answer is known false-positive errors may be more frequent. Clustering or other grouping methods used in unsupervised methods attempt to give some structure to a dataset. Typically, supervised and unsupervised strategies are complementary, as it is the case in NP evolutionary genome-mining (**Figure 3**).

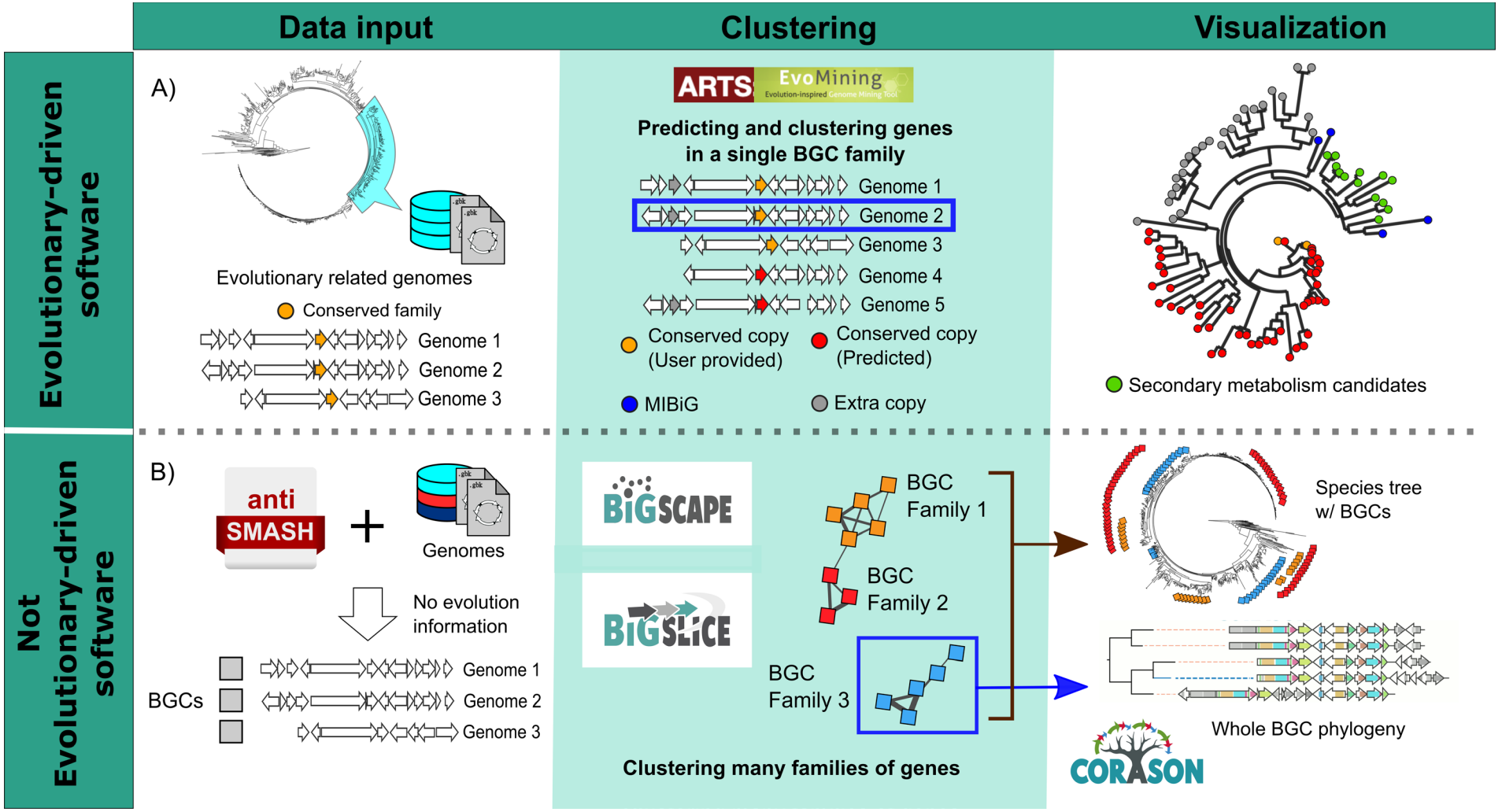
Data input
Clustering
Visualization
Evolutionary-driven software
Not Evolutionary-driven software
A)
Evolutionary related genomes
Conserved family
Genome 1
Genome 2
Genome 3
ARTS
EvoMining
Predicting and clustering genes in a single BGC family
Genome 1
Genome 2
Genome 3
Genome 4
Genome 5
Conserved copy (User provided)
Conserved copy (Predicted)
MIBiG
Extra copy
Secondary metabolism candidates
B)
antiSMASH
Genomes
No evolution information
BGCs
Genome 1
Genome 2
Genome 3
BIGSCAPE
BIGSLICE
BGC Family 1
BGC Family 2
BGC Family 3
Clustering many families of genes
Species tree w/ BGCs
Whole BGC phylogeny
CORASON

Figure 3 (*previous page*). Evolution-driven genome mining tools. **A.** Evolutionary algorithms need as inputs genomes from taxonomically related lineages, where conserved protein families (orange) are selected for further exploration (ARTS/EvoMining). Conserved (orange and red) and extra (gray) copies of these families are identified and compared by a phylogenetic distance against proteins from NP databases (blue). Finally, the tree used in the phylogenetic distance is provided as a visualization, where predictions are included (green). **B.** Algorithms with an evolutionary visualization but without evolutionary driven distances does not restrict their input genomes to be phylogenetically related. Gene clusters obtained from these algorithms are gathered in gene cluster families (GCF) by classification methods. Finally, evolutionary visualizations can be provided, either as a whole-BGC network of phylogenetic tree (BiG-SCAPE/CORASON) or as the occurrence of each GCF throughout a species tree (BiG-SLICE).

Within NP research, supervised problems are used to identify and classify domains, genes, and BGCs. ClusterFinder[85] was one of the first algorithms that attempted to classify regions of the genome as NP BGC (or not) by calculating a moving average of a "biosynthetic score", calculated based on domain- and gene-level agreement with profile Hidden Markov Models of biosynthetic enzymes. Although ClusterFinder[91] does not directly leverage evolutionary theory in its algorithm, it is indirectly inferring the evolutionary processes that shaped BGC regions throughout the genome. Many of these algorithms have been trained primarily (or exclusively) on bacterial data, and thus accurate and reliable identification of fungal BGCs remains a challenge. Fortunately, recent work has begun to take fungal-specific genes and genetic structure into account to address this issue[92–94]. A similar scenario in plants[95] has now been encountered since the realization that BGCs actually exist in this large and prominent group of NP producing organisms.

Identifying shared and novel features within and between taxonomic lineages is attempted by unsupervised algorithms, such as BiG-SCAPE, BiG-SLICE and CORASON. For example, BiG-SCAPE, and more recently BiG-SLICE, clusters BGCs into gene cluster families (GCFs) without requiring prior knowledge of these families. This is done after calculating distance scores between BGCs on the basis of shared protein families and BGC organization. After clustering, it can be useful to sort and/or connect these GCFs with each other into bigger "clans", that are related but more distantly so than members of the same GCF. This broader context can be used to track evolutionary events of related BGCs and investigate how these events are distributed across gene and/or strain phylogenies. An alternative-yet-complementary approach employed by CORASON involves phylogenetic trees of shared enzymatic features, including in some instances whole-BGCs phylogenies. Importantly, these processes use *supervised* classifications of genes and domains to perform *unsupervised* clustering into GCFs, so they too require high quality (i.e. validated, or at least carefully curated) genomic and chemical databases.

In contrast, EvoMining and ARTS, represent the first (and to our knowledge, thus far the only) heuristic algorithms that incorporate evolutionary thinking as part of the supervised approach itself, attempting to infer what is central metabolism and what may be secondary metabolism, with a certain degree of diversification hinting towards the appearance of an specialized pathway. Evolution is inferred as a distance metric, which can be seen as similar to a support vector machine algorithm[96–98], but implemented using a tree to determine appropriate groupings (and thus classifications) for biosynthetic enzymes. Put in another way, it seeks to identify which query enzymes cluster more

closely with central metabolism and which cluster more closely with secondary or specialized metabolism. Extra gene copies are assessed by EvoMining as potential recruitments into NP biosyntheses, and these gene families may differ from one taxonomic lineage to another (**Figure 3A**).

After classification into BGC families (e.g. with BiG-SLICE and/or BiG-SCAPE), further evolutionary context can be added in the visualization stage with CORASON according to the phylogenetic history of genes within the BGC or the strain-level phylogeny of the producing organism itself. In turn, CORASON identifies gene clusters in a genomes database and sorts them according to their evolutionary relationships. Tools such as MicroReact[92] can also allow for visual exploration of large phylogenetic trees annotated with metadata. EvoMining and ARTS both start with labeled sets (genes that are either the primary copy or specialized metabolism copies that belong to other databases, e.g. CARD/MIBiG) and employ supervised methods where evolutionary distance is used to classify putative BGCs. As a consequence, their predictions are intuitively displayed phylogenetically. Other software suites that perform pangenomic visualization (e.g. Anvio [99]) are also useful in that they allow identification of families with potential gene expansion and/or recruitment events. Many recent tools aim to sort and visualize relations between BGCs: for example, MultiGeneBlast[100] (implemented in antiSMASH), finds gene homologs in BGC comparisons. Given otherwise identified BGCs (e.g. by antiSMASH or other tools), BiG-SCAPE[45] can classify them into BGC families and other visualization tools such as clinker[101], FlaGs[102] and TREND[103] allow for interactive visualizations (**Figure 3B**).

Table 3. Big Data algorithms for exploring natural products diversity and evolution.

Algorithm	Validation dataset	Type of data	Method	Date
ARTS 2.0 [61]	Bacterial kingdom genomes and metagenomes	Genomes	Duplication and BGC proximity, Phylogeny and resistance screen	May 2020
BiG-SCAPE [45]	Clusters from ~3,000 genomes	BGCs	Jaccard Index plus Maximum Likelihood FastTree	November 2019
EvoMining 2.0 [27]	~100 conserved families from ~1,000 genomes	Biosynthetic genes	Duplication and gene proximity to MIBiG, Phylogeny	December 2019
BiG-SLICE [88]	BiG-FAM (1,225,071)	BGCs	Balanced Iterative Reducing and Clustering using Hierarchies	August 2020
CORASON [45]	~3,000	Genomes or BGCs (visualization)	Blast plus FastTree	November 2019
clinker [101]	NA	BGCs (visualization)	Hierarchical clustering	January 2021
FlaGs [102]	324	BGCs (visualization)	BGC's Hidden Markov Model	September 2020
TREND[103]	NA	BGCs (visualization)	Hierarchical clustering	April 2020
MicroReact [92]	NA	Trees with metadata (visualization)	libraries:Chart.js, Leaflet, Phylocanvas, React, Sigma	November 2016
Anvi'o [99]	NA	Pangenomes (visualization)	Hidden Markov Models	October 2015

3. **Genomic and enzymatic evolution of Natural Products**

3.1 Evolution of the genome of NP-producing organisms

Multiple studies have been conducted on the evolution of NP producers, providing useful indications for targeted bioprospecting. Biosynthetic potential and diversity appear to be related to the ecological niche of the producers, as was confirmed in multiple instances [14,104–112]. In some cases, though, phylogeny is more important, as observed in microbial taxa where secondary metabolism is most similar in closely related organisms rather than those isolated from the same source[110,113]. Such investigations showcase possible promising targets for NP research, be they specific known[14,113] or understudied taxa[14,51,110] or different environments/niches[104,105,107,109,111]. As such, it is clear therefore that evolution can be applied for the discovery of novel natural products, which can powerful if properly embraced.

Comparative genomic analyses have shown that most bacterial taxa harbor only a few BGCs while some dedicate a large proportion of their genomes to specialized or secondary metabolism [46,87,105,106,108–110,112,114–116].The quantity and diversity of BGC content differs among the taxa, with extreme cases reported[47,104]. How disperse the phylogenetic distribution of a BGC is, can allude to the various effects selection has had on its related pathways[117]. Most notably, horizontal gene transfer (HGT) is a relatively frequent phenomenon in BGCs, which is one likely explanation for their extended distribution across distant taxa and their observed diversity[6,13,64,105,106,108,114,118–121]. While HGT is observed frequently in BGCs compared to other genetic elements, it is important to note that the evolutionary timescales involved are still quite large[6,106,122]and depend on both population structure and genetic identity of donor and recipient[6,106,122]. Vertical inheritance of BGCs within the same lineage is the dominant means through which biosynthetic information is transferred [6,123] This is a key distinction that should be made when studying the evolution of BGCs, as the more subtle vertical evolutionary dynamics happen from generation to generation, while HGT events are typically observed at timescales closer to thousands, millions, or billions of years.

Thus far, all analyses mentioned in this subsection were not conducted on a Big Data scale. Indeed, the information discovered so far is being confirmed by multiple independent inquiries, yet still issues of small taxonomic coverage and sampling biases remain. In 2014, three articles were published that followed a more global approach to NP producer genomics. Cimermancic[91] and co-authors analyzed more than 1000 genomes from across the bacterial kingdom and created a "global map" of biosynthesis, encompassing ~33,000 predicted BGCs. Doroghazi[44] and co-authors focused on one phylum and, using different metrics and methods than Cimermancic, reached a similar conclusion by collecting information on the producers capacity and potential. At the same time, Medema[120] and co-authors examined a large number of known BGCs and proved that the rates of evolutionary events within such units are much higher than in clusters of primary metabolism. Since these studies were first published, the available data has multiplied and so too have the methods for processing them; more universal-scope analyses will soon follow and give the answers to questions that remain open, including

how and when biosynthetic diversity evolved[116] or the capacity of nature to keep providing us with new compounds[124].

The above-mentioned studies have focused on microbes that have been cultured under laboratory conditions. However, the number of unculturable organisms is vast and metagenomic analyses have begun to unravel their hidden biosynthetic potential, indicating promising new sources for NP bioprospecting (see next paragraph). Furthermore, investigating evolutionary patterns based on environmental samples can shed light on the functions of the NPs found in nature as well as their raison d'etre within their microcosm[125]. This is important as NP evolution occurs at the population level, as highlighted by recent examples where population genomics frameworks have been adopted to mine NPs in genomic data, both in fungi and bacteria[29,104,126–129]. Such approaches have even proven valuable at the bacterial colony-level of a domesticated model laboratory strain, i.e. *Streptomyces coelicolor*[130,131].

Soil metagenomic surveys in urban greenspaces, grassland meadows, and areas covering up to continent-wide scale have reported microbial diversity patterns[85,132–135] . These patterns are drastically affected by the environment and massive sequencing efforts are required to comprehensively capture their diversity, even at kilometer scale. High throughput functional studies involving creation of large-insert metagenomic libraries provides a novel approach to examine the functional and phylogenetic diversity of sampled ecosystems[136–138]. Economically attractive approaches using amplicon sequencing have been used to prove the domain-level diversity of environmental NPs. Such approaches have provided clues to answer the long standing question of which sites should be surveyed to maximize the discovery of novel natural products [64,85,109,139–142]. Massive amounts of shotgun metagenomic data are already easily available from public repositories. Analyzing these Big Data to infer significant NP patterns has now become the next bottleneck and faster algorithms and easy to use tools are badly required to mine the potential resource. Additionally, detailed documentation, standardized sampling procedures, and still more metadata are required to be incorporated into public databases in order to exploit patterns and extract useful information.

3.2. BGC and multidomain enzyme evolution

The evolutionary history of BGCs can be studied by building separate and/or concatenated trees of their genes and protein products. These can have very different topologies than the species trees of the NP producers themselves, suggesting unconventional sequence transmission events, such as Horizontal Gene Transfer (see previous section), gene conversion, intra-genomic recombination[120], and others. Together, these trees and functional information of NP genes can be used as a foundation to predict the activity of yet-unknown compounds and suggest potential links between fitness and the evolutionary forces at work.

Natural products exhibit extremely diverse chemistry. Their evolutionary complexity is no less complex. Domains evolve in the context of genes, genes in the context of BGCs, and BGCs in the context of their the producers' genomes[6,143]. Further, how these metabolites contribute to the fitness of their producing organisms depends largely on their

environmental niche, which is often completely unknown or has poorly-understood factors and boundaries[144]. Because of this interdependence between multiple levels of organization, evolution does not affect clusters uniformly[120]. Indicatively, trans-acyltransferase (trans-AT) AT domains have evolved independently from cis-AT AT domains: the latter cluster into NP-specific clades and are known to be acquired vertically, while the prior are present in many different phyla and appear to be transferred horizontally[145]. Based on the clades formed in trans-AT AT and KS trees, it appears their evolution is strongly linked to their elongation substrate specificities [106,120,145,146]. Indeed, computational pipelines such as transPACT[147] place KS sequence information into a phylogenetic framework to predict substrate specificity for unknown sequences. cis-AT and trans-AT PKS variants can produce similar metabolites even though they have distinct evolutionary histories. This case of evolution may be influenced by the modularity of Type I PKS clusters that can be more plastic due to intragenic recombinations and may allow for adaptability in a wide range of ecological niches [145]

Although much of NP evolution is thought of at the level of BGCs or genes, important evolutionary changes can also happen at even smaller scales. Substrate specificity of different NP enzymes is often dictated by the three-dimensional organization of their active sites and/or protein-protein interaction surfaces, so subtle changes to the protein sequence of these areas can steer specificity (and promiscuity) in multiple evolutionary directions. In some cases, these changes correlate with phylogeny, so knowledge of the evolutionary mechanisms behind BGCs can allow for collecting reliable information from domain phylogeny. NRPS domains also show evolutionary patterns linking phylogeny and chemistry[145]. Similar to the trans-AT KS domains of the PKS clusters, A-domains of NRPSs cluster into clades according to substrate specificity, while C-domains are highly conserved and follow a BGC-specific pattern [20,106,120]. Computational methods such as SANDPUMA[148] and others have used this phylogenetic signal to reliably predict the substrate specificity of A-domains. Recently, “Substrate level” evolutionary signals, like in trans-AT KS and NRPS A-domains, can be used to predict substrate specificity, while “pathway level” evolutionary signals, like in NRPS C-domains can be used to predict BGC-level patterns of similar molecules[47].

4. What lies ahead? Needs and opportunities for evolutionary genome mining of NPs.

Evolutionary genome mining of natural products in the Big Data era has inherited the tradition of phylogenetics, in the sense that natural history coupled with genetic and chemical observations can provide mechanistic insight. With this heritage comes the promise of discovering “The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism”, which has important implications in gene expression and the distinctions between “cryptic” and “silent” BGCs.[84] Although genomic and metabolomic speciality databases have made considerable progress, we envisage an ever-growing need for novel speciality datasets merging different layers of information. A promising current endeavor is the assemblage of metabologenomics databases, where genetic information and predictions are merged with chemical data (e.g. Paired Omics database[81]). Nevertheless, the systematic inclusion of other data types, including

evolutionary relationships, remains a challenge. One notable evolutionary database has been recently released for Actinobacteria [68], but those with larger scale and broader taxonomic coverage are much needed. These high-variety databases promise new insights in the NP field as a whole. Similarly, the accompanying algorithms needed to efficiently compute high volume datasets will allow us to perform these analyses at scale and keep pace with the technological advances that generate data at high velocity. In the near future we expect these data to go beyond only genomes, metabolomes, and metagenomes and begin to encompass ecological and functional metadata[149].

Biosynthetic enzyme domains are the focus of current, and likely future, algorithms. This presents unique challenges for enzyme families whose classifications are problematic and/or understudied in the community. For instance, chemists have provided insights into why sequence-based phylogenies are insufficient for certain enzymes: transition-state intermediaries can be highly reactive and plastic, and therefore sequence space is less constrained than in enzymes with well-defined active sites[150]. Examples of this include the terpene cyclases, cytochrome P450s, hydrolases and type III polyketide synthases, amongst others. In these examples, analyses could benefit from alternative methods to establish relationships useful to provide classification and dataset structure. In turn, this may provide more informative training sets within well-structured databases, increasing the quality of predictions surrounding these important classes of natural products biosynthetic enzymes. It should be noted that classification of some of these enzymes within abovementioned DBs, such as antiSMASH DB, does not necessarily mean that this problem has been sorted out (see validation; previous sections). Pangenomic analyses[99,151] to identify expanded enzyme families within lineages may provide an interesting possibility to classify enzyme families on evolutionary grounds.

Here, by reviewing the nascent history of evolutionary genome mining of natural products as a sub-discipline, it has become apparent that a prerequisite for the development of successful algorithms is the realization and characterization of genetic events driving the evolution of biosynthetic enzymes in their genomic context (e.g. BGCs). As such, we highlight the following evolutionary concepts with the promise to link evolution to genetic and chemical mechanisms. It has become clearer that “natural” evolution of natural products can be governed by dynamic processes that result in functional replacements. For example, in convergent evolution of chemically related scaffolds with diverse biomolecular activities[152], whose biosynthesis is directed by non-related BGCs that produce functionally similar molecules. It has also become clearer that biosynthetic pathways can be encoded by physically unrelated loci (in contrast to BGCs), which may consist of sub-clusters[153], and that the same BGC can produce diverse natural products with different biological functions in response to the environmental conditions[154]. This intragenomic cross-talk might be seen as a simplified version of the metabolic exchange between different organisms within a microbiome, for which evolutionary experimental and conceptual frameworks have been developed [155–157]. Both levels of metabolic cross-talk represent an immanent Big Data challenge: to genomically mine large datasets to correlate physically unlinked loci and propose metabolic relationships[72,109] How to best embrace evolutionary processes, many of which we are only beginning to understand, in Big Data genome mining for natural products remains an exciting yet challenging

endeavor; one that will surely provide many possibilities for the future of this emerging sub-discipline.

References

1. Sugden, A., Ash, C., Hanson, B. & Zahn, L. Happy Birthday, Mr. Darwin. *Science* **323**, 727–727 (2009).
2. Goldman, A. D. & Liberles, D. A. The Journal of Molecular Evolution Turns 50. *J. Mol. Evol.* **89**, 119–121 (2021).
3. Lynch, M. *et al.* Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
4. Wideman, J. G., Novick, A., Muñoz-Gómez, S. A. & Doolittle, W. F. Neutral evolution of cellular phenotypes. *Curr. Opin. Genet. Dev.* **58–59**, 87–94 (2019).
5. Matthew B. Hamilton. *Population Genetics, 2nd Edition* | *Wiley*. (2021).
6. Chevrette, M. G. *et al.* Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).
7. Jensen, P. R. Natural Products and the Gene Cluster Revolution. *Trends Microbiol.* **24**, 968–977 (2016).
8. Wolfe, K. H. & Li, W.-H. Molecular evolution meets the genomics revolution. *Nat. Genet.* **33**, 255–265 (2003).
9. Masatoshi Nei & Sudhir Kumar. *Molecular Evolution and Phylogenetics*. (Oxford University Press, 2000).
10. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 4576–4579 (1990).
11. Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and Prospects. *Angew. Chem. Int. Ed.* **56**, 3770–3821 (2017).

12. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. *Chem. Rev.* **119**, 12524–12547 (2019).
13. Larsen, J. S., Pearson, L. A. & Neilan, B. A. Genome Mining and Evolutionary Analysis Reveal Diverse Type III Polyketide Synthase Pathways in Cyanobacteria. *Genome Biol. Evol.* **13**, (2021).
14. Gutiérrez-García, K. *et al.* Phylogenomics of 2,4-Diacetylphloroglucinol-Producing Pseudomonas and Novel Antiglycation Endophytes from Piper auritum. *J. Nat. Prod.* **80**, 1955–1963 (2017).
15. Adamek, M. *et al.* Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in Amycolatopsis species. *BMC Genomics* **19**, 426 (2018).
16. Lind, A. L. *et al.* Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species. *PLOS Biol.* **15**, e2003583 (2017).
17. Bushley, K. E. & Turgeon, B. G. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol. Biol.* **10**, 26 (2010).
18. Piatkowski, B. T. *et al.* Phylogenomics reveals convergent evolution of red-violet coloration in land plants and the origins of the anthocyanin biosynthetic pathway. *Mol. Phylogenet. Evol.* **151**, 106904 (2020).
19. Wilson, A. E. & Tian, L. Phylogenomic analysis of UDP-dependent glycosyltransferases provides insights into the evolutionary landscape of glycosylation in plant metabolism. *Plant J.* **100**, 1273–1288 (2019).
20. Shimizu, Y., Ogata, H. & Goto, S. Type III Polyketide Synthases: Functional Classification and Phylogenomics. *ChemBioChem* **18**, 50–65 (2017).
21. Jenke-Kodama, H., Sandmann, A., Müller, R. & Dittmann, E. Evolutionary Implications of Bacterial Polyketide Synthases. *Mol. Biol. Evol.* **22**, 2027–2039 (2005).
22. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution. *Nat. Rev. Genet.* **8**, 675–688 (2007).

23. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687 (2005).
24. Pál, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nat. Rev. Genet.* **7**, 337–348 (2006).
25. Alanjary, M. *et al.* The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.* **45**, W42–W48 (2017).
26. 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).
27. Sélem-Mojica, N., Aguilar, C., Gutiérrez-García, K., Martínez-Guerrero, C. E. & Barona-Gómez, F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genomics* (2019) doi:10.1099/mgen.0.000260.
28. Alvarez-Ponce, D. Richard Dickerson, Molecular Clocks, and Rates of Protein Evolution. *J. Mol. Evol.* **89**, 122–126 (2021).
29. Rokas, A., Wisecaver, J. H. & Lind, A. L. The birth, evolution and death of metabolic gene clusters in fungi. *Nat. Rev. Microbiol.* **16**, 731–744 (2018).
30. Rokas, A., Mead, M. E., Steenwyk, J. L., Raja, H. A. & Oberlies, N. H. Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat. Prod. Rep.* **37**, 868–878 (2020).
31. Drott, M. T. *et al.* Microevolution in the pansecondary metabolome of Aspergillus flavus and its potential macroevolutionary implications for filamentous fungi. *Proc. Natl. Acad. Sci.* **118**, (2021).
32. Weng, J.-K. The evolutionary paths towards complexity: a metabolic perspective. *New Phytol.* **201**, 1141–1149 (2014).
33. Moghe, G. D. & Last, R. L. Something Old, Something New: Conserved Enzymes and the Evolution of Novelty in Plant Specialized Metabolism. *Plant Physiol.* **169**, 1512–1523

(2015).

34. Megahed, F. M. & Jones-Farmer, L. A. Statistical Perspectives on “Big Data”. in *Frontiers in Statistical Quality Control 11* (eds. Knoth, S. & Schmid, W.) 29–47 (Springer International Publishing, 2015). doi:10.1007/978-3-319-12355-4_3.
35. Barona-Gómez, F. Re-annotation of the sequence > annotation: opportunities for the functional microbiologist. *Microb. Biotechnol.* **8**, 2–4 (2015).
36. Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S. & Rubin, D. L. Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digit. Med.* **2**, 1–6 (2019).
37. Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013).
38. Jin, X., Wah, B. W., Cheng, X. & Wang, Y. Significance and Challenges of Big Data Research. *Big Data Res.* **2**, 59–64 (2015).
39. Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
40. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
41. Belknap, K. C., Park, C. J., Barth, B. M. & Andam, C. P. Genome mining of biosynthetic and chemotherapeutic gene clusters in Streptomyces bacteria. *Sci. Rep.* **10**, 2003 (2020).
42. Barka, E. A. *et al.* Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol. Mol. Biol. Rev. MMBR* **80**, 1–43 (2016).
43. AbuSara, N. F. *et al.* Comparative Genomics and Metabolomics Analyses of Clavulanic Acid-Producing Streptomyces Species Provides Insight Into Specialized Metabolism. *Front. Microbiol.* **10**, (2019).
44. Doroghazi, J. R. & Metcalf, W. W. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **14**, 611 (2013).
45. Männle, D. *et al.* Comparative Genomics and Metabolomics in the Genus *Nocardia*.

mSystems **5**, e00125-20, /msystems/5/3/msys.00125-20.atom (2020).

46. Ziemert, N. *et al.* Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci.* **111**, E1130–E1139 (2014).
47. Hifnawy, M. S. *et al.* The genus Micromonospora as a model microorganism for bioactive natural product discovery. *RSC Adv.* **10**, 20939–20959 (2020).
48. Goldstein, S. L. & Klassen, J. L. Pseudonocardia Symbionts of Fungus-Growing Ants and the Evolution of Defensive Secondary Metabolism. *Front. Microbiol.* **11**, (2020).
49. 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).
50. Undabarrena *et al.* Rhodococcus comparative genomics reveals a phylogenomic-dependent non- ribosomal peptide synthetase distribution: insights into biosynthetic gene cluster connection to an orphan metabolite. *Microbial Genomics* (2021) DOI 10.1099/mgen.0.000621. (In press)
51. Chevrette, M. G., Hoskisson, P. A. & Barona-Gómez, F. Enzyme Evolution in Secondary Metabolism. in *Comprehensive Natural Products III* 90–112 (Elsevier, 2020). doi:10.1016/B978-0-12-409547-2.14712-2.
52. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
53. Noda-Garcia, L., Liebermeister, W. & Tawfik, D. S. Metabolite–Enzyme Coevolution: From Single Enzymes to Metabolic Pathways and Networks. *Annu. Rev. Biochem.* **87**, 187–216 (2018).
54. Noda-Garcia, L. & Tawfik, D. S. Enzyme evolution in natural products biosynthesis: target- or diversity-oriented? *Curr. Opin. Chem. Biol.* **59**, 147–154 (2020).
55. Dittmann, E., Gugger, M., Sivonen, K. & Fewer, D. P. Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol.* **23**, 642–652 (2015).
56. Liu, Z. *et al.* Formation and diversification of a paradigm biosynthetic gene cluster in plants.

Nat. Commun. **11**, 5354 (2020).

57. Fan, P. *et al.* Evolution of a plant gene cluster in Solanaceae and emergence of metabolic diversity. *eLife* **9**, e56717 (2020).
58. Liu, Z. *et al.* Drivers of metabolic diversification: how dynamic genomic neighbourhoods generate new biosynthetic pathways in the Brassicaceae. *New Phytol.* **227**, 1109–1123 (2020).
59. Tang, M.-C., Zou, Y., Watanabe, K., Walsh, C. T. & Tang, Y. Oxidative Cyclization in Natural Product Biosynthesis. *Chem. Rev.* **117**, 5226–5333 (2017).
60. Montalbán-López, M. *et al.* New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **38**, 130–239 (2021).
61. 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).
62. Nakhleh, L. Evolutionary Trees. in *Brenner's Encyclopedia of Genetics* 549–550 (Elsevier, 2013). doi:10.1016/B978-0-12-374984-0.00504-0.
63. Avni, E. & Snir, S. A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes. *Sci. Rep.* **10**, 12425 (2020).
64. 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).
65. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* gkz882 (2019) doi:10.1093/nar/gkz882.
66. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45**, D555–D559 (2017).
67. Blin, K., Shaw, S., Kautsar, S. A., Medema, M. H. & Weber, T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.* **49**, D639–D643 (2021).

68. Schniete, J. K. *et al.* ActDES – a curated Actinobacterial Database for Evolutionary Studies. *Microb. Genomics* (2021) doi:10.1099/mgen.0.000498.
69. Palaniappan, K. *et al.* IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* gkz932 (2019) doi:10.1093/nar/gkz932.
70. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, D490–D497 (2021).
71. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
72. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 1–11 (2020) doi:10.1038/s41587-020-0718-6.
73. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
74. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).
75. Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
77. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
78. 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).
79. Hoskisson, P. A. & Seipke, R. F. Cryptic or Silent? The Known Unknowns, Unknown Knowns, and Unknown Unknowns of Secondary Metabolism. *mBio* **11**, e02642-20.
80. 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.* (2020) doi:10.1101/gr.268169.120.

81. Alanjary, M., Steinke, K. & Ziemert, N. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Res.* **47**, W276–W282 (2019).
82. Adamek, M., Alanjary, M. & Ziemert, N. Applied evolution: phylogeny-based approaches in natural products research. *Nat. Prod. Rep.* **36**, 1295–1312 (2019).
83. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat. Methods* **15**, 5–6 (2018).
84. Yang, J. Y. & Ersoy, O. K. Combined Supervised and Unsupervised Learning in Genomic Data Mining. 143 (2003).
85. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
86. van der Lee, T. A. J. & Medema, M. H. Computational strategies for genome-based natural product discovery and engineering in fungi. *Fungal Genet. Biol.* **89**, 29–36 (2016).
87. Wolf, T., Shelest, V., Nath, N. & Shelest, E. CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics* **32**, 1138–1143 (2016).
88. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* **2**, (2016).
89. 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).
90. Krause, L. *et al.* GISMO—gene identification using a support vector machine for ORF classification. *Nucleic Acids Res.* **35**, 540–549 (2007).
91. Walker, A. S. & Clardy, J. A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* (2021)

doi:10.1021/acs.jcim.0c01304.

92. 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).
93. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
94. Medema, M. H., Takano, E. & Breitling, R. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. *Mol. Biol. Evol.* **30**, 1218–1223 (2013).
95. Gilchrist, C. L. M. & Chooi, Y.-H. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab007.
96. Saha, C. K., Sanches Pires, R., Brolin, H., Delannoy, M. & Atkinson, G. C. FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa788.
97. Gumerov, V. M. & Zhulin, I. B. TREND: a platform for exploring protein function in prokaryotes based on phylogenetic, domain architecture and gene neighborhood analyses. *Nucleic Acids Res.* **48**, W72–W76 (2020).
98. Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *GigaScience* **10**, giaa154 (2021).
99. Chevrette, M. G. & Currie, C. R. Emerging evolutionary paradigms in antibiotic discovery. *J. Ind. Microbiol. Biotechnol.* **46**, 257–271 (2019).
100. Chevrette, M. G. *et al.* The antimicrobial potential of Streptomyces from insect microbiomes. *Nat. Commun.* **10**, 516 (2019).
101. Miller, I. J., Chevrette, M. G. & Kwan, J. C. Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations. *Mar. Drugs* **15**, 165 (2017).
102. Caldera, E. J., Chevrette, M. G., McDonald, B. R. & Currie, C. R. Local Adaptation of

Bacterial Symbionts within a Geographic Mosaic of Antibiotic Coevolution. *Appl. Environ. Microbiol.* **85**, (2019).

103. Iglesias, A., Latorre-Pérez, A., Stach, J. E. M., Porcar, M. & Pascual, J. Out of the Abyss: Genome and Metagenome Mining Reveals Unexpected Environmental Distribution of Abyssomicins. *Front. Microbiol.* **11**, (2020).

104. Sharrar, A. M. *et al.* Bacterial Secondary Metabolite Biosynthetic Potential in Soil Varies with Phylum, Depth, and Vegetation Type. *mBio* **11**, (2020).

105. Silva, S. G., Blom, J., Keller-Costa, T. & Costa, R. Comparative genomics reveals complex natural product biosynthesis capacities and carbon metabolism across host-associated and free-living *Aquimarina* (*Bacteroidetes, Flavobacteriaceae*) species. *Environ. Microbiol.* **21**, 4002–4019 (2019).

106. Yang, Y. *et al.* Genomic characteristics and comparative genomics analysis of the endophytic fungus Sarocladium brachiariae. *BMC Genomics* **20**, 782 (2019).

107. Gutiérrez-García, K. *et al.* Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and Caulobacter spp. That Encode Niche-Specific Biosynthetic Gene Clusters. *Genome Biol. Evol.* **11**, 319–334 (2019).

108. Stubbendieck, R. M. *et al.* Competition among Nasal Bacteria Suggests a Role for Siderophore-Mediated Interactions in Shaping the Human Nasal Microbiota. *Appl. Environ. Microbiol.* **85**, (2019).

109. Chevrette, M. G. *et al.* Taxonomic and Metabolic Incongruence in the Ancient Genus Streptomyces. *Front. Microbiol.* **10**, (2019).

110. Brito, Â. *et al.* Comparative Genomics Discloses the Uniqueness and the Biosynthetic Potential of the Marine Cyanobacterium Hyella patelloides. *Front. Microbiol.* **11**, (2020).

111. 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).

112. Hoffmann, T. *et al.* Correlating chemical diversity with taxonomic distance for discovery

of natural products in myxobacteria. *Nat. Commun.* **9**, 803 (2018).

113. Gluck-Thaler, E. *et al.* The Architecture of Metabolism Maximizes Biosynthetic Diversity in the Largest Class of Fungi. *Mol. Biol. Evol.* **37**, 2838–2856 (2020).

114. Baldeweg, F., Hoffmeister, D. & Nett, M. A genomics perspective on natural product biosynthesis in plant pathogenic bacteria. *Nat. Prod. Rep.* **36**, 307–325 (2019).

115. Koonin, E. V. Archaeal ancestors of eukaryotes: not so elusive any more. *BMC Biol.* **13**, (2015).

116. 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).

117. Vior, N. M. *et al.* Discovery and Biosynthesis of the Antibiotic Bicyclomycin in Distantly Related Bacterial Classes. *Appl. Environ. Microbiol.* **84**, (2018).

118. McDonald, B. R. & Currie, C. R. Lateral Gene Transfer Dynamics in the Ancient Bacterial Genus Streptomyces. *mBio* **8**, (2017).

119. Chase, A. B., Sweeney, D., Muskat, M. N., Guillén-Matus, D. & Jensen, P. R. Vertical inheritance governs biosynthetic gene cluster evolution and chemical diversification. *bioRxiv* 2020.12.19.423547 (2021) doi:10.1101/2020.12.19.423547.

120. Bérdy, J. Bioactive Microbial Metabolites. *J. Antibiot. (Tokyo)* **58**, 1–26 (2005).

121. Traxler, M. F. & Kolter, R. Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.* **32**, 956–970 (2015).

122. Andam, C. P., Choudoir, M. J., Vinh Nguyen, A., Sol Park, H. & Buckley, D. H. Contributions of ancestral inter-species recombination to the genetic diversity of extant Streptomyces lineages. *ISME J.* **10**, 1731–1741 (2016).

123. Li, Y. *et al.* Population Genomics Insights into Adaptive Evolution and Ecological Differentiation in Streptomycetes. *Appl. Environ. Microbiol.* **85**, e02555-18.

124. Tidjani, A.-R. *et al.* Massive Gene Flux Drives Genome Diversity between Sympatric

Streptomyces Conspecifics. *mBio* **10**, e01533-19.

125. McDonald, B. R. *et al.* Biogeography and Microscale Diversity Shape the Biosynthetic Potential of Fungus-growing Ant-associated Pseudonocardia. *bioRxiv* 545640 (2019) doi:10.1101/545640.

126. Zacharia, V. M. *et al.* Genetic Network Architecture and Environmental Cues Drive Spatial Organization of Phenotypic Division of Labor in Streptomyces coelicolor. *mBio* **0**, e00794-21.

127. Zhang, Z. *et al.* Antibiotic production in Streptomyces is organized by a division of labor through terminal genomic differentiation. *Sci. Adv.* **6**, eaay5781 (2020).

128. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).

129. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).

130. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

131. Wang, H. *et al.* Soil Bacterial Diversity Is Associated with Human Population Density in Urban Greenspaces. *Environ. Sci. Technol.* **52**, 5115–5124 (2018).

132. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).

133. Nasrin, S. *et al.* Chloramphenicol Derivatives with Antibacterial Activity Identified by Functional Metagenomics. *J. Nat. Prod.* **81**, 1321–1332 (2018).

134. Santana-Pereira, A. L. R. *et al.* Discovery of Novel Biosynthetic Gene Cluster Diversity From a Soil Metagenomic Library. *Front. Microbiol.* **11**, (2020).

135. Dror, B., Wang, Z., Brady, S. F., Jurkevitch, E. & Cytryn, E. Elucidating the Diversity and Potential Function of Nonribosomal Peptide and Polyketide Biosynthetic Gene Clusters in

the Root Microbiome. *mSystems* **5**, (2020).

136. Elfeki, M., Alanjary, M., Green, S. J., Ziemert, N. & Murphy, B. T. Assessing the Efficiency of Cultivation Techniques To Recover Natural Product Biosynthetic Gene Populations from Sediment. *ACS Chem. Biol.* **13**, 2074–2081 (2018).

137. Lemetre, C. *et al.* Bacterial natural product biosynthetic domain composition in soil correlates with changes in latitude on a continent-wide scale. *Proc. Natl. Acad. Sci.* **114**, 11615–11620 (2017).

138. Reddy, B. V. B. *et al.* Natural Product Biosynthetic Gene Diversity in Geographically Distinct Soil Microbiomes. *Appl. Environ. Microbiol.* **78**, 3744–3752 (2012).

139. Waglechner, N., McArthur, A. G. & Wright, G. D. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat. Microbiol.* **4**, 1862–1871 (2019).

140. Firn, R. D. & Jones, C. G. Natural products ? a simple model to explain chemical diversity. *Nat. Prod. Rep.* **20**, 382 (2003).

141. Nguyen, T. *et al.* Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat. Biotechnol.* **26**, 225–233 (2008).

142. Masschelein, J., Jenner, M. & Challis, G. L. Antibiotics from Gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights. *Nat. Prod. Rep.* **34**, 712–783 (2017).

143. Chevrette,Marc & Helfrich. transPACT v1.0. *bioRxiv* (2021) doi:10.5281/zenodo.4148258.

144. 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).

145. Schorn, M. A. *et al.* A community resource for paired genomic and metabolomic data

mining. *Nat. Chem. Biol.* 1–6 (2021) doi:10.1038/s41589-020-00724-z.

146. Tracanna, V. *et al.* Dissecting Disease-Suppressive Rhizosphere Microbiomes by Functional Amplicon Sequencing and 10× Metagenomics. *mSystems* **0**, e01116-20.

147. Austin, M. B., O'Maille, P. E. & Noel, J. P. Evolving biosynthetic tangos negotiate mechanistic landscapes. *Nat. Chem. Biol.* **4**, 217–222 (2008).

148. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5–e5 (2018).

149. Grenade, N. L., Howe, G. W. & Ross, A. C. The convergence of bacterial natural products from evolutionarily distinct pathways. *Curr. Opin. Biotechnol.* **69**, 17–25 (2021).

150. Del Carratore, F. *et al.* Computational identification of co-evolving multi-gene modules in microbial biosynthetic gene clusters. *Commun. Biol.* **2**, 1–10 (2019).

151. Martinet, L. *et al.* A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. *mBio* **10**, e01230-19.

152. Cibrián-Jaramillo, A. & Barona-Gómez, F. Increasing Metagenomic Resolution of Microbiome Interactions Through Functional Phylogenomics and Bacterial Sub-Communities. *Front. Genet.* **7**, (2016).

153. Wiegand, S. *et al.* Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol.* **5**, 126–140 (2020).

Acknowledgments

We are grateful to Jorge Navarro-Muñoz for useful discussions and Erika V. Cruz for help with figures. Support for M.G.C. provided by grant 2020-67012-31772 (accession 1022881) from the USDA National Institute of Food and Agriculture. F.B.G. and N.S.M. are supported by Conacyt, Mexico (grant No. 285746) and the Royal Society of the United Kingdom, Newton Advanced Fellowship (NAF\R2\180631) to F.B.G. A.G. is grateful for the support of the Deutsche Forschungsgemeinschaft (DFG; Project ID # 398967434-TRR 261). S.M. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2124 – 390838134. N.Z. is funded by the German Center for Infection Research (TTU09.716).