marine drugs

Review

Targeted Large-Scale Genome Mining and Candidate Prioritization for Natural Product Discovery

Jessie James Limlingan Malit [1,2,†], Hiu Yu Cherie Leung [1,2,†] and Pei-Yuan Qian [1,2,*]

[1] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China; jjmalit@connect.ust.hk (J.J.L.M.); hyleungaq@connect.ust.hk (H.Y.C.L.)
[2] Department of Ocean Science and Hong Kong Branch of the Southern Marine Science and Engineering Guangdong Laboratory, The Hong Kong University of Science and Technology, Hong Kong, China
[*] Correspondence: boqianpy@ust.hk
[†] These authors contributed equally to this work.

Abstract: Large-scale genome-mining analyses have identified an enormous number of cryptic biosynthetic gene clusters (BGCs) as a great source of novel bioactive natural products. Given the sheer number of natural product (NP) candidates, effective strategies and computational methods are keys to choosing appropriate BGCs for further NP characterization and production. This review discusses genomics-based approaches for prioritizing candidate BGCs extracted from large-scale genomic data, by highlighting studies that have successfully produced compounds with high chemical novelty, novel biosynthesis pathway, and potent bioactivities. We group these studies based on their BGC-prioritization logics: detecting presence of resistance genes, use of phylogenomics analysis as a guide, and targeting for specific chemical structures. We also briefly comment on the different bioinformatics tools used in the field and examine practical considerations when employing a large-scale genome mining study.

Keywords: genome mining; secondary metabolites; natural products; bioactive compounds; antibiotics; genomics

check for updates

Citation: Malit, J.J.L.; Leung, H.Y.C.; Qian, P.-Y. Targeted Large-Scale Genome Mining and Candidate Prioritization for Natural Product Discovery. *Mar. Drugs* **2022**, *20*, 398. https://doi.org/10.3390/md20060398

Academic Editor: Bill J. Baker

Received: 18 May 2022
Accepted: 14 June 2022
Published: 16 June 2022

1. Introduction

Natural products (NPs), which are produced by bacteria, plants, fungi, animals, and other living cells, have been an inspiration for pharmaceutics and other biological agents such as herbicides and insecticides. In theory, NPs are structurally optimized to interact with biological targets after several hundred millennia of evolution [1] and, therefore, have a higher chance of being bioavailable than traditional small molecule drugs [2]. NPs are often categorized into primary and secondary metabolites. Primary metabolites include nutrients and building blocks for cellular maintenance, whereas secondary metabolites endow their producers with selective advantages in ecological niches and help them survive under dynamic environments [3]. For instance, some secondary metabolites function as antibiotics, pigments, and growth hormones [4]. Thus, secondary metabolite NPs have been targeted as novel drug leads with potent bioactivities, with more than 30% of approved small molecule drugs being bacterial secondary metabolites [5].

The synthesis of these compounds is catalyzed by enzymes. For bacterial and fungal secondary metabolites that had their genetic basis elucidated, the genes encoding such enzymes tended to cluster close together in the genome. These clusters are termed biosynthetic gene clusters (BGCs) [6]. Other relevant genes, such as those involved in the regulation, transport, and resistance to the self-produced compounds, can be found in or alongside the BGCs [7]. Biosynthesis enzymes of secondary metabolites have been extensively studied because of the unprecedented chemical transformations they can catalyze in different compound scaffolds [8]. Bacterial secondary metabolites can be classified according to their chemical scaffolds and signature biosynthesis enzymes that synthesize their core structures. Classes of secondary metabolites include polyketides, peptide-derived NPs, isoprenoids, terpenoids, alkaloids, nucleotide-based NPs, phenylpropanoids, and glycosylated NPs [9], which can be further divided into subclasses according to their additional features or chemical modifications.

Traditionally, the discovery of novel secondary metabolites with potent bioactivities from the pool of NPs produced by a bacterium or a microbial consortium involves a culture-first process. In this approach, the NP production occurs chronologically as follows: isolation of microorganism(s) of interest, culturing the microbe(s) under different conditions to induce NP production, and purification of desired products through chemical extraction. To test the presence of NPs with desirable bioactivities, bioassays are conducted on bacterial culture, extracted fractions, or purified products to study their effects on various organisms or human disease models, a process known as bioactivity-guided screening [10]. This traditional NP discovery pipeline does not require prior knowledge of a producer's genome nor its biosynthesis capability, and, thus, it is termed the "top-down" approach [11]. Although the "top-down" approach has facilitated the discovery of numerous therapeutic NPs, it suffers from several pitfalls. For instance, it leads to the re-isolation or rediscovery of known compounds produced by different bacterial species, because multiple species can produce the same NP [12,13]. Moreover, some secondary metabolites are difficult to be detected, purified, and examined through bioassays because they are produced at trace concentrations. Cultivation under laboratory conditions may not provide all the environmental stimuli required to produce the desired secondary metabolites. A significant percentage of microflora have been nonculturable in the laboratory, making it difficult to identify their NPs [14].

An alternative to the "top-down" approach mentioned above is to search bacterial genomes for the NP biosynthesis machineries [15]. This genome-mining-based approach, known as the "bottom-up" approach, leverages the abundance and availability of genomic data, bioinformatics analysis, and genetic manipulation tools to search for BGCs encoding novel NPs [11]. These data-mining approaches often utilize signature genes and conserved protein domains of enzymes responsible for the biosynthesis of different types of NPs in their search. These enzymes have highly conserved amino acid sequences, despite their ability to generate diverse families of compounds; hence, they can be used as seeds or reference sequences for identifying gene homologs and, therefore, novel BGCs [16]. Before cheap and fast genome sequencing techniques became available, detecting such conserved sequences of target NPs in genomic libraries was achieved by using probes in Southern hybridization experiments or degenerate primers for PCR-based detection [17]. Today, gene homologs can be detected through in silico approaches [18], many of which utilize BLAST searches [19] or profile-based analyses such as HMMER searches to characterize the target NP BGCs and search for novel BGCs in genomic data [20]. The genomic approach also allows for the genetic manipulation of biosynthesis pathways to improve yield, activation of silent BGCs, and heterologous expression of BGCs in a host; it also provides a direct link between a metabolite and its genetic origins [21].

2. Genomics for Natural Product Discovery

Several software packages have been specifically developed to search for secondary metabolite BGCs in bacterial genomes. The most widely used and powerful software pipeline currently is the antibiotics and Secondary Metabolite Analysis Shell (antiSMASH), which uses manually curated gene cluster rules to identify core and additional biosynthesis genes with profile Hidden Markov Models (pHMMs) [22]. The latest version of antiSMASH supports the identification of more than 70 types of BGCs encoding for different chemical scaffolds [23]. The software can also predict the chemical structure of a core scaffold by using the substrate specificity predictions of protein domains found in polyketide synthase (PKS) and nonribosomal peptide synthase (NRPS) modules, and assumed collinearity [22]. ClusterBlast, one of the incorporated sequence alignment tools in antiSMASH, facilitates the comparison of BGCs; this comparison is necessary for the dereplication process, wherein previously known and characterized BGCs from the mining results are removed [24]. Other genome-mining tools developed for BGC detection include PRISM, which also uses pHMMs to detect BGCs and additionally predicts biological activity through machine learning [25]; and MetaBGC, which identifies BGCs from the sequence data of microbial communities [26]. The Secondary Metabolite Bioinformatics Portal (SMBP) provides a list of genome-mining software and includes links to software, tools, and databases relevant to NP discovery with omics data [27].

Technological advancements and the rapid increase in the amount of publicly available genome information in the previous decades have spurred many large-scale in silico genome-mining analyses. Instead of mining a single genome at a time, we can now incorporate big data approaches to mine a large pool of genomes or metagenomes. The large-scale, extensive search in whole-genome databases for novel NPs is an approach that we herein refer to as "global genome mining". In addition to the manually curated BGC databases, and those from the National Center for Biotechnology Information (NCBI) [28], several publicly available online databases can facilitate the global genome mining analysis of BGCs. The antiSMASH database is a repository of antiSMASH-annotated BGCs from more than 20,000 bacterial genomes and contains more than 150,000 BGCs. The Integrated Microbial Genomes Atlas of Biosynthetic Gene Clusters (IMG-ABC) includes more than 400,000 BGCs detected in archaea, bacteria, fungi and metagenome bins [29]. The Minimum Information about a Biosynthetic Gene Cluster (MIBiG) database aims to collect all characterized BGCs with known functions and the secondary metabolites they produce [6]. This database is especially useful in identifying BGCs in sequenced genomes that can produce the same or similar sets of compounds according to sequence homology, thus preventing the rediscovery of known compounds. The MIBiG currently contains the BGC information of 1923 secondary metabolites. A more comprehensive list of databases that can aid large-scale BGC studies can be found in a recent review [30]. Handling these huge datasets requires the development of downstream bioinformatics platforms to properly display the uncharted NP diversity hidden in cryptic BGCs of genomes (Figure 1). Examples of such bioinformatics tools that can handle large-scale biological sequence data include EFI-EST [31], FastTree [32], and cd-hit [33]; they also have the ability to be integrated into the global genome mining pipeline for data analysis and visualization.

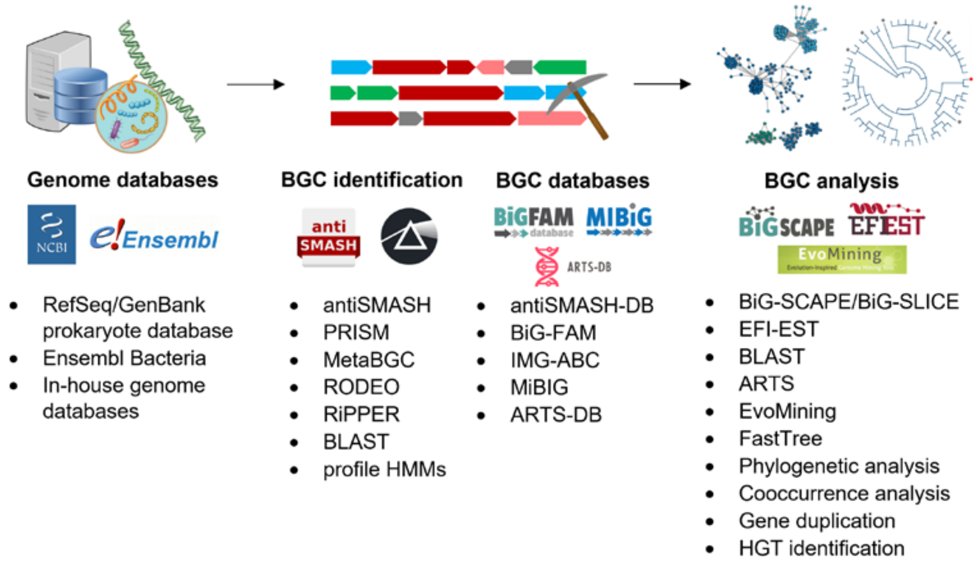

Figure 1. General workflow and examples of bioinformatic tools for natural product discovery guided by large-scale genome mining.

Several omics-based strategies have been developed for novel secondary metabolite discovery, including transcriptomics, proteomics, metabolomics, and meta-omics [16,34,35], as well as machine learning tools [36–38]. Their integration into the NP discovery pipeline has been discussed in previous reviews [16,34–38]. Given the immense, dark chemical space hidden in these cryptic BGCs and the large volume of hits that can result from a single round of global genome mining, one of the most significant challenges now is the prioritization of candidate BGCs that are worthy of further study. In the present review, we focus on natural product discovery studies that conduct large-scale targeted genome-mining approaches on manually curated or publicly available genome datasets. We present studies that have successfully uncovered novel NPs by using NP resistance determinants, phylogenomic analysis, or enzyme-catalyzed structural modifications to prioritize BGCs (Table 1). The aim of this review is to provide ideas and inspirations for researchers conducting large-scale natural-product genome-mining studies on how to identify candidate BGCs for further study.

Table 1. Genetic elements and NP features targeted by resistance, phylogenomic, structure, and RiPP-guided genome-mining strategies and the natural products they identified.

Resistance-Gene-Guided			
Resistance Gene(s)	**Natural Product**	**Source Organism**	**Reference**
pentapeptide repeat protein (PRP) sequences	alkylpyrone-407	*Cystobacterineae* strain MCy9487	[39]
	pyxidicycline A	*Pyxidicoccus fallax* An d48	[40]
dihydroxyacid dehydratase	aspterric acid	*Aspergillus terreus* NIH2624	[41]
tripartite efflux system PleABC	prosekin	Pseudomonas prosekii LMG 26867	[42]
lanosterol 14α-demethylase	lanomycin	*Pyrenophora dematioidea* TTI-1096	[43]
fatty acid synthase	thiotetroamide	Streptomyces afghaniensis NRRL 5621	[44]
D-stereospecific peptidase	bogorol	*Brevibacillus laterosporus* DSM 25	[45]
Phylogenomics-Guided			
Sequences Used for Phylogenetic Analysis	**Natural Product**	**Source Organism**	**Reference**
Whole BGCs of different families	aryl polyenes	*Escherichia coli* CFT073	[46]
"Expanded-then-recruited" enzyme families; 3-carboxyvinyl-phosphoshikimate transferase	arseno-organic metabolites	*Streptomyces lividans* 66	[47]
Each shared gene found in glycopeptide antibiotic-producing BGCs	corbomycin	*Streptomyces* sp. WAC01529	[48]
ATP-grasp ligase	MdnA7	*Cyanothece* sp. PCC 7822	[49]
LuxR	cepacin A	*Burkholderia ambifaria* BCC0191	[50]
Whole BGCs containing *tauD* expansion	detoxin S1	*Streptomyces* sp. NRRL S-325	[51]
terpene synthase	hydropyrene	*Streptomyces clavuligerus* ATCC 27064	[52]
chain length factor (CLF) protein	oryzanaphthopyran A	*Streptacidiphilus oryzae* CGMCC 4.2012	[53]
Structure-Guided			
Targeted Chemical Structure	**Natural Product**	**Source Organism**	**Reference**
cationic amino acid residues	brevicidine	*Brevibacillus laterosporus* DSM 25	[54]
chemical transformations catalyzed by cytochrome P450 on cyclodipeptides	cyctetryptomycin B	*Saccharopolyspora hirsuta* DSM 44795	[55]
prenyl groups on cyclodipeptides	griseocazine D1	*Streptomyces griseocarneus* 132	[56]
thioether bonds	freyrasin	*Paenibacillus polymyxa* ATCC 842	[57]

Table 1. *Cont.*

chemical transformations catalyzed by the DUF–SH didomain	guangnanmycin	*Streptomyces* sp. CB01883	[58]
phosphonic acid	argolaphos A	*Streptomyces monomycini* NRRLB-24309	[59]
Global Genome Mining for RiPPs			
Combining the structure-guided strategy with precursor peptide sequence search			
RiPP Family	**Natural Product**	**Source Organism**	**Reference**
lanthipeptide	birimositide	*Streptomyces rimosus* subsp. *rimosus* WC3908	[60]
cyanobactin	tolypamide	*Tolypothrix* sp. PCC 7601	[61]
polyoxazole-thiazole-based cyclopeptide	aurantizolicin	*Streptomyces auranticaus* JA 4570	[62]
thiopeptide	saalfelduracin	*Amycolatopsis saalfeldensis* NRRL B-24474	[63]
thioamitides	thiovarsolin A	*Streptomyces varsoviensis* DSM 40346	[64]
sactipeptide	streptosactin	*Streptococcus thermophilus* JIM 8232	[65]
lanthipeptide	flavucin, agalacticin, etc.	*Corynebacterium lipophiloflavum* DSM 44291	[66]

3. Resistance-Gene-Guided Genome Mining

One approach in performing global genome mining for novel bioactive compounds is targeting biosynthesis-associated resistance determinants. Although most BGCs only contain genes responsible for synthesizing the desired secondary metabolite, some BGCs include genes involved in protecting the producing organism from the metabolite's harmful effects [7]. These genes, called resistance genes, encode enzymes that impart immunity or resistance to the producer and generally work in three different ways. In product detoxification, the protein binds to and, in many cases, also modifies the produced secondary metabolite, thereby preventing the metabolite from binding to its target in a producer strain. An enzyme can also rapidly remove an NP from a cellular location or by binding it to a transporter. Lastly, a BGC can contain a duplicate resistant version of the NP's target, or an enzyme that catalyzes the modification of the target, to render it resistant to a produced NP [7]. Several bioactive compounds and their modes of action have been identified by using resistance-guided genome mining (Figure 2).

As mentioned, resistance genes can be colocalized with the BGC of the NP it protects it from. Griselimycin [67], salinosporamide A [68], and platensimycin [69] are examples of NPs discovered to contain resistance determinants in their BGCs. In genome mining, we can use previously confirmed biosynthesis-associated resistance genes as molecular "beacons", illuminating BGCs that could produce novel compounds with similar bioactivities as the known NP. For instance, to search for novel topoisomerase inhibitors with anticancer and antibiotic qualities, Panter et al. [40] mined myxobacterial genomes for BGCs containing topoisomerase-targeting pentapeptide repeat protein (PRP) sequences. The choice of PRP was primarily inspired by cystobactamids, which contained PRP sequences in their BGC to protect the native gyrase from the cystobactamids. Their study led to the identification of pyxidicyclines–Type II polyketides with an unusual nitrogen-containing tetracene moiety. These compounds indeed inhibited bacterial topoisomerase IV, leading to its bactericidal effects, with an IC_{50} range of 6.25–1.6 μg/mL. The compounds had an MIC value of 0.06 μg/mL against HCT-116 cells, as well. The mode of action was believed to be through the inhibition of human topoisomerase I [40]. The same research group used PRP sequences again to prioritize Type III PKS candidate BGCs out of the 116 identified from myxobacterial genomes. A candidate BGC from *Cystobacterineae* strain MCy9487 was heterologously expressed in a myxobacterial model strain *Myxococcus xanthus* DK1622, which produced novel alkylpyrones that exhibited bacterial topoisomerase IV and human topoisomerase I inhibition, leading to cytotoxicity in HCT-116 cells but not bactericidal activity [39].

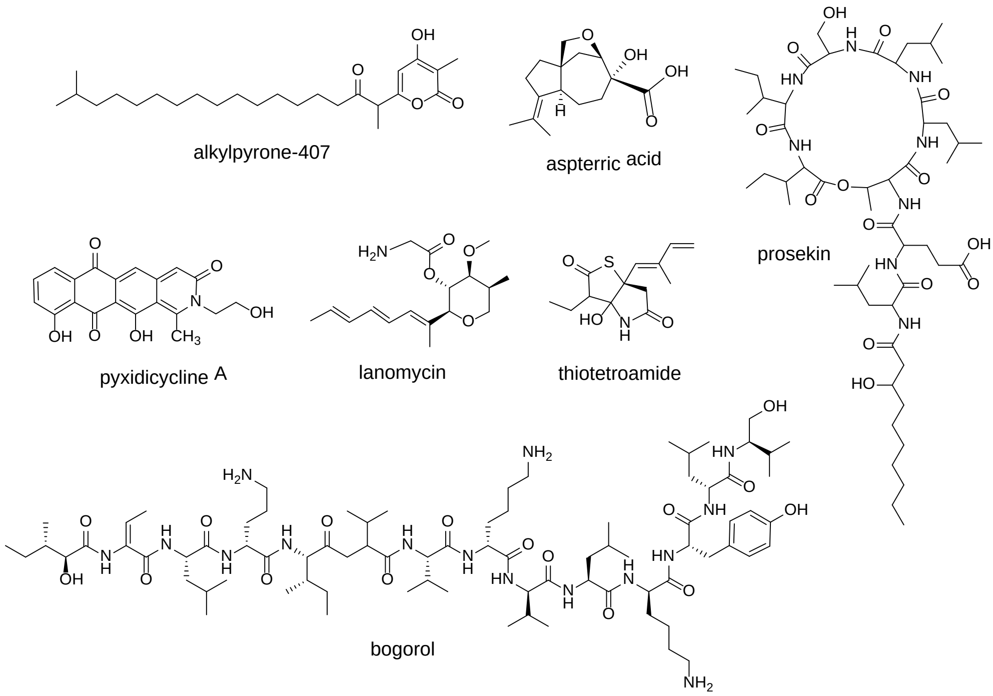

Figure 2. Compounds identified through resistance-gene-guided genome mining and their resistance determinants (in brackets): alkylpyrone-407 and pyxidicycline A (pentapeptide repeat protein (PRP) sequences), aspterric acid (dihydroxyacid dehydratase), prosekin (tripartite efflux system PleABC), lanomycin (lanosterol 14α-demethylase), thiotetroamide (fatty acid synthase), and bogorol (D-stereospecific peptidase).

The enzymes responsible for transport of the produced compound can also be used to select for BGCs of interest. In a study focusing on the biosynthesis pathways of lipopeptides (LPs) produced by *Pseudomonas*, Girard et al. [42] identified a transport system dedicated to LP export by exploring the synteny of the LP-encoding genomes. This tripartite efflux system removes the antimicrobial LP from the *Pseudomonas* and is composed of three proteins, PleABC. These three genes are found near the LP-encoding BGC. The protein encoding for the inner membrane ABC transporter (PleB) was used as bait in BLASTP and antiSMASH analyses, and they obtained 75 putative LP-encoding BGCs from *Pseudomonas* genomes. Through a phylogenetic analysis of *pleB*, they identified four LP-encoding BGCs that contain *pleB* genes of various evolutionary distances from *pleB* in known LP BGCs. The strains containing these BGCs were subjected to crude extraction and chemical characterization, which led to discovery of a novel LP, prosekin.

This resistance-gene-guided genome-mining approach has also been applied to genome mine fungal genomes. Dihydroxyacid dehydratase (DHAD), an enzyme involved in the branched chain amino acid biosynthesis pathway in plants, is essential to plant growth [70]. The absence of an equivalent pathway in animals suggests that this enzyme can be used as a herbicide target. To identify NPs with DHAD-inhibiting activities, Yan et al. [41] hypothesized that DHAD inhibitor genes are likely to appear alongside an inhibitor-resistant DHAD. They scanned fungal genomes for core biosynthesis enzymes that co-localized with DHAD and identified a cluster containing a sesquiterpene cyclase, two cytochrome P450s, and a DHAD homolog. This BCG was heterologously expressed in *Saccharomyces cerevisiae* RC01, producing aspterric acid. Extensive bioactivity testing showed that aspterric acid inhibited DHAD. Spray treatment on wild-type *Arabidopsis thaliana* drastically impeded plant growth and pollen formation. Transgenic *A. thaliana* containing a copy of this cluster′s DHAD gene was unaffected by the aspterric acid treatment, showing that it was indeed an inhibitor-resistant DHAD [41]. Using the same self-resistance gene logic, their group used a cytochrome P450 protein — lanosterol 14α-demethylase (CYP51), as a query in another search for bioactive compounds. CYP51 is involved in the ergosterol biosynthesis pathway and could be an effective target for anticholesterol and antifungal drugs. The group developed an algorithm to identify BGCs that have their core biosynthesis enzymes colocalized with CYP51, and identified a candidate. After initial confirmation of CYP51 activity of the candidate BGC through in vitro assays, heterologous expression was performed, uncovering the biosynthesis pathway of restrictin and lanomycin [43].

In a study to identify novel resistance determinants found in BGCs, our laboratory identified D-stereospecific resistance peptidases that act on D-amino acid–containing nonribosomal peptides (DNRPs). In the global genome mining pipeline that we developed, a networking analysis of biosynthesis elements, transporters, and regulators was applied to 2511 DNRP-encoding BGCs to identify novel resistance determinants. We found that D-stereospecific peptidases were strongly associated with DNRP biosynthesis genes and selected their coincident BGCs for further evaluation. Two candidate strains, *Brevibacillus laterosporus* DSM 25 and *Paenibacillus polymyxa* CICC10580, produced the DNRPs bogorol and tridecaptin, respectively. In the biochemical characterization of the associated D-stereospecific peptidases in each BGC, we showed that the two enzymes can cleave their respective DNRPs [45].

A phenomenon that is suggested to confer self-resistance is gene duplication, as it increases the number of the antibiotic target, thereby freeing up evolutionary pressure to allow the target to mutate into a resistant form. Genes that have undergone duplication can be identified by constructing a core genome that is conserved throughout a bacterial taxon and searching the individual pangenomes for additional gene copies, suggesting a gene duplication or horizontal gene transfer event. In a study by Tang et al. [44], they probed *Salinispora* genomes for duplicated housekeeping genes. They did this by identifying all the conserved orthologous groups (OGs) of protein-coding genes in *Salinispora*, and identified those that were shared among all strains to form the 'core genome'. The OGs were grouped into clusters (COGs) to denote their general biological function. With a list of COGs and their hypothetical general functions, they then searched the pangenomes for additional OGs that allegedly had the same function as the OGs in the core genome, which could be duplicated OGs. Their pipeline correctly detected the presence of a duplicated 20*S* proteasome β-subunit present in the salinosporamide BGC, showing that their method was viable. Although duplication was detected in multiple COGs in their probing of *Salinispora*, they focused on a COG assigned to lipid transport and metabolism. The bacterial fatty acid synthase, in particular, was deemed favorable, as it displayed significant structural differences to its equivalent in mammals, implying low cytotoxicity of the produced compound. They discovered thiolactomycins and thiotetroamides by identifying a hybrid PKS-NRPS BGC that contained a duplicated fatty acid synthase gene. Indeed, the observed antibiotic activities of these families of compounds were owed to fatty acid synthase inhibition [44]. Something to be taken into consideration before using gene duplication phenomenon for NP target discovery, however, is that gene duplication events do not guarantee duplicated genes with new functions, because they may only serve metabolic redundancy and are not recruited to secondary metabolism pathways [71].

Although enzymes that were previously confirmed as resistance determinants provide a good starting point for a resistance-gene-guided genome-mining study, our study and that of Tang et al. showed that determining a specific resistant determinant beforehand is not necessary [44,45]. We can exploit the logic behind resistance mechanisms, instead of focusing on a known resistance gene, to discover new compounds and resistance mechanisms. Still, inspiration can be drawn from BGCs confirmed to contain resistance-related genes [7], or from compounds with an unknown biosynthesis pathway but with an established mode of action. Enzymes involved in their mode of action can be targeted and possibly be co-localized with a novel producing BGC. Resistance-gene databases, such as The Comprehensive Antibiotic Resistance Database [72] and Resfams [73], can be used to identify targets for genome mining. Prioritizing candidate BGCs associated with resistance determinants is also advantageous, as compounds identified with this pipeline are more likely to display bioactivities because their mode of action is hypothesized through the biological function of the resistance genes.

4. Phylogenomics-Guided Genome Mining

As the resistant-target-based mining in the previous section already suggested, enzymes that are involved in secondary metabolite biosynthesis are basically paralogs of enzymes involved in primary metabolite generation [74]. In the same vein, these secondary metabolite biosynthesis enzymes can evolve and, thus, can lead to the catalysis of novel chemical transformations. This evolution generally occurs through the expansion of the enzyme's substrate specificity [47]. As selective pressure directs these enzymes to produce metabolites with enhanced bioactivities, divergent enzymes can be used to identify BGCs encoding for NPs with novel chemical modifications and bioactivities [48]. This logic has been applied to discover novel NPs through the construction of phylogenetic trees to identify BGCs containing divergent biosynthesis enzymes, leading to the production of new compounds (Figure 3).

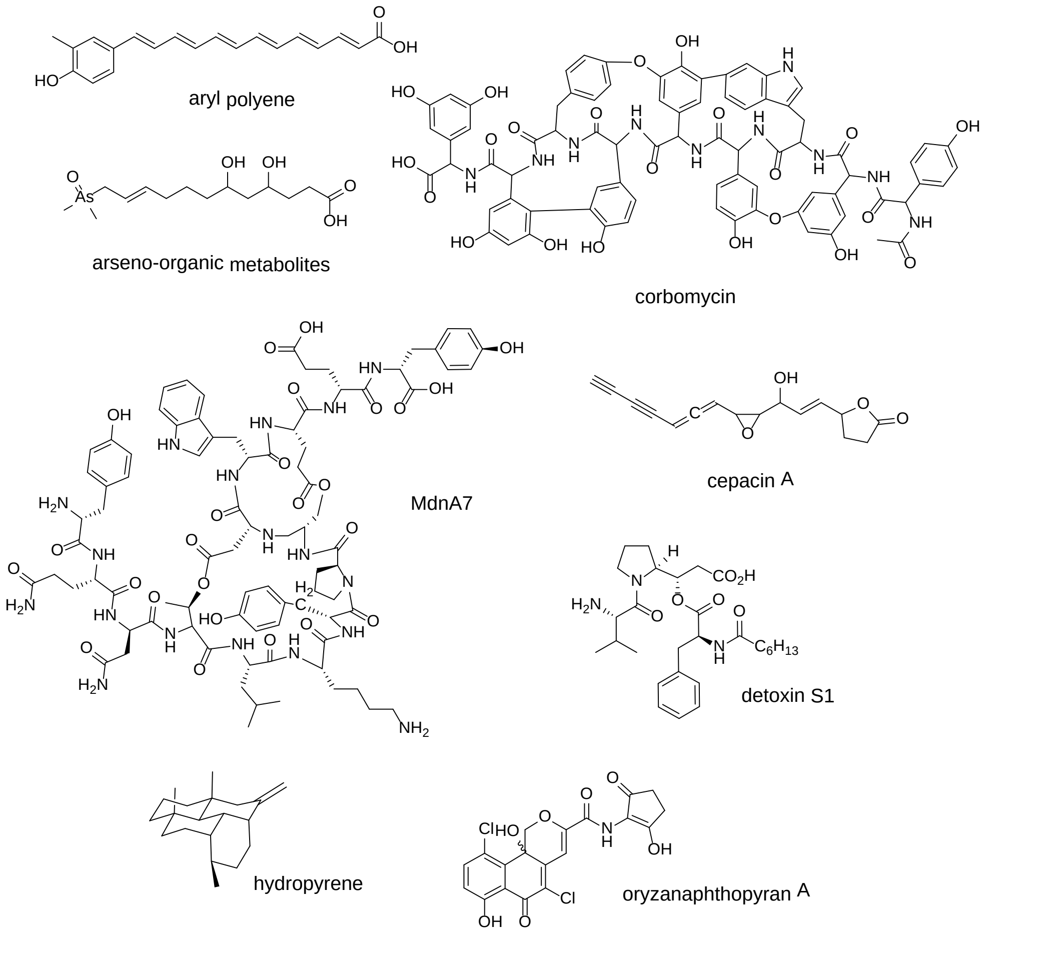

Figure 3. Compounds with novel chemical scaffolds identified through phylogenomics-guided genome mining: aryl polyenes from *Escherichia coli*, arseno-organic metabolites from *Streptomyces lividans*, corbomycin from *Streptomyces* sp. WAC01529, MdnA7 from *Cyanothece* sp. PCC 7822, cepacin A from *Burkholderia ambifaria*, detoxin S1 from *Streptomyces* sp. NRRL S-325, hydropyrene from *Streptomyces clavuligerus* ATCC 27064, and oryzanaphthopyran A from *Streptacidiphilus oryzae* CGMCC 4.2012.

One of the first studies that used a global bacterial genome database for the analysis of secondary metabolism had used phylogenomics-guided mining and was undertaken by Fischbach's laboratory. It developed the ClusterFinder algorithm to detect the BGCs of known and unknown NP classes. At the time of its creation, ClusterFinder raised interest when preexisting tools could only detect BGCs from characterized classes, but the software has since fallen out of favor for NP genome mining due to its limitations and the discovery that many of its hits were not valid secondary metabolite BGCs [23,75–77]. The algorithm uses an HMM-based probabilistic model trained on a curation of 732 BGCs that produce various known compounds. The genome sequence is converted into strings of PFAM domains by the software, and each domain is assessed for the probability that it belongs to a BGC, based on the training set used. Since the algorithm purely depends on PFAM domain frequencies and not on NP-specific genetic signatures, novel BGC classes can be detected. They applied this algorithm to 1154 bacterial genomes spanning the prokaryotic tree of life and performed a global phylogenomic analysis of all prokaryotic BGCs. They further constructed a BGC distance network by using evolutionary distances and identified groups of uncharacterized BGCs that are not detected by other genome-mining tools. They expressed representative BGCs from the group containing the greatest number of BGCs in *Escherichia coli*, producing aryl polyene carboxylic acids [46].

Cruz-Morales et al. [47] performed mining without targeting any particular NP class by identifying enzymes that have undergone enzyme expansion and repurposing from the central metabolism. These divergent enzymes are possibly recruited by secondary metabolite BGCs for the catalysis of chemical transformations devoted to NP biosynthesis. They started with identification of central metabolic enzymes and their orthologs from 230 actinobacterial genomes, and then they identified expanded enzyme families, selecting those with a higher number of orthologs than the average documented for enzyme families in the genome database. The identified expanded enzyme families were then searched against a database of NP-related enzymes from 226 known actinobacterial BGCs, and 23 expanded enzyme families putatively involved in NP biosynthesis were identified. Performing a phylogenetic analysis on each of the 23 enzyme families showed the presence of divergent clades, most of which contained homologs of enzymes from characterized BGCs. From these results, they focused on the 3-phosphoshikimate-1-carboxivinyl transferase family, AroA, and identified its associated BGC. The functional characterization of this BGC facilitated the production of a novel arseno-organic metabolite [47].

For known NP classes, researchers can target a specific gene or genes known to be conserved in known members of that class. Novel microviridins were isolated by using chemo-enzymatic synthesis by Ahmed et al. [49] after they prioritized their BGC hits with phylogenomics analysis. They identified 174 microviridin BGCs across the bacterial domain and then used the conserved MdnC ATP-grasp ligase gene to construct a maximum likelihood phylogenetic tree. The constructed tree revealed the presence of cryptic microviridin BGCs for further evaluation. They selected the microviridin BGC from *Cyanothece* sp. PCC 7822, as it putatively encodes for an unusual number of 10 precursor peptides. It also contains leucine, arginine, and lysine residues on critical positions that interact with serine proteases, thus suggesting that these microviridins could display serine protease inhibition. A combined chemo-enzymatic synthesis approach was used to successfully produce the new microviridins MdnA3, 6, 7, 8, and 9, with MdnA6 inhibiting trypsin at an IC50 of 21.5 μM, becoming the most potent microviridin-based trypsin inhibitor to date [49]. In a study by Culp et al. [48] on glycopeptide antibiotics (GPA), they constructed phylogenetic trees for not one conserved gene, but every common gene and gene segments from 71 GPA BGCs. The phylogenetic tree built from the condensation domains in the C2 module of nonribosomal peptide synthase—an enzyme involved in GPA production—revealed divergent clades from "true" GPAs. The "true" GPA BGCs possessed known GPA resistance genes, while those in the divergent clades lacked them. There were two such divergent clades, and a candidate was selected from one of the clades for BGC characterization. Subsequent fermentation led to the production of a novel compound corbomycin. Together with the characterized compound complestatin from the other clade, the two GPAs were shown to block autolysin activity required for cell wall development, a novel mode of action for GPAs, ultimately inhibiting bacterial growth [48].

A study by Yamada et al. on terpene synthases (TSs) found in bacteria was a case of conducting phylogenomics-guided mining on a gene that relatively lacks significant conserved sequences across species. As opposed to the conserved biosynthesis enzymes of other NPs, TSs originating from bacteria do not have significantly conserved domains and, thus, are not amenable for detection through sequence similarity such as in BLAST. A combined HMM and PFAM search method was thereby developed to mine for TSs found in bacterial genomes [78]. By mining from more than 8 million bacterial proteins, they were able to identify 262 proteins putatively encoding for TSs. They aligned and constructed a phylogenetic tree from these sequences to assign functions to the TSs found in several clades, such as geosmin and epi-isozizaene synthases. They were especially interested in isolated clades without a characterized member and prioritized them for heterologous expression in a *Streptomyces avermitilis* host. This has led them to isolate 13 new compounds—two sesquiterpenes, which were named hydropyrene, and its derivative, hydropyrenol; and 11 diterpenes [52].

A similar approach was used by Chen et al. in their study, wherein a global phylogenetic tree was constructed for the chain length factor (CLF) protein, a domain partly responsible for condensation reactions in Type II PKS synthases. They predicted the chemical class and the uniqueness of the polyketide structure that BGCs containing CLF proteins produced. To validate their results, several candidates that were evolutionarily distant from clades containing characterized proteins were selected for compound production. They characterized a novel polyketide oryzanaphthopyran, which contains an unusual angular naphthopyran scaffold [53].

Many NP global genome mining studies used phylogenetic mapping to examine their results. Most of the studies described in the present review have conducted phylogenetic mapping on candidate BGCs or genes. The exercise of constructing a phylogenetic tree for a quorum sensing protein led to a serendipitous NP BGC discovery, as reported by Mullins et al. [50]. Mullins et al. aimed to investigate the mechanisms behind the biocontrol activities of the noted pathogen killer *Burkholderia ambifaria* by analyzing the genomes of 64 *B. ambifaria* strains. They identified the BGCs in those genomes that putatively coded the known secondary metabolites of *B. ambifaria*. Of the BGCs, they constructed a phylogenetic tree for LuxR, a protein involved in quorum sensing system LuxRI, as it has been observed that the expression of BGCs can be regulated by quorum sensing. They discovered a candidate from the tree that featured a PKS BGC downstream of a LuxRI system. The characterization of this BGC led to production of cepacin A, a known potent anti-oomycetal that previously had unknown BGC origins. They further demonstrated that cepacin A inhibited the growth of the oomycete *Pythium ultimum*, which causes the damping-off disease in germinating crops [50].

Navarro-Muñoz et al. [51] developed two applications to aid in the analysis of huge BGC datasets: the Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE) and the Core Analysis of Syntenic Orthologues to Prioritize Natural Product Gene Clusters (CORASON). BiG-SCAPE groups BGCs identified by antiSMASH into gene cluster families (GCFs) by sequence similarity, whereas CORASON establishes phylogenetic relationships between the detected BGCs and generated GCFs. They used these two applications to mine 3080 actinobacterial genomes for detoxin and rimosamide BGCs. These BGCS were organized into GCFs by BiG-SCAPE, uncovering a conserved set of core genes, specifically *tauD*, which is present in all known detoxin/rimosamide-related BGCs. CORASON then constructed phylogenetic trees from the *tauD*-containing BGCs and revealed unexplored clades for further analysis. An LC–MS/MS metabolomics dataset that contained strains harboring BGCs from these unexplored clades was analyzed through molecular networking, resulting in the identification of 99 putatively novel detoxin or rimosamide analogs. They focused their characterization efforts on three detoxin BGC clades containing interesting characteristics for NP production: One of the clades has BGCs containing putative cytochrome P450 and enoyl-CoA hydratase/isomerase genes, leading to the production of detoxin S_1, which contained an additional heptanamide side chain. The second clade had the detoxin BGCs adjacent to a spectomycin BGC, resulting in the identification of detoxins N_1–N_3, which feature a N-formylated tyrosine instead of phenylalanine. The third clade composed of *Amycolatopsis* BGCs containing a cytochrome P450 gene, revealing detoxins P_1–P_3, a diverse set of detoxins incorporating different amino acids.

A phylogenomics-based mining approach incorporates different evolutionary theories that can supplement sequence-similarity and rule-based genome-mining approaches. It searches for divergent secondary metabolism enzymes or BGC features and, thus, can aid in identifying compounds with novel chemical scaffolds [46,47]. However, phylogenetic tree construction is not guaranteed to work for all enzymes or NP classes, and not all divergent enzymes catalyze new chemical reactions. In addition, the line between secondary metabolism enzymes and primary metabolism enzymes is not always clear, because of our limited understanding of gene evolution. Given this limitation, it is highly recommended to avoid the most common genes and to select slightly more uncommon genes for phylogenetic analysis [71]. Focusing on established core biosynthesis enzymes of NPs could help alleviate this drawback, as well. This is less likely to lead to the discovery of novel NP scaffolds, although there are examples of this happening.

5. Structure-Guided Genome Mining

In large-scale genome-mining studies, predictions on the structure of the NP core scaffold and predictions on its additional tailoring steps can show the possible chemical diversity of secondary metabolites produced by the uncharacterized BGCs. The predictions can also be used to discover novel enzymes that catalyze unprecedented chemical transformations. Thus, enzymes known to catalyze intriguing chemical reactions on NPs have been used to query and prioritize BGCs for interesting NP structures or novel NPs (Figure 4).

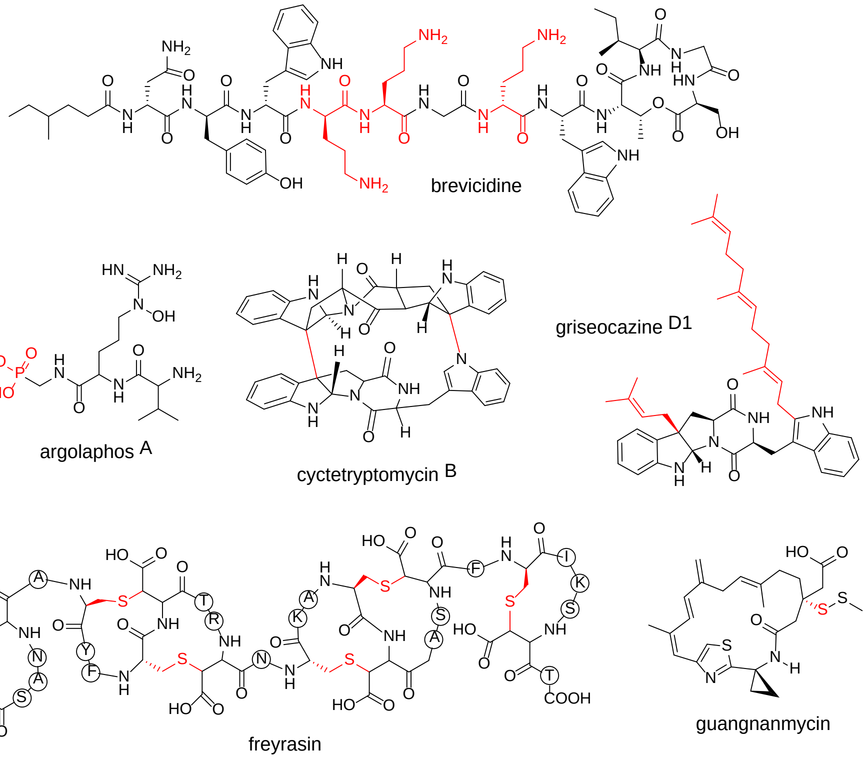

Figure 4. Compounds identified through structure-guided genome mining and specific chemical moieties targeted for (in red): brevicidine (cationic amino acid residues), argolaphos A (phosphonic acid), cyctetryptomycin B (chemical transformations catalyzed by cytochrome P450), griseocazine D1 (prenyl groups), freyrasin (thioether bonds), and guangnanmycin (chemical transformations catalyzed by the DUF–SH didomain).

Over 10,000 actinobacterial species were screened by Ju et al. [59] for the presence of phosphoenolpyruvate mutase gene *pepM*, which is an essential gene in the biosynthesis of phosphonic acids. Draft genomes were first obtained for the 278 strains containing *pepM*. The associated BGCs containing the gene were clustered together through a networking analysis, which uncovered uncharacterized phosphonic acid–producing GCFs. They selected four uncharacterized groups for further study and discovered three antibacterial phosphonopeptides, namely argolaphos AB and valinophos, along with the sulfur-containing phosphonates phosphonocystoximates and H-phosphinates.

Our group performed global genome mining of 162,672 bacterial genomes to search for novel cytochrome P450-associated cyclodipeptide (CDP) synthase BGCs. Cytochrome P450s have been previously shown to be able to catalyze a variety of chemical transformations on CDPs [79]. We found 829 BGCs that had such an association and selected one BGC from *Saccharopolyspora hirsuta* DSM 44795 that contained two cytochrome P450 genes, as we suspected that the two enzymes can perform sequential reactions on the CDP precursor. The heterologous expression of this BGC led to the description of the novel cyclodipeptides cyctetryptomycins A and B, which contain an unusual large macrocyclic core. Further bioassay testing showed that these compounds display potent neuroprotective activity [55]. Following this study, we were inspired by how prenylation can enhance the bioactivities of compounds and aimed to uncover new prenylated compounds by following the same pipeline. In this case, we targeted prenyltransferase-associated CDP synthase BGCs. There were substantially less of these BGCs than those containing cytochrome P450s, with only 26 BGCs detected. Surprisingly, a BGC from *Streptomyces griseocarneus* 132 that contained two prenyltransferase genes downstream of the CDP synthase gene was identified, akin to our previous CDP study. The BGC was then heterologously expressed, producing prenylated cyclodipeptides. We named this new family of compounds griseocazines, and a preliminary structure–activity relationship assay revealed that one of the griseocazines, the multiprenylated griseocazine D1, showed significantly improved neuroprotective activity compared to its nonprenylated counterpart [56].

Guangnanmycins and weishanmycins are leinamycin-type NPs discovered after large-scale genome mining by Pan et al. [58]. Leinamycin (LNM) contains a rare 1,3-dioxo-1,2-dithiolane moiety and has very potent antitumor activity, even in drug-resistant tumors. Its biosynthesis also involves a unique enzymology, involving a Type I PKS that lacks an acyltransferase domain; a bifunctional acyltransferase/decarboxylase responsible for β-alkylation; and an unusual domain of unknown function (DUF) and a cysteine lyase domain (SH), which catalyzes the incorporation of sulfur into the LNM backbone. In an effort to discover novel LNM analogues that also contained such sulfur incorporation, they identified *lnm*-type gene clusters by using the domains DUF–SH as a probe. They identified 19 putative *lnm*-type BGCs from 48,780 bacterial genomes and an additional 30 BGCs from their in-house culture collection of 5000 strains. All of the 49 identified strains were subjected to fermentation, and the two compounds were discovered.

In a sactipeptide study from Hudson et al. [57], the sequence similarity network of putatively identified rSAM maturases of sactipeptide and sactipeptide-like SCIFFs (six cysteines in forty-five residues [80]) and a maximum likelihood tree suggested that SCIFF-related rSAM maturases were closer to the QhpD protein family than they are to known sactipeptide rSAM maturases. QhpD enzymes catalyze S-Cβ and S-Cγ thioether linkages, but not the S-Cα ones that define the sactipeptide class. As it had been unclear previously whether SCIFFs possessed S-Cα linkages [81], this suggestion of SCIFFs possibly being not true sactipeptides inspired Hudson et al. to select a putative SCIFF, freyrasin, from *Paenibacillus polymyxa* ATCC 842 for characterization, and they were able to show that it did not have S-Cα links. They suggested the renaming of these sactipeptide-like peptides with non-S-Cα thioether bonds to radical non-alpha thioether peptides, or ranthipeptides [82]. This is another example of the benefit of large-scale mining allowing for serendipitous discoveries wherein unexpected patterns in the data occur and inform further investigations.

Our group used the predicted NP structure of candidates, directly selecting for NP features with theoretical bioactive functions to guide BGC prioritization of cationic nonribosomal peptides (CNRPs). Carrying positive charges, CNRPs can interact with the negatively charged bacterial cell membrane and have been demonstrated to have antibiotic activity against Gram-negative bacteria [83]. To analyze their diversity and discover novel CNRPs, we analyzed 7935 complete bacterial genomes for CNRP-encoding BGCs and identified 11,286 CNRP BGCs. We reduced the number of BGCs to 807, focusing on those with two or more positively charged residues in their CNRPs and discarding BGCs of siderophores and of shorter peptides. We then constructed a peptide-similarity network for these CNRPs, showing their diversity. Using this network as a guide, we identified clusters of uncharacterized putative N-acylated CNRPs—a group of lipopeptides, with members reported to be effective against Gram-negative bacterial infections. The presence of acyl groups on a hydrophobic moiety on N-acylated CNRPs can also impart membranolytic and cell-penetrating functions [84]. From the 261 putative N-acylated CNRPs, we decided to target those harbored in Bacilli because the taxon is a known producer for many CNRPs. We selected candidates with three or more positively charged amino acid residues, as they would theoretically show increased efficacy in disrupting the Gram-negative cell membrane. We finally produced three novel N-acylated cyclic depsipeptides: brevicidine, laterocidine, and paenibacterin B, which are broad spectrum Gram-negative antibiotics. We treated *E. coli* with the newly discovered CNRPs and observed the disruption of the outer cell membrane with atomic force microscopy [54].

As shown by these studies, a desired chemical moiety or enzymatic reaction can be used to target specific BGCs. Usually, a gene that produces an enzyme that catalyzes such chemical transformation is used as a bait, but conversely a predicted structure based on conserved protein domains could also be used, as in the case of NRPS and ribosomally produced peptides (RiPPs) (discussed in next section). The motivations for these structure-first prioritization approaches are usually the potent bioactivity imparted by the targeted chemical modifications, and to find analogs for compounds with a unique biosynthesis logic, or it could be simply to discover novel enzyme catalytic reactions. After filtering, phylogenetic and networking analyses can be used in conjunction with this approach in order to group BGCs with similar genetic architectures, thus exposing uncharacterized BGCs and allowing for the identification of divergent enzymes, both of which increase the likelihood of obtaining novel compounds.

6. Global Genome Mining for RiPPs

The mining of structure-modifying enzymes is especially useful for RiPP genome mining. RiPP BGCs consist of genes encoding precursor peptides and tailoring enzymes that catalyze post-translational modifications (PTMs) on the precursors, transforming them into mature peptide structures [82,85]. In contrast to nonribosomal peptides and polyketide synthetases, RiPPs have no universally conserved enzymes, enzyme domains, or other genetic features [64,86], but many recognized RiPP classes or families, such as lanthipeptide, sactipeptide, thiopeptide, and lasso peptide, have class-specific core enzymes and precursor peptide sequence motifs [82]. For genome-mining studies of RiPPs, the ribosomal origin of their precursor peptides facilitates the reliable prediction of their core structure solely on the basis of gene sequence [87,88]. Using the core tailoring enzymes of known RiPP classes to search for novel RiPPs is essentially the same approach as those described in the previous section on structure-guided genome mining. However, RiPP precursor peptides provide a challenge because their sequences are usually short and variable. RiPP precursors are therefore often unannotated in BLAST (see Reference [89], as an example). Hence, many genome-mining tools and software packages, including RODEO [89], RiPPER [64], and antiSMASH [22,23], have been designed specifically to identify these short precursor sequences.

Many RiPP genome-mining approaches start with mining one or more core enzymes of an RiPP class, especially a class-specific tailoring enzyme, and then they search the surrounding genomic context for putative precursor sequences. An example of this approach to RiPP mining is the software RODEO, a tool that has since been incorporated into antiSMASH. RODEO characterizes the surrounding genomic context of a queried RiPP tailoring enzyme. It uses pHMMs to identify possible additional RiPP biosynthesis enzymes, followed by a precursor peptide search using RODEO's ORF detection. The putative RiPP tailoring genes or precursors can then be clustered and analyzed, e.g., with a sequence similarity network. With this approach, the developers of RODEO have conducted searches for novel NPs from lanthipeptide [60], sactipeptide [57], thiopeptide [63], and lasso peptide [89] classes.

After clustering, most of these studies selected candidates from the resultant hits according to interesting or novel structures. They also considered the availability of the source strains. This resulted in the isolation of novel members of the respective RiPP classes (Figure 5). In their lasso peptide study, Tietz et al. [89] investigated candidates with novel predicted topologies, producing six novel lasso peptides and discovering novel and rare chemical modifications on lasso peptides—a new subclass featuring a novel topological position of a disulfide bond, and a lasso peptide that incorporated a citrulline. In their lanthipeptide study, Walker et al. [60] successfully produced a two-component lanthipeptide, birimositide, from *Strepotmyces rimosus* [90]. Hudson et al. [57] chose a candidate sactipeptide that had been previously bioinformatically predicted to have an interesting chemical structure and produced thuricin Z/huazacin, the fifth known sactipeptide [81,91].

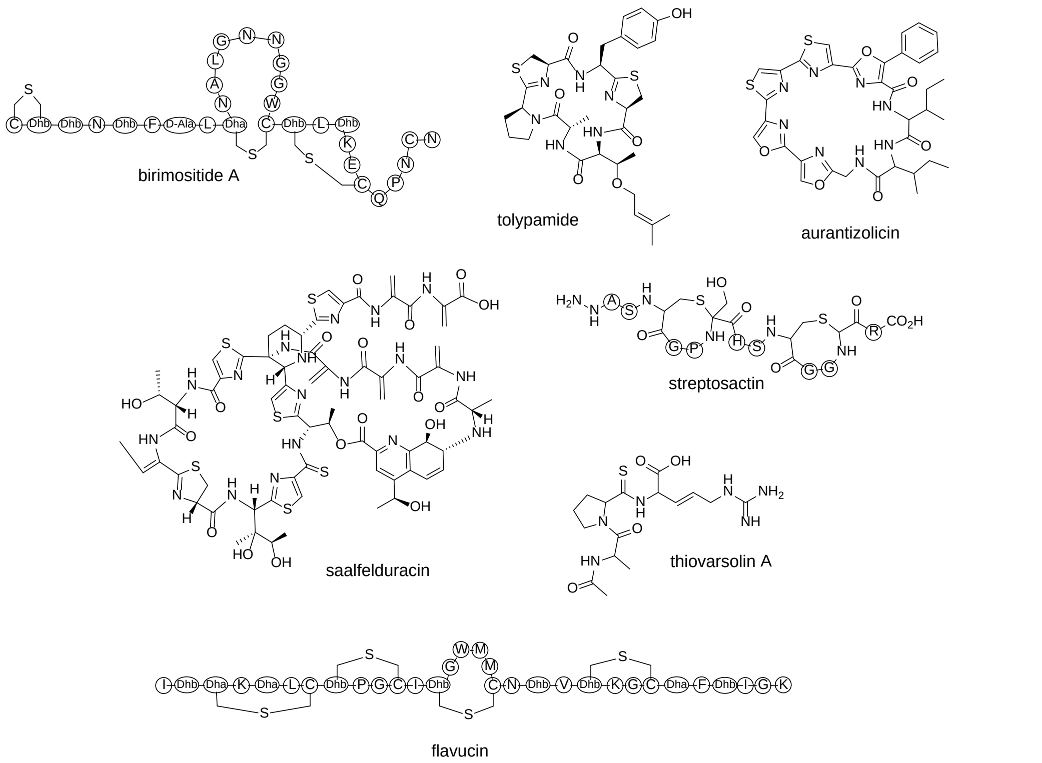

Figure 5. Novel RiPPs identified through large-scale genome mining with unusual biosynthesis and chemical moieties: birimositide *Streptomyces rimosus* subsp. *rimosus* WC3908, tolypamide from *Tolypothrix* sp. PCC 7601, aurantizolicin from *Streptomyces auranticaus* JA 4570, saalfelduracin from *Amycolatopsis saalfeldensis* NRRL B-24474, streptosactin from *Streptococcus thermophilus* JIM 8232, thiovarsolin A from *Streptomyces varsoviensis*, and flavucin from *Corynebacterium lipophiloflavum* DSM 44291.

As demonstrated by these studies, a working hypothesis is not required for candidate selection post-mining. Candidates can be selected simply on the basis of what appears rare, interesting, or unique, or vice versa, those which most resemble previously discovered compounds (see Reference [92], for example). In 2016, Skinnider et al. [62] mined for all known RiPP classes from 65,421 prokaryotic genomes and selected the rarest RiPP class from their clustering results, which had the least number of predicted members and only three documented members. This rare class was the YM-216391 family, with all three of the known members showing cytotoxicity activity at nanomolar levels and azole-rich macrocyclic structures. They isolated a new member of this rare class, named aurantizolcin.

In their thiopeptide study, Schwalen et al. [63] further prioritized their putative thiopeptide hits post-RODEO. They targeted BGCs with a co-incidence of two enzymes—YcaO, which is a core enzyme of the thiopeptide class, and TfuA. TfuA–YcaO coincidence is required for a thioamidation PTM in *Methanosarcina acetivorans* [93], and was also found responsible for the thioamidated backbones of thioamitides [82,94–96]. Schwalen et al. examined a TfuA-like protein that appeared in 2% of their putative thiopeptide BGCs and selected a candidate from *Amycolatopsis saalfeldensis* NRRL B-24474; they successfully extracted the novel thiopeptide saalfelduracin that possesses thioamidation PTM.

Santos-Aberturas et al. [64] directly used the aforementioned TfuA–YcaO coincidence to search for thioamidated RiPPs. They retrieved all the TfuA domain proteins from the actinobacterial phylum in the NCBI non-redundant protein sequence database, explored the genomic area around the *tfuA*-like gene for the *ycaO* gene, and used the genome mining tool RiPPer to search for precursor peptides. After constructing a peptide-similarity network from their hits, they focused on one orphan group and produced thioamidated peptides, named as thiovarsolins, from *Streptomyces varsoviensis*.

Besides core enzymes and precursor sequences, other features in a BGC, such as operons, can be targeted. Bushin et al. [65] aimed to identify BGCs that featured streptide BGC-like rSAMs and were also under the control of *shp*/*rgg* quorum-sensing operons. Using a dataset of 2785 streptococcal genomes, they manually curated 592 rSAM-RiPP gene clusters and grouped the BGCs by the enzyme sequence similarity of rSAM. They functionally characterized members from the generated enzyme-similarity network and revealed unprecedented structural modifications possible on RiPPs, such as the formation of a tetrahydro [5,6] benzindole motif [97]; intramolecular β-thioether linkages [98]; aliphatic ether crosslinks [99]; arginine to tyrosine [100] and lysine to tryptophan [101] cross-links; and multiple sactionine macrocycles, such as in streptosactin [65].

Purushothaman et al. [61] appeared to skip the in silico precursor sequence mining process and instead used the tailoring enzyme of interest to narrow down their candidates. They were targeting a tailoring enzyme found in some cyanobactin BGCs, F-type prenyltransferases, and constructed a sequence similarity network from the entire InterPro family of around 100 cyanobactin prenyltransferases. Of these prenyltransferases, they selected a candidate TolF after constructing a Bayesian phylogenetic tree with known cyanobactin prenyltransferases. The Bayesian tree revealed that TolF was evolutionarily related to known cyanobactin prenyltransferase TruF1, but with only ~48% sequence similarity. The prospect of a functionally divergent TruF1-like prenyltransferase was intriguing because TruF1 catalyzed interesting reactions but was difficult to be characterized due to low solubility and low in vitro activity. They identified the other genes within the BGC of TolF and successfully produced novel cyanobactin tolypamide, in addition to a stable TolF prenyltransferase. Characterizing TolF revealed that it differed from TruF1 in the way it prenylates in the forward direction instead of reverse, thus granting researchers mechanistic insight into Ser- and Thr-prenylating enzymes when TruF1 was unavailable for extensive study [61].

RiPP BGCs tend to consist of a few genes. Some RiPP tailoring enzymes are substrate promiscuous, a phenomenon that some researchers have exploited. Van Heel et al. [66] selected lanthipeptide candidate BGCs that closely resembled those of the model lanthipeptide nisin. It was previously found that, when some lanthipeptide (or even non-lantipeptide) core peptides are fused to the nisin leader peptide, the hybrid lanthipeptide precursor peptide was accepted by the nisin biosynthesis system as a substrate, resulting in such hybrid BGCs producing non-nisin lanthipeptides [102]. Van Heel et al. wanted to exploit the promiscuity of nisin tailoring enzymes to engineer a convenient heterologous production system for producing novel lanthipeptides. They selected lanthipeptide candidate BGCs with only the same tailoring enzymes as nisin BGCs and successfully produced 31 peptides by using the heterologous production system. Moreover, five of the proteins showed antimicrobial activity, such as flavucin and agalacticin.

7. Conclusions

With the advancements in genome-mining technologies and development of the big data field, global genome mining for NP discovery has attracted considerable interest. The resultant large number of hits in a large-scale study often requires researchers to prioritize candidates for downstream investigations, because it is beyond the capabilities of many laboratories to screen every candidate. This review covers diverse approaches and tools for prioritizing candidates through global genome mining and discusses their advantages and limitations.

We introduced the studies that used resistance genes, phylogenomics, and NP chemical structures to guide the discovery of novel BGCs in large-scale genome mining. Resistance determinants are not found in all NP BGCs, but prioritizing candidate BGCs associated with resistance determinants is nevertheless a way to detect novel compounds. The approach of using a particular self-resistance enzyme or motif as molecular "beacons" has led to discoveries of NPs that putatively target the respective self-resistance enzymes [40,41]. Two different approaches that requires no prior knowledge of a resistance enzyme are to search for gene duplications and to search for cooccurrence of transport-related enzymes [42,44], as these phenomena suggest the possible presence of self-resistance mechanisms. The gene-duplication logic can be useful in identifying compounds with a specific bioactivity in mind, as the NP's function is presumably related to that of the duplicated gene.

Phylogenomic-guided genome mining can search for divergent enzymes that have possibly evolved from central metabolism to catalyze novel chemical transformations in secondary metabolites. An annotated, functionally characterized enzyme is prerequisite to identifying its divergent clades in a phylogenetic tree [48,51]. However, prior knowledge of a particular enzyme is not necessary in some approaches. Whole BGCs can be phylogenetically analyzed without targeting for specific biosynthesis enzymes, such as by using GCF tools such as BiG-SCAPE [51]. Both resistance-gene-based logic and phylogenomic identification of divergent enzymes can be applied with and without prior knowledge of a particular enzyme.

If the chemical structure of an NP family is known, genome mining can be guided by specific structural features, particularly those in NPs with unprecedented chemistries or potent bioactivities. The presence of specific chemical moieties known to enhance bioactivities, such as the cationic amino acids and thioamides mentioned in this review [54,63], can be directly used for candidate prioritization. Targeting a biosynthesis enzyme with high substrate promiscuity or the ability to catalyze different reactions can be beneficial, as shown in the studies on nisin lanthipeptide heterologous expression [66], cytochrome P450 [51,55], and radical SAMs [57,65].

These studies show that, beyond sequence-similarity-based mining, genes can be targeted according to hypotheses about their functions, evolution, and co-evolution with other genes. The targeted item can also be something other than proteins or peptides, such as the quorum-sensing operons [50,65]. However, a solid hypothesis is not always a prerequisite for starting a large-scale search, because serendipitous discoveries from analyzing a large pool of candidates can be used to generate new hypotheses. An unexpected association of a candidate's biosynthesis enzymes with known enzymes can prompt its selection, because characterizing it may bring insight into biosynthesis enzyme mechanisms. Selecting putative hits with novel chemical structures or BGC features from global genome mining can expand our understanding of biosynthesis processes and aid in the discovery of novel NP class motifs. Regardless, multiple hypotheses and approaches can be used concurrently to maximize the likelihood of identifying BGCs of interest.

8. Future Perspective

There are still many possibilities in the design of NP global genome mining studies in the software and in the overall pipeline. Novel iterative tools to deal with the extensive data from global genome mining have been created, such as BiG-SLiCE [103] and the BiG-FAM database [104], being continuations from the work of BiG-SCAPE, to cluster BGCs into families at faster speeds. The IMG-ABC and MIBiG databases have been updated periodically in recent years, with the former using the newer antiSMASH v5 to replace the previous BGC predictions and the latter applying manual curation to improve their respective schemas and data [6,77]. ClusterFinder has been removed as a default analysis option in antiSMASH v5 onward, and, thus, false positive BGCs resulting from it that are highly unlikely to produce secondary metabolites will not be detected by using the default settings [105]. BGCs detected by ClusterFinder has also been removed from the IMG-ABC database; thus, it should more accurately represent valid predicted BGCs [75–77].

Databases specific to different modes of genome mining have also become available. The ARTS-DB contains a repository of putative resistance genes detected by ARTS. ARTS, as a standalone application, can also be used to screen for resistant housekeeping genes in genomes from all bacterial taxa based on gene duplication, horizontal gene transfer (HGT) events, and colocalization with a BGC [106]. To date, ARTS-DB contains putative resistance genes identified through ARTS from more than 70,000 genomes and metagenome-assembled genomes [107]. For phylogenomics-guided genome mining, EvoMining is a tool that allows users to visualize expansion and recruitment events in enzyme families through phylogenetic reconstruction. Users can use their own enzyme, genome, and BGC databases. However, pre-computed databases are also available [71]. The accessibility of different NP genome-mining-specific databases and the availability of the different pipelines to study large-scale BGC data contribute greatly to the performance of global-genome-mining studies. Artificial intelligence (AI) technologies, such as deep neural networks, can offer new perspectives in data mining [108,109].

Besides the genome-mining process, automating other parts of the NP discovery pipeline will make global studies more viable. One common hurdle, again due to the size of the data, is difficulties in detecting the novel compound. The automation of RiPP detection by computationally predicting possible NP structures and fragments has been attempted. This is achieved by generating a library of possible MS spectra of the desired RiPP, from which MS peaks of bacterial culture can be matched to. In a prior-mentioned 2016 study by Skinnider et al. [62], the team conducted gene-guided prediction of their candidate compound's structure, which allowed them to automate the LC–MS/MS peak searching process by using ion mass predictions based on the predicted structure. Other examples of automated RiPP detection efforts are RiPPquest, pep2path, and pepSAVI-MS [110–112]. Automated peptide matching is commonly used in proteomics [113], illustrating how existing methods can be applied to the global genome mining of NPs when being optimized for increased speed, reduced computational load, and high effectiveness.

The age-old bottleneck in the NP discovery pipeline of producing and extracting the NP from bacterial culture has led to the development of various alternative avenues to overcome it. One solution is to synthesize the bioinformatically predicted peptide artificially [114,115]. In RiPP discovery, the promiscuity of tailoring enzymes has been exploited to catalyze chemical transformations on novel precursor peptides, pairing well-expressed tailoring enzymes from characterized BGCs with the novel precursor peptides of interest. This hybrid strategy has been successful in both in vivo and in vitro expression: for in vivo expression, hybrid BGCs containing genes from different BGCs are constructed and expressed [66]; and for in vitro production, the final product is synthesized by in vitro reactions, using purified enzymes [116]. The genetic cloning process for BGCs is another part of the pipeline that can be tackled. After selecting specific BGCs and biosynthesis enzyme genes through global genome mining, the gene can be artificially synthesized. This can bypass the efforts in obtaining and isolating the source microbial strain and enable immediate heterologous expression that simplifies the validation of their biochemical functions. This is particularly useful for sequences derived from metagenome assemblies without strain isolation, as well as for sequences of less-studied bacterial taxa where codon optimization for heterologous expression is being considered.

The characteristics and functions of the source strains that produce bioactive natural products are important information in genome-mining-based discovery of bioactive compounds. For example, thermophilic bacteria may produce more thermostable target peptides or proteins, and, thus, multiple laboratories have prioritized candidate BGCs from thermophilic bacteria, which were achieved through a filtering step post-mining [116–118] or only mining from thermophilic bacteria [119]. In addition, candidate BGCs are frequently selected from bacterial strains or organisms that are little-known to avoid NP rediscovery. On the other hand, bacteria that are well-known to be rich natural product producers [86], or bacterial families known to produce a particular type of NP [54], are often the targets for more intensive mining. Well-known taxa which enjoy well-established genetic engineering tools are also favorable for downstream work. One can even combine these two principles in selecting rare and conventional bacteria, as illustrated by Chevrette et al. [120] pursuing *Streptomyces* found on insects, a popular NP powerhouse genus in a little-studied ecological niche. Such biological awareness of the strain can aid in BGC prioritization indirectly.

Author Contributions: Conceptualization, writing—original draft preparation, and writing—review and editing, J.J.L.M., H.Y.C.L., and P.-Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (2018YFA0903200), Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangdong, SMSEGL20SC01), Major Project of Basic and Applied Basic Research of Guangdong Province (2019B030302004), Key Special Project for Introduced Talents Team of the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (GML2019ZD409), and a CRF grant from the HKSAR government (C6026-19G-A).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Chuanhai Wu for his valuable comments and suggestions on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Waglechner, N.; McArthur, A.G.; Wright, G.D. Phylogenetic Reconciliation Reveals the Natural History of Glycopeptide Antibiotic Biosynthesis and Resistance. *Nat. Microbiol.* **2019**, *4*, 1862–1871. [CrossRef] [PubMed]
2. Zhang, M.M.; Wang, Y.; Lui Ang, E.; Zhao, H. Engineering Microbial Hosts for Production of Bacterial Natural Products. *Nat. Prod. Rep.* **2016**, *33*, 963–987. [CrossRef] [PubMed]
3. Hug, J.J.; Krug, D.; Müller, R. Bacteria as Genetically Programmable Producers of Bioactive Natural Products. *Nat. Rev. Chem.* **2020**, *4*, 172–193. [CrossRef]
4. Breitling, R.; Ceniceros, A.; Jankevics, A.; Takano, E. Metabolomics for Secondary Metabolite Research. *Metabolites* **2013**, *3*, 1076–1083. [CrossRef]
5. Newman, D.J.; Cragg, G.M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770–803. [CrossRef]
6. Kautsar, S.A.; Blin, K.; Shaw, S.; Navarro-Muñoz, J.C.; Terlouw, B.R.; van der Hooft, J.J.J.; van Santen, J.A.; Tracanna, V.; Suarez Duran, H.G.; Pascal Andreu, V.; et al. MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function. *Nucleic Acids Res.* **2020**, *48*, D454–D458. [CrossRef]
7. O'Neill, E.C.; Schorn, M.; Larson, C.B.; Millán-Aguiñaga, N. Targeted Antibiotic Discovery through Biosynthesis-Associated Resistance Determinants: Target Directed Genome Mining. *Crit. Rev. Microbiol.* **2019**, *45*, 255–277. [CrossRef]
8. Schmidt, E.W. Trading Molecules and Tracking Targets in Symbiotic Interactions. *Nat. Chem. Biol.* **2008**, *4*, 466–473. [CrossRef]
9. Walsh, C.T.; Tang, Y. *Natural Product Biosynthesis*; Royal Society of Chemistry: London, UK, 2017; ISBN 978-1-78801-131-0.
10. Katz, L.; Baltz, R.H. Natural Product Discovery: Past, Present, and Future. *J. Ind. Microbiol. Biotechnol.* **2016**, *43*, 155–176. [CrossRef]

11. Luo, Y.; Cobb, R.E.; Zhao, H. Recent Advances in Natural Product Discovery. *Curr. Opin. Biotechnol.* **2014**, *30*, 230–237. [CrossRef]
12. Genilloud, O.; González, I.; Salazar, O.; Martín, J.; Tormo, J.R.; Vicente, F. Current Approaches to Exploit Actinomycetes as a Source of Novel Natural Products. *J. Ind. Microbiol. Biotechnol.* **2011**, *38*, 375–389. [CrossRef] [PubMed]
13. Wohlleben, W.; Mast, Y.; Stegmann, E.; Ziemert, N. Antibiotic Drug Discovery. *Microb. Biotechnol.* **2016**, *9*, 541–548. [CrossRef] [PubMed]
14. Reen, F.J.; Romano, S.; Dobson, A.D.W.; O'Gara, F. The Sound of Silence: Activating Silent Biosynthetic Gene Clusters in Marine Microorganisms. *Mar. Drugs* **2015**, *13*, 4754–4783. [CrossRef] [PubMed]
15. Bachmann, B.O.; Van Lanen, S.G.; Baltz, R.H. Microbial Genome Mining for Accelerated Natural Products Discovery: Is a Renaissance in the Making? *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 175–184. [CrossRef]
16. Ziemert, N.; Alanjary, M.; Weber, T. The Evolution of Genome Mining in Microbes—A Review. *Nat. Prod. Rep.* **2016**, *33*, 988–1005. [CrossRef]
17. Weber, T.; Welzel, K.; Pelzer, S.; Vente, A.; Wohlleben, W. Exploiting the Genetic Potential of Polyketide Producing Streptomycetes. *J. Biotechnol.* **2003**, *106*, 221–232. [CrossRef]
18. Lee, N.; Hwang, S.; Kim, J.; Cho, S.; Palsson, B.; Cho, B.-K. Mini Review: Genome Mining Approaches for the Identification of Secondary Metabolite Biosynthetic Gene Clusters in Streptomyces. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1548–1556. [CrossRef]
19. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
20. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37. [CrossRef]
21. Zerikly, M.; Challis, G.L. Strategies for the Discovery of New Natural Products by Genome Mining. *ChemBioChem* **2009**, *10*, 625–633. [CrossRef]
22. Medema, M.H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M.A.; Weber, T.; Takano, E.; Breitling, R. AntiSMASH: Rapid Identification, Annotation and Analysis of Secondary Metabolite Biosynthesis Gene Clusters in Bacterial and Fungal Genome Sequences. *Nucleic Acids Res.* **2011**, *39*, W339–W346. [CrossRef] [PubMed]
23. Blin, K.; Shaw, S.; Kloosterman, A.M.; Charlop-Powers, Z.; van Wezel, G.P.; Medema, M.H.; Weber, T. AntiSMASH 6.0: Improving Cluster Detection and Comparison Capabilities. *Nucleic Acids Res.* **2021**, *49*, W29–W35. [CrossRef] [PubMed]
24. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H.U.; Bruccoleri, R.; Lee, S.Y.; Fischbach, M.A.; Müller, R.; Wohlleben, W.; et al. AntiSMASH 3.0—a Comprehensive Resource for the Genome Mining of Biosynthetic Gene Clusters. *Nucleic Acids Res.* **2015**, *43*, W237–W243. [CrossRef] [PubMed]
25. Skinnider, M.A.; Johnston, C.W.; Gunabalasingam, M.; Merwin, N.J.; Kieliszek, A.M.; MacLellan, R.J.; Li, H.; Ranieri, M.R.M.; Webster, A.L.H.; Cao, M.P.T.; et al. Comprehensive Prediction of Secondary Metabolite Structure and Biological Activity from Microbial Genome Sequences. *Nat. Commun.* **2020**, *11*, 6058. [CrossRef]
26. Sugimoto, Y.; Camacho, F.R.; Wang, S.; Chankhamjon, P.; Odabas, A.; Biswas, A.; Jeffrey, P.D.; Donia, M.S. A Metagenomic Strategy for Harnessing the Chemical Repertoire of the Human Microbiome. *Science* **2019**, *366*, eaax9176. [CrossRef]
27. Weber, T.; Kim, H.U. The Secondary Metabolite Bioinformatics Portal: Computational Tools to Facilitate Synthetic Biology of Secondary Metabolite Production. *Synth. Syst. Biotechnol.* **2016**, *1*, 69–79. [CrossRef]
28. Sayers, E.W.; Bolton, E.E.; Brister, J.R.; Canese, K.; Chan, J.; Comeau, D.C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2022**, *50*, D20–D26. [CrossRef]
29. Hadjithomas, M.; Chen, I.-M.A.; Chu, K.; Huang, J.; Ratner, A.; Palaniappan, K.; Andersen, E.; Markowitz, V.; Kyrpides, N.C.; Ivanova, N.N. IMG-ABC: New Features for Bacterial Secondary Metabolism Analysis and Targeted Biosynthetic Gene Cluster Discovery in Thousands of Microbial Genomes. *Nucleic Acids Res.* **2017**, *45*, D560–D565. [CrossRef]
30. Chevrette, M.G.; Gavrilidou, A.; Mantri, S.; Selem-Mojica, N.; Ziemert, N.; Barona-Gómez, F. The Confluence of Big Data and Evolutionary Genome Mining for the Discovery of Natural Products. *Nat. Prod. Rep.* **2021**, *38*, 2024–2040. [CrossRef]
31. Gerlt, J.A.; Bouvier, J.T.; Davidson, D.B.; Imker, H.J.; Sadkhin, B.; Slater, D.R.; Whalen, K.L. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks. *Biochim. Biophys. Acta-Proteins Proteom.* **2015**, *1854*, 1019–1037. [CrossRef]
32. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef] [PubMed]
33. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
34. Machado, H.; Tuttle, R.N.; Jensen, P.R. Omics-Based Natural Product Discovery and the Lexicon of Genome Mining. *Curr. Opin. Microbiol.* **2017**, *39*, 136–142. [CrossRef] [PubMed]
35. Palazzotto, E.; Weber, T. Omics and Multi-Omics Approaches to Study the Biosynthesis of Secondary Metabolites in Microorganisms. *Curr. Opin. Microbiol.* **2018**, *45*, 109–116. [CrossRef]
36. Kloosterman, A.M.; Medema, M.H.; van Wezel, G.P. Omics-Based Strategies to Discover Novel Classes of RiPP Natural Products. *Curr. Opin. Biotechnol.* **2021**, *69*, 60–67. [CrossRef]
37. Prihoda, D.; Maritz, J.M.; Klempir, O.; Dzamba, D.; Woelk, C.H.; Hazuda, D.J.; Bitton, D.A.; Hannigan, G.D. The Application Potential of Machine Learning and Genomics for Understanding Natural Product Diversity, Chemistry, and Therapeutic Translatability. *Nat. Prod. Rep.* **2021**, *38*, 1100–1108. [CrossRef]

38. Zhong, Z.; He, B.; Li, J.; Li, Y.-X. Challenges and Advances in Genome Mining of Ribosomally Synthesized and Post-Translationally Modified Peptides (RiPPs). *Synth. Syst. Biotechnol.* **2020**, *5*, 155–172. [CrossRef]
39. Hug, J.J.; Panter, F.; Krug, D.; Müller, R. Genome Mining Reveals Uncommon Alkylpyrones as Type III PKS Products from Myxobacteria. *J. Ind. Microbiol. Biotechnol.* **2019**, *46*, 319–334. [CrossRef]
40. Panter, F.; Krug, D.; Baumann, S.; Müller, R. Self-Resistance Guided Genome Mining Uncovers New Topoisomerase Inhibitors from Myxobacteria. *Chem. Sci.* **2018**, *9*, 4898–4908. [CrossRef]
41. Yan, Y.; Liu, Q.; Zang, X.; Yuan, S.; Bat-Erdene, U.; Nguyen, C.; Gan, J.; Zhou, J.; Jacobsen, S.E.; Tang, Y. Resistance-Gene-Directed Discovery of a Natural-Product Herbicide with a New Mode of Action. *Nature* **2018**, *559*, 415–418. [CrossRef]
42. Girard, L.; Geudens, N.; Pauwels, B.; Höfte, M.; Martins, J.C.; De Mot, R. Transporter Gene-Mediated Typing for Detection and Genome Mining of Lipopeptide-Producing Pseudomonas. *Appl. Environ. Microbiol.* **2022**, *88*, e01869-21. [CrossRef] [PubMed]
43. Liu, N.; Abramyan, E.D.; Cheng, W.; Perlatti, B.; Harvey, C.J.B.; Bills, G.F.; Tang, Y. Targeted Genome Mining Reveals the Biosynthetic Gene Clusters of Natural Product CYP51 Inhibitors. *J. Am. Chem. Soc.* **2021**, *143*, 6043–6047. [CrossRef] [PubMed]
44. Tang, X.; Li, J.; Millán-Aguiñaga, N.; Zhang, J.J.; O'Neill, E.C.; Ugalde, J.A.; Jensen, P.R.; Mantovani, S.M.; Moore, B.S. Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-Directed Genome Mining. *ACS Chem. Biol.* **2015**, *10*, 2841–2849. [CrossRef] [PubMed]
45. Li, Y.-X.; Zhong, Z.; Hou, P.; Zhang, W.-P.; Qian, P.-Y. Resistance to Nonribosomal Peptide Antibiotics Mediated by d-Stereospecific Peptidases. *Nat. Chem. Biol.* **2018**, *14*, 381–387. [CrossRef]
46. Cimermancic, P.; Medema, M.H.; Claesen, J.; Kurita, K.; Wieland Brown, L.C.; Mavrommatis, K.; Pati, A.; Godfrey, P.A.; Koehrsen, M.; Clardy, J.; et al. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **2014**, *158*, 412–421. [CrossRef]
47. Cruz-Morales, P.; Kopp, J.F.; Martínez-Guerrero, C.; Yáñez-Guerra, L.A.; Selem-Mojica, N.; Ramos-Aboites, H.; Feldmann, J.; Barona-Gómez, F. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol. Evol.* **2016**, *8*, 1906–1916. [CrossRef]
48. Culp, E.J.; Waglechner, N.; Wang, W.; Fiebig-Comyn, A.A.; Hsu, Y.-P.; Koteva, K.; Sychantha, D.; Coombes, B.K.; Van Nieuwenhze, M.S.; Brun, Y.V.; et al. Evolution-Guided Discovery of Antibiotics That Inhibit Peptidoglycan Remodelling. *Nature* **2020**, *578*, 582–587. [CrossRef]
49. Ahmed, M.N.; Reyna-González, E.; Schmid, B.; Wiebach, V.; Süssmuth, R.D.; Dittmann, E.; Fewer, D.P. Phylogenomic Analysis of the Microviridin Biosynthetic Pathway Coupled with Targeted Chemo-Enzymatic Synthesis Yields Potent Protease Inhibitors. *ACS Chem. Biol.* **2017**, *12*, 1538–1546. [CrossRef]
50. Mullins, A.J.; Murray, J.A.H.; Bull, M.J.; Jenner, M.; Jones, C.; Webster, G.; Green, A.E.; Neill, D.R.; Connor, T.R.; Parkhill, J.; et al. Genome Mining Identifies Cepacin as a Plant-Protective Metabolite of the Biopesticidal Bacterium Burkholderia Ambifaria. *Nat. Microbiol.* **2019**, *4*, 996–1005. [CrossRef]
51. Navarro-Muñoz, J.C.; Selem-Mojica, N.; Mullowney, M.W.; Kautsar, S.A.; Tryon, J.H.; Parkinson, E.I.; De Los Santos, E.L.C.; Yeong, M.; Cruz-Morales, P.; Abubucker, S.; et al. A Computational Framework to Explore Large-Scale Biosynthetic Diversity. *Nat. Chem. Biol.* **2020**, *16*, 60–68. [CrossRef]
52. Yamada, Y.; Kuzuyama, T.; Komatsu, M.; Shin-ya, K.; Omura, S.; Cane, D.E.; Ikeda, H. Terpene Synthases Are Widely Distributed in Bacteria. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 857–862. [CrossRef] [PubMed]
53. Chen, S.; Zhang, C.; Zhang, L. Investigation of the Molecular Landscape of Bacterial Aromatic Polyketides by Global Analysis of Type II Polyketide Synthases. *Angew. Chem. Int. Ed.* **2022**, *61*, e202202286. [CrossRef]
54. Li, Y.-X.; Zhong, Z.; Zhang, W.-P.; Qian, P.-Y. Discovery of Cationic Nonribosomal Peptides as Gram-Negative Antibiotics through Global Genome Mining. *Nat. Commun.* **2018**, *9*, 3273. [CrossRef] [PubMed]
55. Malit, J.J.L.; Liu, W.; Cheng, A.; Saha, S.; Liu, L.-L.; Qian, P.-Y. Global Genome Mining Reveals a Cytochrome P450-Catalyzed Cyclization of Crownlike Cyclodipeptides with Neuroprotective Activity. *Org. Lett.* **2021**, *23*, 6601–6605. [CrossRef]
56. Malit, J.J.L.; Wu, C.; Tian, X.; Liu, W.; Huang, D.; Sung, H.H.-Y.; Liu, L.-L.; Williams, I.D.; Qian, P.-Y. Griseocazines: Neuroprotective Multiprenylated Cyclodipeptides Identified through Targeted Genome Mining. *Org. Lett.* **2022**, *24*, 2967–2972. [CrossRef]
57. Hudson, G.A.; Burkhart, B.J.; DiCaprio, A.J.; Schwalen, C.J.; Kille, B.; Pogorelov, T.V.; Mitchell, D.A. Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New Cα, Cβ, and Cγ-Linked Thioether-Containing Peptides. *J. Am. Chem. Soc.* **2019**, *141*, 8228–8238. [CrossRef]
58. Pan, G.; Xu, Z.; Guo, Z.; Hindra; Ma, M.; Yang, D.; Zhou, H.; Gansemans, Y.; Zhu, X.; Huang, Y.; et al. Discovery of the Leinamycin Family of Natural Products by Mining Actinobacterial Genomes. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E11131–E11140. [CrossRef]
59. Ju, K.-S.; Gao, J.; Doroghazi, J.R.; Wang, K.-K.A.; Thibodeaux, C.J.; Li, S.; Metzger, E.; Fudala, J.; Su, J.; Zhang, J.K.; et al. Discovery of Phosphonic Acid Natural Products by Mining the Genomes of 10,000 Actinomycetes. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12175–12180. [CrossRef]
60. Walker, M.C.; Eslami, S.M.; Hetrick, K.J.; Ackenhusen, S.E.; Mitchell, D.A.; van der Donk, W.A. Precursor Peptide-Targeted Mining of More than One Hundred Thousand Genomes Expands the Lanthipeptide Natural Product Family. *BMC Genom.* **2020**, *21*, 387. [CrossRef]
61. Purushothaman, M.; Sarkar, S.; Morita, M.; Gugger, M.; Schmidt, E.W.; Morinaka, B.I. Genome-Mining-Based Discovery of the Cyclic Peptide Tolypamide and TolF, a Ser/Thr Forward O-Prenyltransferase. *Angew. Chem. Int. Ed.* **2021**, *60*, 8460–8465. [CrossRef]

62. Skinnider, M.A.; Johnston, C.W.; Edgar, R.E.; Dejong, C.A.; Merwin, N.J.; Rees, P.N.; Magarvey, N.A. Genomic Charting of Ribosomally Synthesized Natural Product Chemical Space Facilitates Targeted Mining. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E6343. [CrossRef] [PubMed]
63. Schwalen, C.J.; Hudson, G.A.; Kille, B.; Mitchell, D.A. Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics. *J. Am. Chem. Soc.* **2018**, *140*, 9494–9501. [CrossRef] [PubMed]
64. Santos-Aberturas, J.; Chandra, G.; Frattaruolo, L.; Lacret, R.; Pham, T.H.; Vior, N.M.; Eyles, T.H.; Truman, A.W. Uncovering the Unexplored Diversity of Thioamidated Ribosomal Peptides in Actinobacteria Using the RiPPER Genome Mining Tool. *Nucleic Acids Res.* **2019**, *47*, 4624–4637. [CrossRef]
65. Bushin, L.B.; Covington, B.C.; Rued, B.E.; Federle, M.J.; Seyedsayamdost, M.R. Discovery and Biosynthesis of Streptosactin, a Sactipeptide with an Alternative Topology Encoded by Commensal Bacteria in the Human Microbiome. *J. Am. Chem. Soc.* **2020**, *142*, 16265–16275. [CrossRef] [PubMed]
66. Van Heel, A.J.; Kloosterman, T.G.; Montalban-Lopez, M.; Deng, J.; Plat, A.; Baudu, B.; Hendriks, D.; Moll, G.N.; Kuipers, O.P. Discovery, Production and Modification of Five Novel Lantibiotics Using the Promiscuous Nisin Modification Machinery. *ACS Synth. Biol.* **2016**, *5*, 1146–1154. [CrossRef]
67. Kling, A.; Lukat, P.; Almeida, D.V.; Bauer, A.; Fontaine, E.; Sordello, S.; Zaburannyi, N.; Herrmann, J.; Wenzel, S.C.; König, C.; et al. Targeting DnaN for Tuberculosis Therapy Using Novel Griselimycins. *Science* **2015**, *348*, 1106–1112. [CrossRef]
68. Kale, A.J.; McGlinchey, R.P.; Lechner, A.; Moore, B.S. Bacterial Self-Resistance to the Natural Proteasome Inhibitor Salinosporamide A. *ACS Chem. Biol.* **2011**, *6*, 1257–1264. [CrossRef]
69. Peterson, R.M.; Huang, T.; Rudolf, J.D.; Smanski, M.J.; Shen, B. Mechanisms of Self-Resistance in the Platensimycin and Platencin Producing Streptomyces Platensis MA7327 and MA7339 Strains. *Chem. Biol.* **2014**, *21*, 389–397. [CrossRef]
70. Amorim Franco, T.M.; Blanchard, J.S. Bacterial Branched-Chain Amino Acid Biosynthesis: Structures, Mechanisms, and Drugability. *Biochemistry* **2017**, *56*, 5849–5865. [CrossRef]
71. Sélem-Mojica, N.; Aguilar, C.; Gutiérrez-García, K.; Martínez-Guerrero, C.E.; Barona-Gómez, F. EvoMining Reveals the Origin and Fate of Natural Product Biosynthetic Enzymes. *Microb. Genom.* **2019**, *5*, e000260. [CrossRef]
72. Alcock, B.P.; Raphenya, A.R.; Lau, T.T.Y.; Tsang, K.K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A.-L.V.; Cheng, A.A.; Liu, S.; et al. CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **2020**, *48*, D517–D525. [CrossRef] [PubMed]
73. Gibson, M.K.; Forsberg, K.J.; Dantas, G. Improved Annotation of Antibiotic Resistance Determinants Reveals Microbial Resistomes Cluster by Ecology. *ISME J.* **2015**, *9*, 207–216. [CrossRef] [PubMed]
74. Medema, M.H.; Fischbach, M.A. Computational Approaches to Natural Product Discovery. *Nat. Chem. Biol.* **2015**, *11*, 639–648. [CrossRef] [PubMed]
75. Baltz, R.H. Natural Product Drug Discovery in the Genomic Era: Realities, Conjectures, Misconceptions, and Opportunities. *J. Ind. Microbiol. Biotechnol.* **2019**, *46*, 281–299. [CrossRef] [PubMed]
76. Baltz, R.H. Genome Mining for Drug Discovery: Progress at the Front End. *J. Ind. Microbiol. Biotechnol.* **2021**, *48*, kuab044. [CrossRef]
77. Palaniappan, K.; Chen, I.-M.A.; Chu, K.; Ratner, A.; Seshadri, R.; Kyrpides, N.C.; Ivanova, N.N.; Mouncey, N.J. IMG-ABC v.5.0: An Update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D422–D430. [CrossRef]
78. Komatsu, M.; Tsuda, M.; Ōmura, S.; Oikawa, H.; Ikeda, H. Identification and Functional Analysis of Genes Controlling Biosynthesis of 2-Methylisoborneol. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 7422–7427. [CrossRef]
79. Harken, L.; Li, S.-M. Modifications of Diketopiperazines Assembled by Cyclodipeptide Synthases with Cytochrome P450 Enzymes. *Appl. Microbiol. Biotechnol.* **2021**, *105*, 2277–2285. [CrossRef]
80. Haft, D.H.; Basu, M.K. Biological Systems Discovery in Silico: Radical S-Adenosylmethionine Protein Families and Their Target Peptides for Posttranslational Modification. *J. Bacteriol.* **2011**, *193*, 2745–2755. [CrossRef]
81. Chen, Y.; Wang, J.; Li, G.; Yang, Y.; Ding, W. Current Advancements in Sactipeptide Natural Products. *Front. Chem.* **2021**, *9*, 595991. [CrossRef]
82. Montalbán-López, M.; Scott, T.A.; Ramesh, S.; Rahman, I.R.; van Heel, A.J.; Viel, J.H.; Bandarian, V.; Dittmann, E.; Genilloud, O.; Goto, Y.; et al. New Developments in RiPP Discovery, Enzymology and Engineering. *Nat. Prod. Rep.* **2021**, *38*, 130–239. [CrossRef] [PubMed]
83. Yeung, A.T.Y.; Gellatly, S.L.; Hancock, R.E.W. Multifunctional Cationic Host Defence Peptides and Their Clinical Applications. *Cell. Mol. Life Sci.* **2011**, *68*, 2161. [CrossRef] [PubMed]
84. Epand, R.M.; Vogel, H.J. Diversity of Antimicrobial Peptides and Their Mechanisms of Action. *Biochim. Biophys. Acta-Biomembr.* **1999**, *1462*, 11–28. [CrossRef]
85. Arnison, P.G.; Bibb, M.J.; Bierbaum, G.; Bowers, A.A.; Bugni, T.S.; Bulaj, G.; Camarero, J.A.; Campopiano, D.J.; Challis, G.L.; Clardy, J.; et al. Ribosomally Synthesized and Post-Translationally Modified Peptide Natural Products: Overview and Recommendations for a Universal Nomenclature. *Nat. Prod. Rep.* **2013**, *30*, 108–160. [CrossRef] [PubMed]
86. Kloosterman, A.M.; Shelton, K.E.; van Wezel, G.P.; Medema, M.H.; Mitchell, D.A. RRE-Finder: A Genome-Mining Tool for Class-Independent RiPP Discovery. *mSystems* **2020**, *5*, 267. [CrossRef] [PubMed]
87. Zhang, Y.; Chen, M.; Bruner, S.D.; Ding, Y. Heterologous Production of Microbial Ribosomally Synthesized and Post-Translationally Modified Peptides. *Front. Microbiol.* **2018**, *9*, 1801. [CrossRef]

88. Malit, J.J.L.; Wu, C.; Liu, L.-L.; Qian, P.-Y. Global Genome Mining Reveals the Distribution of Diverse Thioamidated RiPP Biosynthesis Gene Clusters. *Front. Microbiol.* **2021**, *12*, 987. [CrossRef]
89. Tietz, J.I.; Schwalen, C.J.; Patel, P.S.; Maxson, T.; Blair, P.M.; Tai, H.-C.; Zakai, U.I.; Mitchell, D.A. A New Genome-Mining Tool Redefines the Lasso Peptide Biosynthetic Landscape. *Nat. Chem. Biol.* **2017**, *13*, 470–478. [CrossRef]
90. Singh, M.; Chaudhary, S.; Sareen, D. Roseocin, a Novel Two-Component Lantibiotic from an Actinomycete. *Mol. Microbiol.* **2020**, *113*, 326–337. [CrossRef]
91. Mo, T.; Ji, X.; Yuan, W.; Mandalapu, D.; Wang, F.; Zhong, Y.; Li, F.; Chen, Q.; Ding, W.; Deng, Z.; et al. Thuricin Z: A Narrow-Spectrum Sactibiotic That Targets the Cell Membrane. *Angew. Chem. Int. Ed.* **2021**, *58*, 18793–18797. [CrossRef]
92. Lee, H.; Choi, M.; Park, J.-U.; Roh, H.; Kim, S. Genome Mining Reveals High Topological Diversity of ω-Ester-Containing Peptides and Divergent Evolution of ATP-Grasp Macrocyclases. *J. Am. Chem. Soc.* **2020**, *142*, 3013–3023. [CrossRef] [PubMed]
93. Nayak, D.D.; Mahanta, N.; Mitchell, D.A.; Metcalf, W.W. Post-Translational Thioamidation of Methyl-Coenzyme M Reductase, a Key Enzyme in Methanogenic and Methanotrophic Archaea. *Elife* **2017**, *6*, e29218. [CrossRef] [PubMed]
94. Izawa, M.; Kawasaki, T.; Hayakawa, Y. Cloning and Heterologous Expression of the Thioviridamide Biosynthesis Gene Cluster from Streptomyces Olivoviridis. *Appl. Environ. Microbiol.* **2013**, *79*, 7110–7113. [CrossRef] [PubMed]
95. Kawahara, T.; Izumikawa, M.; Kozone, I.; Hashimoto, J.; Kagaya, N.; Koiwai, H.; Komatsu, M.; Fujie, M.; Sato, N.; Ikeda, H.; et al. Neothioviridamide, a Polythioamide Compound Produced by Heterologous Expression of a Streptomyces Sp. Cryptic RiPP Biosynthetic Gene Cluster. *J. Nat. Prod.* **2018**, *81*, 264–269. [CrossRef] [PubMed]
96. Kjaerulff, L.; Sikandar, A.; Zaburannyi, N.; Adam, S.; Herrmann, J.; Koehnke, J.; Müller, R. Thioholgamides: Thioamide-Containing Cytotoxic RiPP Natural Products. *ACS Chem. Biol.* **2017**, *12*, 2837–2841. [CrossRef]
97. Bushin, L.B.; Clark, K.A.; Pelczer, I.; Seyedsayamdost, M.R. Charting an Unexplored Streptococcal Biosynthetic Landscape Reveals a Unique Peptide Cyclization Motif. *J. Am. Chem. Soc.* **2018**, *140*, 17674–17684. [CrossRef]
98. Caruso, A.; Bushin, L.B.; Clark, K.A.; Martinie, R.J.; Seyedsayamdost, M.R. Radical Approach to Enzymatic β-Thioether Bond Formation. *J. Am. Chem. Soc.* **2019**, *141*, 990–997. [CrossRef]
99. Clark, K.A.; Bushin, L.B.; Seyedsayamdost, M.R. Aliphatic Ether Bond Formation Expands the Scope of Radical SAM Enzymes in Natural Product Biosynthesis. *J. Am. Chem. Soc.* **2019**, *141*, 10610–10615. [CrossRef]
100. Caruso, A.; Martinie, R.J.; Bushin, L.B.; Seyedsayamdost, M.R. Macrocyclization via an Arginine-Tyrosine Crosslink Broadens the Reaction Scope of Radical S-Adenosylmethionine Enzymes. *J. Am. Chem. Soc.* **2019**, *141*, 16610–16614. [CrossRef]
101. Schramma, K.R.; Seyedsayamdost, M.R. Lysine-Tryptophan-Crosslinked Peptides Produced by Radical SAM Enzymes in Pathogenic Streptococci. *ACS Chem. Biol.* **2017**, *12*, 922–927. [CrossRef]
102. Majchrzykiewicz, J.A.; Lubelski, J.; Moll, G.N.; Kuipers, A.; Bijlsma, J.J.E.; Kuipers, O.P.; Rink, R. Production of a Class II Two-Component Lantibiotic of Streptococcus Pneumoniae Using the Class I Nisin Synthetic Machinery and Leader Sequence. *Antimicrob. Agents Chemother.* **2010**, *54*, 1498–1505. [CrossRef] [PubMed]
103. 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* **2021**, *10*, giaa154. [CrossRef] [PubMed]
104. Kautsar, S.A.; Blin, K.; Shaw, S.; Weber, T.; Medema, M.H. BiG-FAM: The Biosynthetic Gene Cluster Families Database. *Nucleic Acids Res.* **2021**, *49*, D490–D497. [CrossRef] [PubMed]
105. Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S.Y.; Medema, M.H.; Weber, T. AntiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline. *Nucleic Acids Res.* **2019**, *47*, W81–W87. [CrossRef]
106. Mungan, M.D.; Alanjary, M.; Blin, K.; Weber, T.; Medema, M.H.; Ziemert, N. ARTS 2.0: Feature Updates and Expansion of the Antibiotic Resistant Target Seeker for Comparative Genome Mining. *Nucleic Acids Res.* **2020**, *48*, W546–W552. [CrossRef]
107. Mungan, M.D.; Blin, K.; Ziemert, N. ARTS-DB: A Database for Antibiotic Resistant Targets. *Nucleic Acids Res.* **2022**, *50*, D736–D740. [CrossRef]
108. De los Santos, E.L.C. NeuRiPP: Neural Network Identification of RiPP Precursor Peptides. *Sci. Rep.* **2019**, *9*, 13406. [CrossRef]
109. Hannigan, G.D.; Prihoda, D.; Palicka, A.; Soukup, J.; Klempir, O.; Rampula, L.; Durcak, J.; Wurst, M.; Kotowski, J.; Chang, D.; et al. A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction. *Nucleic Acids Res.* **2019**, *47*, e110. [CrossRef]
110. Mohimani, H.; Kersten, R.D.; Liu, W.-T.; Wang, M.; Purvine, S.O.; Wu, S.; Brewer, H.M.; Pasa-Tolic, L.; Bandeira, N.; Moore, B.S.; et al. Automated Genome Mining of Ribosomal Peptide Natural Products. *ACS Chem. Biol.* **2014**, *9*, 1545–1551. [CrossRef]
111. Medema, M.H.; Paalvast, Y.; Nguyen, D.D.; Melnik, A.; Dorrestein, P.C.; Takano, E.; Breitling, R. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. *PLoS Comput. Biol.* **2014**, *10*, e1003822. [CrossRef]
112. Kirkpatrick, C.L.; Broberg, C.A.; McCool, E.N.; Lee, W.J.; Chao, A.; McConnell, E.W.; Pritchard, D.A.; Hebert, M.; Fleeman, R.; Adams, J.; et al. The "PepSAVI-MS" Pipeline for Natural Product Bioactive Peptide Discovery. *Anal. Chem.* **2017**, *89*, 1194–1201. [CrossRef] [PubMed]
113. Mohimani, H.; Liu, W.-T.; Mylne, J.S.; Poth, A.G.; Colgrave, M.L.; Tran, D.; Selsted, M.E.; Dorrestein, P.C.; Pevzner, P.A. Cycloquest: Identification of Cyclopeptides via Database Search of Their Mass Spectra against Genome Databases. *J. Proteome Res.* **2011**, *10*, 4505–4512. [CrossRef] [PubMed]
114. Vila-Farres, X.; Chu, J.; Inoyama, D.; Ternei, M.A.; Lemetre, C.; Cohen, L.J.; Cho, W.; Reddy, B.V.B.; Zebroski, H.A.; Freundlich, J.S.; et al. Antimicrobials Inspired by Nonribosomal Peptide Synthetase Gene Clusters. *J. Am. Chem. Soc.* **2017**, *139*, 1404–1407. [CrossRef] [PubMed]

115. Vila-Farres, X.; Chu, J.; Ternei, M.A.; Lemetre, C.; Park, S.; Perlin, D.S.; Brady, S.F. An Optimized Synthetic-Bioinformatic Natural Product Antibiotic Sterilizes Multidrug-Resistant Acinetobacter Baumannii-Infected Wounds. *mSphere* **2018**, *3*, e00528-17. [CrossRef]
116. Hudson, G.A.; Zhang, Z.; Tietz, J.I.; Mitchell, D.A.; vander Donk, W.A. In Vitro Biosynthesis of the Core Scaffold of the Thiopeptide Thiomuracin. *J. Am. Chem. Soc.* **2015**, *137*, 16012–16015. [CrossRef]
117. DiCaprio, A.J.; Firouzbakht, A.; Hudson, G.A.; Mitchell, D.A. Enzymatic Reconstitution and Biosynthetic Investigation of the Lasso Peptide Fusilassin. *J. Am. Chem. Soc.* **2019**, *141*, 290–297. [CrossRef]
118. Mahanta, N.; Liu, A.; Dong, S.; Nair, S.K.; Mitchell, D.A. Enzymatic Reconstitution of Ribosomal Peptide Backbone Thioamidation. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3030. [CrossRef]
119. Koos, J.D.; Link, A.J. Heterologous and in Vitro Reconstitution of Fuscanodin, a Lasso Peptide from Thermobifida Fusca. *J. Am. Chem. Soc.* **2019**, *141*, 928–935. [CrossRef]
120. Chevrette, M.G.; Carlson, C.M.; Ortega, H.E.; Thomas, C.; Ananiev, G.E.; Barns, K.J.; Book, A.J.; Cagnazzo, J.; Carlos, C.; Flanigan, W.; et al. The Antimicrobial Potential of Streptomyces from Insect Microbiomes. *Nat. Commun.* **2019**, *10*, 516. [CrossRef]