Integration of machine learning and pan-genomics expands the biosynthetic landscape of RiPP natural products

Alexander M. Kloosterman[1], Peter Cimermancic[2], Somayah S. Elsayed[1], Chao Du[1], Michalis Hadjithomas[3$], Mohamed S. Donia[5], Michael A. Fischbach[4], Gilles P. van Wezel[1#], and Marnix H. Medema[6#].

1. Institute of Biology, Leiden University, Netherlands. 2. Verily Life Sciences, South San Francisco, CA. 3. DOE Joint Genome Institute, Walnut Creek, CA. 4. Department of Bioengineering, Stanford University, CA. 5. Department of Molecular Biology, Princeton University, NJ. 6. Bioinformatics group, Wageningen University, Netherlands

# Corresponding authors: g.wezel@biology.leidenuniv.nl, marnix.medema@wur.nl

$ Current address: LifeMine Therapeutics, Cambridge, MA

Abstract

Most clinical drugs are based on microbial natural products, with compound classes including polyketides (PKS), non-ribosomal peptides (NRPS), fluoroquinones and ribosomally synthesized and post-translationally modified peptides (RiPPs). While variants of biosynthetic gene clusters (BGCs) for known classes of natural products are easy to identify in genome sequences, BGCs for new compound classes escape attention. In particular, evidence is accumulating that for RiPPs, subclasses known thus far may only represent the tip of an iceberg. Here, we present decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER), a RiPP genome mining algorithm aimed at the discovery of novel RiPP classes. DecRiPPter combines a Support Vector Machine (SVM) that identifies candidate RiPP precursors with pan-genomic analyses to identify which of these are encoded within operon-like structures that are part of the accessory genome of a genus. Subsequently, it prioritizes such regions based on the presence of new enzymology and based on patterns of gene cluster and precursor peptide conservation across species. We then applied decRiPPter to mine 1,295 *Streptomyces* genomes, which led to the identification of 42 new candidate RiPP families that could not be found by existing programs. One of these was studied further and elucidated as a novel subfamily of lanthipeptides, designated Class V. Two previously unidentified modifying enzymes are proposed to create the hallmark lanthionine bridges. Taken together, our work highlights how novel natural product families can be discovered by methods going beyond sequence similarity searches to integrate multiple pathway discovery criteria.

Code and data availability

The source code of DecRiPPter is freely available online at https://github.com/Alexamk/decRiPPter. Results of the data analysis are available online at http://www.bioinformatics.nl/~medem005/decRiPPter_strict/index.html and http://www.bioinformatics.nl/~medem005/decRiPPter_mild/index.html (for the strict and mild filters, respectively). All training data and code used to generate these, as well as outputs of the data analyses, are available on Zenodo at doi:10.5281/zenodo.3834818.

Introduction

The introduction of antibiotics in the 20th century contributed hugely to extend the human life span. However, the increase in antibiotic resistance and the concomitant steep decline in the number of new compounds discovered via high-throughput screening[1,2], means that we again face huge challenges to treat infections by multi-drug resistant bacteria[3]. The low return of investment of high throughput screening is due to dereplication, in other words, the rediscovery of bioactive compounds that have been identified before[4,5]. A revolution in our understanding was brought about by the development of next-generation sequencing technologies. Actinobacteria are the most prolific producers of bioactive compounds, including some two-thirds of the clinical antibiotics[6,7]. Mining of the genome sequences of these bacteria revealed a huge repository of previously unseen biosynthetic gene clusters (BGCs), highlighting that their potential as producers of bioactive molecules had been grossly underestimated[6,8,9]. However, these BGCs are often not expressed under laboratory conditions, most likely because the environmental cues that activate their expression in their original habitat are missing[10,11]. To circumvent these issues, a common strategy is to select a candidate BGC and force its expression by expression of the pathway-specific activator or via expression of the BGC in a heterologous host[12]. However, these methods are time-consuming, while it is hard to predict the novelty and utility of the compounds they produce.

To improve the success of genome mining-based drug discovery, many bioinformatic tools have been developed for identification and prioritization of BGCs. These tools often rely on conserved genetic markers present in BGCs of certain natural products, such as polyketides (PKS), non-ribosomal peptide synthetases (NRPS) and terpenes[13–15]. While these methods have unearthed vast amounts of uncharacterized BGCs, they further expand on previously characterized classes of natural products. This raises the question of whether entirely novel classes of natural products could still be discovered. A few genome mining methods, such as ClusterFinder[16] and EvoMining[17,18], have tried to tackle this problem. These methods either use criteria true of all BGCs or build around the evolutionary properties of gene families found in BGCs, rather than using specific BGC-class-specific genetic markers. While the lack of clear genetic markers may result in a higher number of false positives, these methods have indeed charted previously uncovered biochemical space and led to the discovery of new natural products.

One class of natural products whose expansion has been fueled by the increased amount of genomic sequences available is that of the ribosomally synthesized and post-translationally modified peptides (RiPPs)[19]. RiPPs are characterized by a unifying biosynthetic theme: a small gene encodes a short precursor peptide, which is extensively modified by a series of enzymes that typically recognize the N-terminal part of the precursor called the leader peptide, and finally cleaved to yield the mature product[20]. Despite this common biosynthetic logic, RiPP modifications are highly diverse. The latest comprehensive review categorizes RiPPs into roughly 20 different classes[19], such as lanthipeptides, lasso peptides and thiopeptides. Each of these classes is characterized by one or more specific modifications, such as the thioether bridge in lanthipeptides or the knot-like structure of lasso peptides. Despite the extensive list of known classes and modifications, new RiPP classes are still being found. Newly identified RiPP classes often carry unusual modifications, such as D-amino acids[21], addition of unnatural amino acids[22,23], β-amino acids[24], or new variants of thioether crosslinks[25,26]. These discoveries strongly indicate that the RiPP genomic landscape remains far from completely charted, and that novel types of RiPPs with new and unique biological activities may yet be uncovered. However, RiPPs pose a unique and major challenge to genome-based pathway identification attempts: unlike in the case of NRPSs and PKSs, there are no universally conserved enzyme families or enzymatic domains that are found across all RiPP pathways. Rather, each class of RiPPs comprises its own unique set of enzyme families to post-translationally modify the precursor peptides belonging to that class. Hence, while biosynthetic gene clusters (BGCs) for known RiPP classes can be identified using conventional genome mining algorithms, a much more elaborate strategy is required to automate the identification of novel RiPP classes.

Several methods have made progress in tackling this challenge. 'Bait-based' approaches such as RODEO[27,28] and RiPPer[29] identify RiPP BGCs by looking for homologues of RiPP tailoring enzymes (RTEs) of interest, and facilitate identifying these RTEs in novel contexts to find many new RiPP BGCs. However, these methods still require a known query RTE from a known RiPP subclass. Another tool recently described, NeuRiPP, is capable of predicting precursors independent of RiPP subclass, but is limited to precursor analysis[30]. Yet another tool, DeepRiPP, can detect novel RiPP BGCs that are chemically far removed from known examples, but is mainly designed to identify new members of known classes[31]. In the end, an algorithm for the discovery of BGCs encoding novel RiPP classes will need to integrate various sources of information to reliably identify genomic regions that are likely to encode RiPP precursors along with previously undiscovered RTEs.

Here, we present decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER), an integrative algorithm for the discovery of novel classes of RiPPs, without requiring prior knowledge of their specific modifications or core enzymatic machinery. DecRiPPter employs a Support Vector Machine (SVM) classifier that predicts RiPP precursors regardless of RiPP subclass, and combines this with pan-genomic analysis to identify which putative precursor genes are located within specialized genomic regions that encode multiple enzymes and are part of the accessory genome of a genus. Sequence similarity networking of the resulting precursors and gene clusters then facilitates further prioritization. Applying this method to the gifted natural product producer genus *Streptomyces*, we identified 42 new RiPP family candidates. Experimental characterization of a widely distributed candidate RiPP BGC led to the discovery of a novel lanthipeptide that was produced by a previously unknown enzymatic machinery.

Results

RiPP BGC discovery by detection of genomic islands with characteristics typical of RiPP BGCs

Given the promise of RiPPs as a source for novel natural products, we set out to construct a platform to facilitate identification of novel RiPP classes. Since no criteria could be used that are specific for individual RiPP classes, we used three criteria that generally apply to RiPP BGCs: 1) they contain one or more ORFs for a precursor peptide; 2) they contain genes encoding modifying machinery in an operon-like gene cluster together with precursor gene(s); 3) they have a sparse distribution within the wider taxonomic group in which they are found. To focus on novel RiPP classes, we added a fourth criterion: 4) they have no direct similarity to BGCs of known classes (Figure 1).

For the first criterion, we trained an SVM classifier to distinguish between RiPP precursors and other peptides. A collection of 175 known RiPP precursors, gathered from RiPP clusters from the MIBiG repository[32] was used as a positive training set. For the negative training set, we generated a set of 20,000 short non-precursor sequences, consisting of 10,000 randomly selected short proteins (<175 amino acids long) from Uniprot without measurable similarity to RiPP precursors (representative of gene encoding proteins but not RiPP precursors), and 10,000 translated intergenic sequences between a stop codon and the next start codon of sizes 30-300 nt taken from 10 genomes across the bacterial tree of life (representative of spurious ORFs that do not encode proteins). From both positive and negative training set sequences, 36 different features were extracted describing the amino acid composition and physicochemical properties of the protein/peptide sequences, as well as localized enrichment of amino acids prone to modification by RTEs. Based on these, a support vector machine was trained (see details in Methods section). To make sure that this classifier could predict precursors independent of RiPP subclass, we trained it on all possible subsets of the positive training set in which one of the RiPP subclasses was entirely left out (a strategy we termed leave-one-class-out cross-validation). Typically, the classifier was still capable of predicting the class that was left out, with an area-under-receiver operating characteristics curve of 0.955**.**

For the second criterion, we made use of the fact that the majority of RiPP BGCs appear to contain the genes encoding the precursor and the core biosynthetic enzymes in the same strand orientation within close intergenic distance (81.6% of MIBiG RiPPs). Therefore, candidate gene clusters are formed from the genes that appear to reside in an operon with predicted precursor genes, based on intergenic distance and the COG scores calculated (see description of third criterion below, the Methods section and Figure S1). These gene clusters are then analyzed for protein domains that could constitute the modifying machinery (Figure 1b). Rather than restricting ourselves to specific protein domains, we constructed a broad dataset of Pfam and TIGRFAM domains that are linked to an E.C. number using InterPro mappings[33]. This dataset was extended with a previously curated set of Pfam domains found to be prevalent in the positive training set of the ClusterFinder algorithm[34], and manually curated, resulting in a set of 4,131 protein domains. We also constructed Pfam[35] and TIGRFAM[36] domain datasets of transporters, regulators and peptidases, as well as a dataset consisting of known RiPP modifying domains to provide more detailed annotation and allow specific filtering of RiPP BGCs based on the presence of each of these types of Pfam domains (Supplemental Document 2).

For the third criterion, we sought to distinguish specialized genomic regions from conserved genomic regions. Indeed, most BGCs are sparingly distributed among genomes, with even closely related strains showing differences in their BGC repertoires[37–39]. We therefore developed an algorithm that separates the 'core' genome from the 'accessory' genome, by comparing all genes in a group of query genomes from the same taxon (typically a genus), and identifying the frequency of occurrence of each gene within that group of genomes (Figures 1c and S2). For the purpose of comparing genes between genomes, we reasoned that it was more straightforward to identify groups of functionally closely related genes that also include recent paralogues, due to the complexities of dealing with orthology relationships across large numbers of genomes (especially for biosynthetic genes that are known to have a discontinuous taxonomic distribution and may undergo frequent duplications[40]). Therefore, decRiPPter first identifies the distribution of sequence identity values of protein-coding genes that can confidently be assigned to be orthologues, and uses this distribution to find groups of genes across genomes with orthologue-like mutual similarity. To identify a set of high-confidence orthologues, decRiPPter looks for genomic loci between which at least three contiguous genes are each other's bidirectional best hits (BBHs, using DIAMOND[41]) between all possible genome pairs of the group of genomes analyzed, and assigned the center genes of these loci orthologue status, termed a true conserved orthologous gene (trueCOG)[42]. Since many orthologues are missed by only considering orthologues based on BBHs[43], and to also include recent paralogues, we then further expanded the list of homologues with orthologue-like similarity by dynamically determining a cutoff between each genome pair based on the similarity of the trueCOGs shared between those genomes. This cutoff is used to find all highly similar gene pairs, which are then clustered with the Markov Clustering Algorithm (MCL[44]) into 'clusters of orthologous genes' (COGs). The number of COG members found for each gene is divided by the number of genomes in the query to get a COG score ranging from 0 to 1, reflecting how widespread the gene is across the set of query genomes. To validate our calculations, we analyzed the COG-scores of the highly conserved single-copy BUSCO gene set from OrthoDB[45,46], as well as the COG-scores of the genes in the gene clusters predicted by antiSMASH. In line with our expectations, homologs of the BUSCO gene set averaged COG-scores of 0.95 (Figure S5), while the COG-scores of the antiSMASH gene clusters were much lower, averaging 0.311 +- 0.249 for all BGCs, and 0.234 +- 0.166 for RiPP BGCs (Figure S6). While the COG-scoring method requires a group of genomes to be analyzed rather than a single genome, we believe that the extra calculation significantly contributes in filtering false positives (see Table 1 and Figure S4). In addition, the COG scores aid in the gene cluster identification based on the assumption that gene clusters are generally sets of genes with similar absence/presence patterns across species (see Methods section).

For the final criterion, the algorithm dereplicates the identified clusters by comparing them to known RiPP BGCs. All putative BGCs are clustered based on domain content and precursor similarity using sequence similarity networking[47], and compared to known RiPP BGCs from MIBiG[48]. In addition, the overlap between predicted RiPP BGCs and gene clusters found by antiSMASH[49] is determined (Figure 1).

decRiPPter identifies 42 candidate novel RiPP classes in *Streptomyces*

While RiPPs are found in many different microorganisms, their presence in streptomycetes reflects perhaps the most diverse array of RiPP classes within a single genus. Streptomycetes produce a broad spectrum of RiPPs, namely lanthipeptides[50], lasso peptides[27], linear azol(in)e-containing peptides (LAPs)[51], thiopeptides[52], thioamide-containing peptides[29] and bottromycins[53]. Their potential as RiPP producers is further highlighted by a recent study showcasing the diversity of lanthipeptide BGCs in *Streptomyces* and other actinobacteria[54]. Given the large variety of different families of natural products produced by this genus, we hypothesized it to be a likely source of novel RiPP classes, and sought to exhaustively mine it.

We started by running the pipeline described above on all publicly available *Streptomyces* genomes (1,295 genomes) from NCBI (Supplemental Document 3). Due to computational limits, the genomes were split into ten randomly selected groups to calculate the frequency of distribution of each gene (COG-scores). In general, the number of genomes that could be grouped together and the resulting cutoffs were found to vary with the amount of minimum trueCOGs required (Figure S3A). To make sure that as many genomes as possible could be compared at once, we set the cutoff for minimum number of trueCOGs at 10. Despite the low cutoff, the distribution of similarity scores between genome pairs still resembled a Gaussian distribution (Figure S3B). The bimodal distribution of the resulting COG-scores showed that the majority of the genes were either conserved in only a small portion of the genomes, or present in almost all genomes (Figure S4).

We then scanned all predicted products of genes as well as predicted ORFs in intergenic regions shorter than 100 amino acids (total $7.19 * 10^7$) with the SVM classifier. While by far most of the queries scored below 0.5, a peak of queries scoring from 0.9 to 1.0 was observed (Figure S7). Seeking to be inclusive at this stage, we set the cutoff at 0.9, resulting in $1.32*10^6$ candidate precursors passing this initial filter, thus filtering out 98.2 % of all candidates. Eliminating candidate precursors whose genes were completely overlapping reduced the number to $8.17*10^5$ precursors (1.1 %). While, most probably, the vast majority of these are not RiPP precursors, it provides a suitably sized set of candidates to then enter the next stages of the decRiPPter workflow.

In our analyses, we found that the majority of RiPP BGCs contain the majority of biosynthetic genes on the same strand orientation as the precursor (MIBiG: 81.6%; antiSMASH RiPP BGCs: 73.1%). We therefore formed gene clusters using only the genes on the same strand as the predicted precursor. To create a training set, we divided all known RiPP BGCs and all antiSMASH RiPP BGCs found in the analyzed genome sequences into sections where each section contained only genes on the same strand. The core section was defined as the section that contained the most biosynthetic genes as detected by antiSMASH or as annotated in the MIBiG database. These sections were used as training sets to finetune distance and COG cutoffs for our gene cluster methods.

In a simple gene cluster method, genes were joined only using the intergenic distances as a cutoff. Using this method, we found that at a distance of 750 nucleotides, all MIBiG core sections were covered, and 91% of all antiSMASH core sections (Figure S8AB). However, using only distance may cause the gene cluster formation to overshoot into regions not associated with the BGC (e.g. Figure S2). We therefore created an alternative method, called the 'island method'. In this method, each gene is first joined with immediately adjacent genes that lie in the same strand orientation and have very small intergenic regions (<=50 nucleotides), to form islands. These islands may subsequently be combined if they have similar average COG-scores (see materials and methods). We found that with this method, we could confidently cover our validation set, while slightly reducing the average size of the gene clusters (3.73 ± 3.75 vs 3.44 ± 3.53; Figure S8CDE). In addition the variation of the COG scores within the gene clusters decreased, suggesting that fewer housekeeping genes would be added to detected biosynthetic gene clusters (Figure S8F).

Overlapping gene clusters were fused, resulting in 7.18 $*10^5$ gene clusters. To organize the results, all clusters were paired if their protein domain content was similar (Jaccard index of protein domains; cutoff: 0.5) and at least one of their predicted precursors showed sequence similarity (NCBI blastp; bitscore cutoff: 30). These cutoffs were used to distinguish between different RiPP subclasses (Figure S9). Clustering these pairs with MCL created 45,727 'families' of gene clusters, containing 312,163 gene clusters, while the remaining 406,105 gene clusters were left ungrouped.

Analysis of overlap between decRiPPter clusters and BGCs predicted by antiSMASH revealed that 5,908 clusters overlapped, constituting 78% of antiSMASH hits, but only 0.8% of decRiPPter clusters (Table 1, row 2). To further narrow down our results, we applied several filters to increase the saturation of RiPP BGCs in our dataset. A mild filter, limiting the average COG score to 0.25 and requiring two biosynthetic genes and a gene encoding a transporter, increased the fraction of overlapping RiPP BGCs to 7.8% (Table 1, row 3). When only clusters associated with genes for a predicted peptidase and a predicted regulator were considered, and the average COG score was limited to 0.1, the fraction increased further to 14.4% (Table 1, row 4). While many antiSMASH RiPP BGCs were filtered out in the process (and, by extension, many unknown RiPP BGCs were likely also filtered out this way), we felt our odds of discovering novel RiPP families were highest when focusing on the dataset with the highest fraction of RiPP BGCs, and therefore applied the strict filter. The remaining 2,471 clusters of genes were clustered as described above. Since our efforts were aimed at finding new gene cluster families, we discarded groups of clusters with fewer than three members, leaving 1,036 gene clusters in 187 families. Families in which more than half of the gene clusters overlapped with antiSMASH non-RiPP BGCs were discarded as well, leaving only known RiPP families and new candidate RiPP families (893 gene clusters, 151 families; Figure 2).

Roughly a third (272) of the remaining gene clusters were members of known families of RiPPs, including lasso peptides, lanthipeptides, thiopeptides, bacteriocins and microcins. In addition, many of the other candidate clusters (55) contained genes common to known RiPP BGCs, such as those encoding YcaO cyclodehydratases and radical SAM-utilizing proteins (Figure 2). These gene clusters were not annotated as RiPP gene clusters by antiSMASH, but the presence of these genes alone or in combination with a suitable precursor can be used as a lead to find novel RiPP gene clusters[24,29].

Each remaining family of gene clusters was manually investigated to filter out likely false positives from the candidates. Common reasons to discard gene clusters were functional annotations of candidate precursors as having a non-precursor function (e.g. homologous to ferredoxin or LysW[55]), annotations of the genes within a gene cluster related to primary metabolism (e.g. genes for cell-wall modifying enzymes), or other abnormalities (e.g. large intergenic gaps or very large gene cluster of more than 40 genes). Several modifying enzymes belonging to the candidate families were homologous to gene products involved in primary metabolism, such as 6-pyruvoyltetrahydropterin synthase or phosphoglycerate mutase. Given the low distribution (COG scores) of the genes encoding these enzymes, it seemed more likely to us that they were adapted from primary metabolism to play a role in secondary metabolism[17]. We therefore only discarded a gene cluster family if multiple clear relations to a known pathway were found. The remaining 42 candidate families were further grouped together into broader classes depending on whether a common enzyme was found (Figure 2).

A large group of families all contained one or more genes for ATP-grasp enzymes. ATP-grasp enzymes are all characterized by a typical ATP-grasp-fold, which binds ATP, which is hydrolyzed to catalyze a number of different reactions. As such, these enzymes have a wide variety of functions in both primary and secondary metabolism, and their genes are present in a many different genomic contexts[56]. Involvement of ATP-grasp enzymes in RiPP biosynthesis has been reported for both microviridin[57] and pheganomycin[23], where they catalyze macrocyclization and peptide ligation, respectively. The ATP-grasp enzymes involved in the biosynthesis of these products did not show direct similarity to any of the ATP-grasp ligases of these candidates, however, suggesting that these belong to yet to be uncovered biosynthetic pathways.

Among the candidate families were three families that contained homologs to *mauE*, and one that additionally contained a homolog of *mauD*. The proteins encoded by these genes, along with other proteins encoded in the *mau* gene cluster, are known to be involved in the maturation of of methylamine dehydrogenase, which is required for methylamine metabolism. MauE in particular has been speculated to play a role in the formation of disulfide bridges in the β-subunit of the protein, while the exact function of MauD remains unclear[58]. As no other orthologs of the *mau* cluster were found within the genomes of *Streptomyces sp.* 2112.3, *Streptomyces viridosporus* T7A or *Streptomyces sp.* CS081A, it is unlikely that these proteins carry out this function. Rather, the presence of these genes in a putative RiPP BGC suggests that they play a role in modification of RTEs or RiPP precursors. Supporting this hypothesis, each of these gene clusters contained a gene predicted to a encode for a precursor containing at least eight cysteine residues (Table S3).

Similarly, homologs of *hypE* and *hypF* were detected in a gene cluster containing another gene encoding an ATP-grasp ligase. Genes encoding these proteins are typically part of the *hyp* operon, which is involved in the maturation of hydrogenase. Specifically, the two proteins cooperate to synthesize a thiocyanate ligand, which is transferred onto an iron center and used as a catalyst[59]. No other homologs of genes in the *hyp* operon were detected, however, suggesting that these protein-coding genes have adopted a novel function.

The remaining 18 families could not be grouped under a single denominator, nor could any single enzyme be found that clearly distinguished these groups as RiPP or non-RiPP BGCs. A wide variety of enzymes was found to be encoded by these gene clusters, including p450 oxidoreductases, flavoproteins, aminotransferases, methyltransferases and phosphatases. In addition (and in line with features dominant in the positive training set), the predicted precursor peptides were often rich in cysteine, serine and threonine residues (Table S3), which contain reactive hydroxyl and sulfide moieties and are present in precursors of various known RiPP subclasses.

All candidate gene clusters presented here carry the features we selected, typical of RiPP BGCs: a low frequency of occurrence among the scanned genomes, a suitable precursor peptide, candidate modifying enzymes, transporters, regulators and peptidases. However, many known RiPP BGCs were removed, suggesting that there may be more uncharacterized RiPP families among the gene clusters we discarded. While the complete dataset could not be covered here, the command-line application of decRiPPter has been set up to allow users to set their own filters. In addition, decRiPPter runs are visualized in an HTML output, in which the results can be further browsed and filtered by Pfam domains and other criteria, allowing users to find candidate families according to their preferences. The results from this analysis of the strict and the mild filter is available at http://www.bioinformatics.nl/~medem005/decRiPPter_strict/index.html and http://www.bioinformatics.nl/~medem005/decRiPPter_mild/index.html, respectively.

Discovery of a novel family of lanthipeptides

To validate the capacity of decRiPPter to find novel RiPP subclasses, we set out to experimentally characterize one of the candidate families (Figure 2; Other; red marker). Gene clusters belonging to this family shared several genes encoding flavoproteins, methyltransferases, oxidoreductases and occasionally a phosphotransferase. Importantly, the predicted precursor peptides encoded by these putative BGCs showed clear conservation of the N-terminal region, while varying more in the C-terminal region (Figure S10). This distinction is typical of RiPP precursors, as the N-terminal leader peptide is used as a recognition site for modifying enzymes, while the C-terminal core peptide can be more variable[20].

One of the gene clusters belonging to this candidate family was identified in *Streptomyces pristinaespiralis* ATCC 25468 (fig 3A; Table 2). *S. pristinaespiralis* is known for the production of pristinamycin, and was selected for experimental work since the strain is genetically tractable[60,61]. The gene cluster was named after its origin (*spr*: *Streptomyces pristinaespiralis* RiPP), and the genes were named after their putative function.

The gene cluster contains four genes encoding putative precursor peptides, although only three of the peptides (SprA1-A3) showed similarity to each other and to the other peptides in the same family (Figure S10). The fourth predicted precursor peptide (encoded by s*prX*) did not align with any of the other peptides and was assumed to be a false positive. The products encoded by *sprA1* and *sprA2* were highly similar to one another compared to the *sprA3* gene product. Occurrence of two distinct genes for precursors within a single RiPP BGC is typical for two-component lanthipeptides[62].

Most of the modifying enzymes present in the gene cluster had not previously been implicated in RiPP biosynthesis. The predicted *sprF2* gene product, however, shows high similarity to cysteine decarboxylases such as EpiD and CypD. These enzymes decarboxylate C-terminal cysteine residues, which is the first step in the formation of C-terminal loop structures called S-[(Z)-2-aminovinyl]-D-cysteine (AviCys) and S-[(Z)-2-aminovinyl]-(3S)-3-methyl-D-cysteine (AviMeCys)[63]. Several RiPP classes have been reported with this modification, including lanthipeptides, cypemycins and thioviridamides, although they are only consistently present in cypemycins and thioviridamides. This type of modification is less common among lanthipeptides, with only nine out of 120 lanthipeptide gene clusters in MIBiG encoding the required decarboxylase. Cysteine-decarboxylating genes are also present in non-RiPP gene clusters (Table S4) and are also associated with other metabolic pathways[64].

A more detailed comparison with the gene clusters in MIBiG showed that two more genes from the thioviridamide gene cluster were homologous to two genes encoding a predicted phosphotransferase (*sprPT*) and a hypothetical protein (*sprH3*), respectively. Taken together with the homologous cysteine decarboxylase, it appeared that our gene cluster was distantly related to the thioviridamide gene cluster[65]. Thioviridamide-like compounds are primarily known for thioamide residues, for which a TfuA-associated YcaO is thought to be responsible[29,66]. However, a YcaO homologue was not encoded by the gene cluster, making it unlikely that this gene cluster should produce thioamide-containing RiPPs.

Two strains were created to help determine the natural product specified by the BGC. For the first strain, the entire gene cluster was replaced by an apramycin resistance cassette (aac3(IV)) by homologous recombination with the pWHM3 vector[67] (*spr*::apra). In case the gene cluster was natively expressed, this strain should allow for easy identification of the natural product by comparative metabolomics. In the second approach, we sought to activate the BGC in case it was not natively expressed. To this end, we targeted the cluster-situated *luxR*-family transcriptional regulatory gene *sprR*. The *sprR* gene was expressed from the strong and constitutive *gapdh* promoter from *S. coelicolor* (p_{gapdh}) on the integrative vector pSET152[68]. The resulting construct (pAK1) was transformed to *S. pristinaespiralis* by protoplast transformation.

To assess the expression of the gene cluster in the transformants, we analyzed changes in the global expression profiles in 2 days and 7 days old samples of NMMP-grown cultures using quantitative proteomics (Figure 3B). Aside from the regulator itself, six out of the sixteen other proteins were detected in the strain containing expression construct pAK1, while only SprPT could be detected in the strain carrying the empty vector pSET152. SprPT was also detected in the proteome of *spr*::apra, however, indicating a false positive. In the wild-type strain, SprT3 and SprR were detected, but only in a single replicate and at a much lower level. Overall, these results suggest that under the chosen growth conditions the gene cluster was expressed at very low amounts in wild-type cells, and was activated when the expression of the likely pathway-specific regulatory gene was enhanced. This makes *spr* a likely cryptic BGC.

To see if a RiPP was produced, the same cultures used for proteomics were separated into mycelial biomass and supernatant. The biomass was extracted with methanol, while HP20 beads were added to the supernatants to absorb secreted natural products. Analysis of the crude methanol extracts and the HP20 eluents with HPLC-MS revealed several peaks eluting between 5.5 and 7 minutes in the methanol extracts (fig 3C), which were not found in extracts from wild-type strain or the strain containing the empty vector. Feature detection with MZMine followed by statistical analysis with MetaboAnalyst revealed seven unique peaks, with m/z between 707.3534 and 918.0807 (Figure S11). The isotope patterns of these peaks showed that the six of the corresponding compounds were triply charged. Careful analysis of derivative peaks with mass increases consistent with Na- or K-addition, led to the conclusion that these peaks corresponded to the $[M+3H]^{3+}$ adduct, suggesting a monoisotopic masses in the range of 2,604.273 and 2,754.242 Da . The highest signal came from the compound with monoisotopic mass of 2,703.245. Four of the other masses seemed to be related to this mass, as they were different in increments of 4, 14, or 16 Da (Table S5). We therefore reasoned that the mass of 2,703.245 Da was the final product, while others were incompletely processed peptides.

To further verify that the identified masses indeed belonged to the RiPP precursors in our gene cluster, we first removed the apramycin resistance cassette from Spr::apra using the pUWLCRE vector[69], creating strain Δ*spr*. The expression construct pAK1 and an empty pSET152 vector were transformed to the strain Δ*spr*. When these strains were grown under the same conditions, the aforementioned peaks were not detected, further suggesting that indeed they belonged to products of this gene cluster (Figure S12).

Most masses were detected in only low amounts. In order to resolve this, we created a similar construct as pAK1, but this time using the low-copy shuttle vector pHJL401 as the vector[70]. The plasmid pAK2 was introduced into *S. pristinaespiralis* and the transformants grown in NMMP for 7 days. Extraction of the mycelial biomass with methanol resulted in a higher abundance of the masses previously detected (Figure S13). Consistent with the MS profiles of pAK1 transformants, also pAK2 transformants produced an abundant peak corresponding to a monoisotopic mass of 2,703.245 Da, as well as a second peak corresponding to a monoisotopic mass of 2,553.260 Da. Most of the other masses could be related to one of these two masses, suggesting these are the final products, related to two distinct precursors (Tables S5 and S6).

We then performed MS-MS analysis of the extracts of the pAK2 transformants to identify the metabolites and their expected modifications, such as Avi(Me)Cys moieties. The fragmentation pattern of the mass of 2,703.245 Da could be assigned to the sprA3 precursor, when several modifications were applied (Figure 3D, Table S7). Similarly, fragments with a mass of 2,553.260 could be matched to the SprA2 precursors considering the same modifications (Figure S14; Table S8).

Among the predicted modifications were N-terminal methylation, which was supported by the presence of the methyltransferase *sprMe* in the gene cluster. Secondly, the C-terminal cysteine was predicted to have undergone oxidative decarboxylation (-46 Da), as expected based on the presence of the gene *sprF2* in the gene cluster. In addition, many of the serines and threonines could only be matched when their masses were altered by -16 or -18 Da. These mass differences are typical of dehydration (-18 Da) of the residues to dehydroalanine and dehydrobutyric acid. Reduction of these dehydrated amino acids (+2 Da) would then give rise to alanine and butyric acid residues, a modification which has been reported for lanthipeptides[71].

To test for the presence of dehydrated serines and threonines, we treated the purified product with dithiothreitol (DTT), which covalently attaches to these residues via 1,4 nucleophilic addition[72]. Treatment with DTT resulted in the addition of up to two adducts, showing the presence of dehydrated residues, although one fewer than expected (Figure S15). The fact that two of the dehydrated residues are adjacent to one another may have resulted in steric hindrance, preventing full conversion.

Surprisingly, no fragments were found of the residues $S^{-18}S^{-18}T^{-18}WC$ in the center of SprA3, or for the N-terminal $T^{-18,\ +28}T^{-18}PVC$ region. Considering the other modifications typical of lanthipeptides, we hypothesized the presence of thioether crosslinks between the dehydrobutyric acids and cysteines. To find further support for this hypothesis, we treated the purified product of SprA3 with iodoacetamide (IAA). Iodoacetamide alkylates free cysteines, while cysteines in thioether bridges remain unmodified[73]. In agreement with our hypothesis, treatment with iodoacetamide did not affect the observed masses, despite the presence of three cysteines in the peptide (Figure S10). In addition, we hydrolyzed the purified peptide with 6M HCl at 110°C for 24h. Under these conditions, the amide bond should be hydrolyzed, while the thioether bond should be unaffected[74]. The resulting mixture of amino acids both contained masses corresponding to a cysteine linked to either a dehydrated serine, or to a twice methylated, dehydrated threonine (Table S10). The C-terminal predicted AviMeCys was not detected, although this may be explained by the presence of the alkene in the moiety, which are likely to react under acidic conditions.

Many of the other masses found were higher when compared to the product of SprA3 by increments of 16 Da, suggesting that the peptide was incompletely processed. The fragmentation patterns of these masses could not be unambiguously resolved (Figure S16). An unmodified serine or threonine could occur at several places within the precursor, and each of the possible outcomes would likely give rise to compounds with identical mass and very similar hydrophobic properties, which would not be separated properly. Overall, these results further reinforce the idea that the compound with monoisotopic mass of 2,703.245 Da belongs to the fully modified product, while the others are derived from it.

The *sprH3/sprPT* gene pair is present in a wide variety of RiPP-like contexts

Taken together, we have shown that the SprA3 precursor contained a number of posttranslational modifications that are typical of lantibiotics. The conversion of serine/threonine to alanine/butyric acid via reduction, the creation of an AviCys moiety and the crosslinks to form thioether bridges are all found in lanthipeptides, and are dependent on dehydration of serine and threonine residues. Four different sets of enzymes, called LanBC, LanM, LanKC and LanKL can catalyze these reactions in the biosynthesis of lanthipeptides and are used to designate the lanthipeptide type.

As stated before, no members of any of these enzyme families were found to be encoded by the gene cluster studied. However, *sprH3* and *sprPT* showed homology to two uncharacterized genes of the thioviridamide BGC. Thioviridamide contains an AviCys moiety, the formation of which requires a dehydrated serine residue. The enzymes responsible for dehydration and subsequent cyclization have not been identified yet[65,75]. Since both gene clusters share a common modification for which the enzyme is unknown, we hypothesized that *sprH3* and *sprPT* should be responsible for dehydration and cyclization, and thus are hallmarks for a new lanthipeptide subtype, which we designate type V.

Lanthipeptide core modifying enzymes catalyze the most prominent reaction in lanthipeptide maturation, and as such, are present in many different genetic contexts[54]. To validate that SprH3 and SprPT are the sought-after modifying enzymes, we studied the distribution of the *SprH3*/*PT* gene pair across *Streptomyces* genomes analyzed by decRiPPter. Using CORASON[76] with the s*prPT* gene as a query yielded 195 homologs in various gene clusters (Figure 4). The *sprPT/sprH3* gene pair was completely conserved across all gene clusters for which an uninterrupted contig of DNA was available. , strongly supporting their functional interaction and joint involvement. Using the *sprH3* gene as a query yielded similar results (data not shown). A total of 391 orthologs of the gene pair were found outside *Streptomyces*, particularly in Actinobacteria (219) and Firmicutes (161; Figure S17). Distantly similar homologs of the gene pair were also identified in Cyanobacteria, Plantomycetes and Proteobacteria.

Among the 195 identified gene clusters in *Streptomyces*, the majority (131) overlapped with a gene cluster detected by decRiPPter, indicating that the gene pair was within short intergenetic distance from predicted precursor gene in the same strand orientation. A large fraction (80) also passed the strictest filtering (see Table I), showing that among these gene clusters were many encoding biosynthetic machinery, peptidases and regulators. In contrast, only nine of the gene clusters overlapped with a BGC identified by antiSMASH. Four of these showed the gene pair in apparent operative linkage with a bacteriocin gene cluster, marked as such by the presence of a DUF692 domain, which is often associated with small prepeptides such as methanobactins. Another four gene clusters detected by decRiPPter were only overlapping due to the gene pair being on the edge of a neighboring gene cluster.

The genetic context of the gene pairs showed a wide variation (Figure 4, right side). While some gene clusters were mostly homologous to the *spr* gene cluster (Figure 4, group g-h), others shared only a few genes (groups a and d), and some only shared the gene pair itself (groups b, c and e). Many other predicted enzyme families were found to be encoded inside these gene clusters, including YcaO-like proteins, glycosyltransferases, sulfotransferases and aminotransferases. The large variation in genetic contexts combined with the clear association with a predicted precursor indicates that this gene pair likely plays a role in many different RiPP-associated genetic contexts, supporting their proposed role as a core gene pair.

Furthermore, we searched for genes encoding enzymes whose functions are dependent on a lanthipeptide dehydration in their substrate, to find if they were associated with the *sprPT/sprH3* gene pair. Both within and outside *Streptomyces*, homologs of *sprF1* and *sprF2* were often found associated with the gene pair (*sprF1*: 251/586; 40.1%; *sprF2*: 281/586; 48.0%; Table S11). Another modification dependent on the presence of dehydrated serine and threonine residues is the conversion of these to alanine and butyric acid, respectively, catalyzed by LtnJ and CrnJ[71]. Outside *Streptomyces*, the genomic surroundings of the *sprPT/sprH3* gene pair occasionally contained homologs of the *ltnJ* gene (40/391; 10.1%), further implying that these genes carry out the canonical dehydration reactions.

A similar modification was observed for SprA2 and SprA3, despite that no homologs of the genes encoding LtnJ or CrnJ were identified within the *spr* gene cluster. However, *sprOR* encodes a putative oxidoreductase, and thus candidates for this modification. Supporting this, orthologs of *sprOR* were found frequently associated with either canonical lanthipeptide BGCs or the *sprPT/sprH3* gene pair (lanthipeptide: 124/462; *sprPT/sprH3*: 137/462; Table S10). One of these lanthipeptide BGCs showed high homology to the lacticin 3147 BGCs from *Lactococcus lactis*. Lacticin 3147 contains several D-alanine residues as a result of conversion of dehydrated serine residues[77]. While all the genes, including the precursors, were well conserved between the two gene clusters, the *ltnJ* gene had been replaced by an *sprOR* homolog, suggesting that their gene products catalyze similar functions (Figure S18).

Conclusion and final perspectives

The continued expansion of available genomic sequence data has allowed for discovery of large reservoirs of natural product BGCs, fueled by sophisticated genome mining methods. These methods must make tradeoffs between novelty and accuracy[12]. Tools primarily aimed at accuracy reliably discover large numbers of known natural product BGCs, but are limited by specific genetical markers. On the other hand, while tools aimed at novelty may discovery new natural products, these tools have to sacrifice on accuracy, resulting in a larger amount of false positives.

Here, we take a new approach to natural product genome mining, aimed specifically at the discovery of novel types of RiPPs. To this end, we built decRiPPter, an integrative approach to RiPP genome mining, based on general features of RiPP BGCs rather than selective presence of specific types of enzymes and domains. To increase the accuracy of our methods, we base detection of the RiPP BGCs on the one thing all RiPP BGCs have in common: a gene encoding a precursor peptide. With this method, we identify 42 candidate novel RiPP families, mined from only 1,295 *Streptomyces* genomes. These families are undetected by antiSMASH, and show no clear markers identifying them as belonging to previously known RiPP BGC classes. While the approach to RiPP genome mining taken here inevitably gives rise to a higher number of false positives, we feel that such a 'low-confidence / high novelty' approach[12] is necessary for the discovery of completely novel RiPP families. Additionally, users are able to set their own filters for the identified gene clusters, allowing them to search candidate RiPP families containing specific enzymes or enzyme types within a much more confined search space compared to manual genome browsing.

The product of one of the candidate classes was characterized as the first member of a new class of lanthipeptides (termed 'type V') that was not detected by any other RiPP genome mining tool. Variants of this gene cluster are widespread across *Streptomyces* species, further expanding one of the most widely studied RiPP families. In addition, two proposed core genes were used to expand the family by finding additional homologs in *Actinobacteria* and *Firmicutes*. Taken together, this work shows that known RiPP families only cover part of the complete genomic landscape, and that many more RiPP families likely remain to be discovered, especially when expanding the search space to the broader bacterial tree of life.

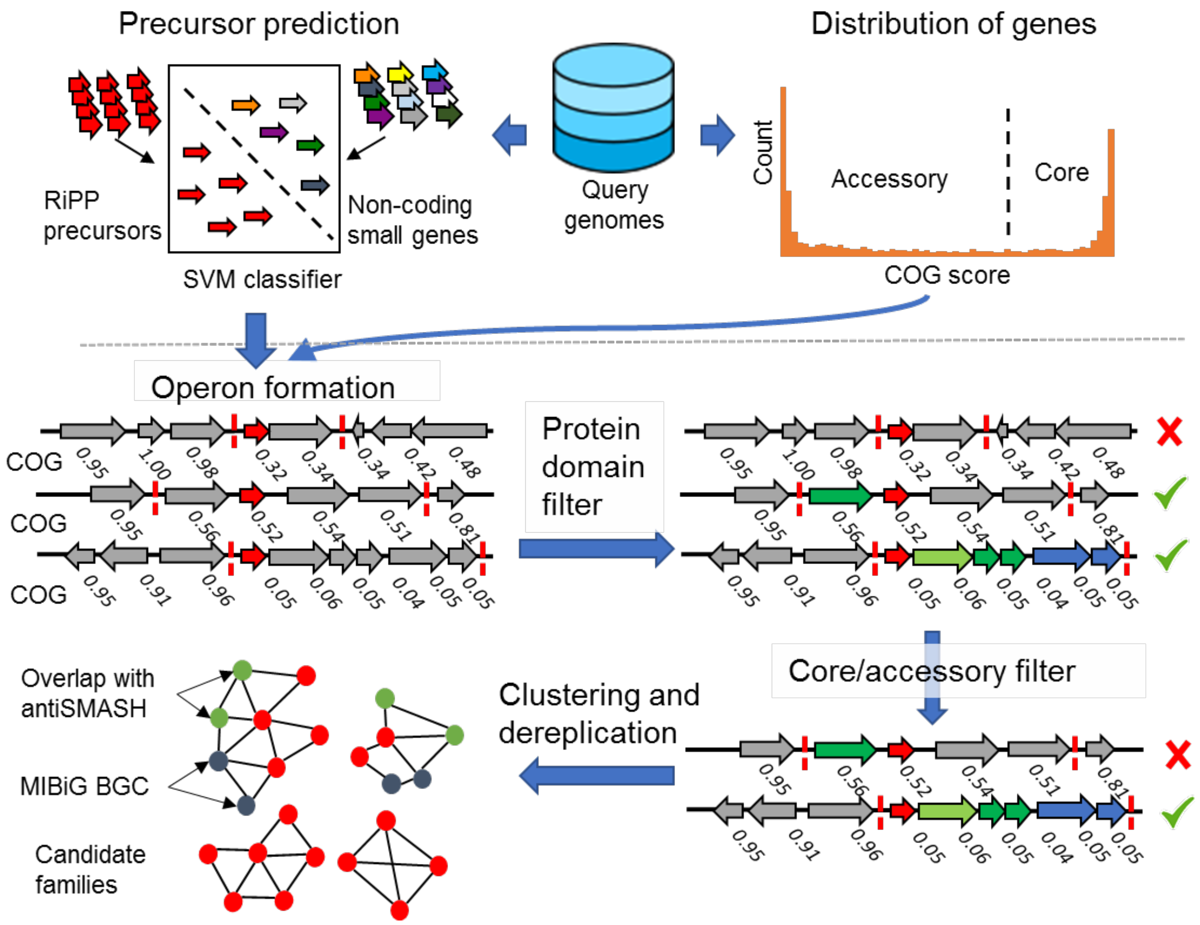

Figure 1. decRiPPter pipeline for the detection of novel RiPP families. From a given group of genomes, all genes smaller than 100 amino acids are analyzed by the SVM classifier, which finds candidate precursors. The gene clusters formed around the precursors are analyzed for specific protein domains. In addition, all COG scores are calculated to act as an additional filter, and to aid in gene cluster detection. The remaining gene clusters are clustered together and with MIBiG gene clusters to dereplicate and organize the results. In addition, overlap with antiSMASH detected BGCs is analyzed (**4**).

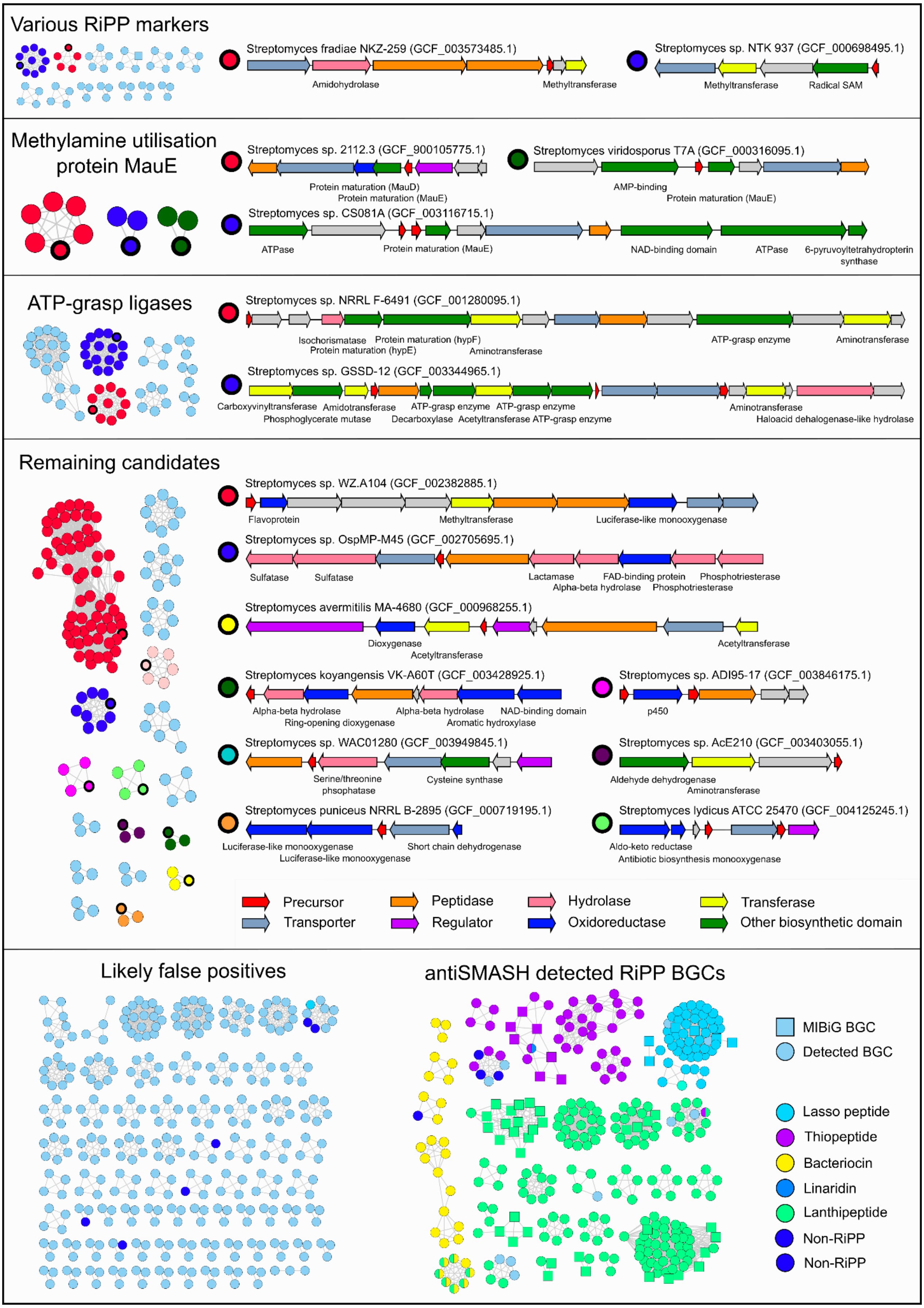
Various RiPP markers
Streptomyces fradiae NKZ-259 (GCF_003573485.1)
Amidohydrolase
Methyltransferase
Streptomyces sp. NTK 937 (GCF_000698495.1)
Methyltransferase
Radical SAM
Methylamine utilisation protein MauE
Streptomyces sp. 2112.3 (GCF_900105775.1)
Protein maturation (MauD)
Protein maturation (MauE)
Streptomyces viridosporus T7A (GCF_000316095.1)
AMP-binding
Protein maturation (MauE)
Streptomyces sp. CS081A (GCF_003116715.1)
ATPase
Protein maturation (MauE)
NAD-binding domain
ATPase
6-pyruvoyltetrahydropterin synthase
ATP-grasp ligases
Streptomyces sp. NRRL F-6491 (GCF_001280095.1)
Isochorismatase
Protein maturation (hypF)
Protein maturation (hypE)
Aminotransferase
ATP-grasp enzyme
Aminotransferase
Streptomyces sp. GSSD-12 (GCF_003344965.1)
Carboxyvinyltransferase
Amidotransferase
Phosphoglycerate mutase
ATP-grasp enzyme
Decarboxylase
ATP-grasp enzyme
Acetyltransferase
ATP-grasp enzyme
Aminotransferase
Haloacid dehalogenase-like hydrolase
Remaining candidates
Streptomyces sp. WZ.A104 (GCF_002382885.1)
Flavoprotein
Methyltransferase
Luciferase-like monooxygenase
Streptomyces sp. OspMP-M45 (GCF_002705695.1)
Sulfatase
Sulfatase
Lactamase
Alpha-beta hydrolase
FAD-binding protein
Phosphotriesterase
Phosphotriesterase
Streptomyces avermitilis MA-4680 (GCF_000968255.1)
Dioxygenase
Acetyltransferase
Acetyltransferase
Streptomyces koyangensis VK-A60T (GCF_003428925.1)
Alpha-beta hydrolase
Ring-opening dioxygenase
Alpha-beta hydrolase
Aromatic hydroxylase
NAD-binding domain
Streptomyces sp. ADI95-17 (GCF_003846175.1)
p450
Streptomyces sp. WAC01280 (GCF_003949845.1)
Serine/threonine phosphatase
Cysteine synthase
Streptomyces sp. AcE210 (GCF_003403055.1)
Aldehyde dehydrogenase
Aminotransferase
Streptomyces puniceus NRRL B-2895 (GCF_000719195.1)
Luciferase-like monooxygenase
Luciferase-like monooxygenase
Short chain dehydrogenase
Streptomyces lydicus ATCC 25470 (GCF_004125245.1)
Aldo-keto reductase
Antibiotic biosynthesis monooxygenase
Precursor
Transporter
Peptidase
Regulator
Hydrolase
Oxidoreductase
Transferase
Other biosynthetic domain
Likely false positives
antiSMASH detected RiPP BGCs
MIBiG BGC
Detected BGC
Lasso peptide
Thiopeptide
Bacteriocin
Linaridin
Lanthipeptide
Non-RiPP
Non-RiPP

Figure 2. decRiPPter finds 42 candidate RiPP families with a large variety of encoded modifying enzymes and precursors . Gene clusters found in 1,295 *Streptomyces* genomes were passed through a strict filter and grouped together (see main text). Arrow colors indicate enzyme family of the product, and the description of gene products is given below the arrows. Roughly a third of the remaining candidates overlapped with or were similar to RiPP BGCs predicted by antiSMASH. Another third of the remaining candidates were discarded as likely false positives (see main text). Of the remaining 42 candidate RiPP families, 15 example gene clusters are displayed.

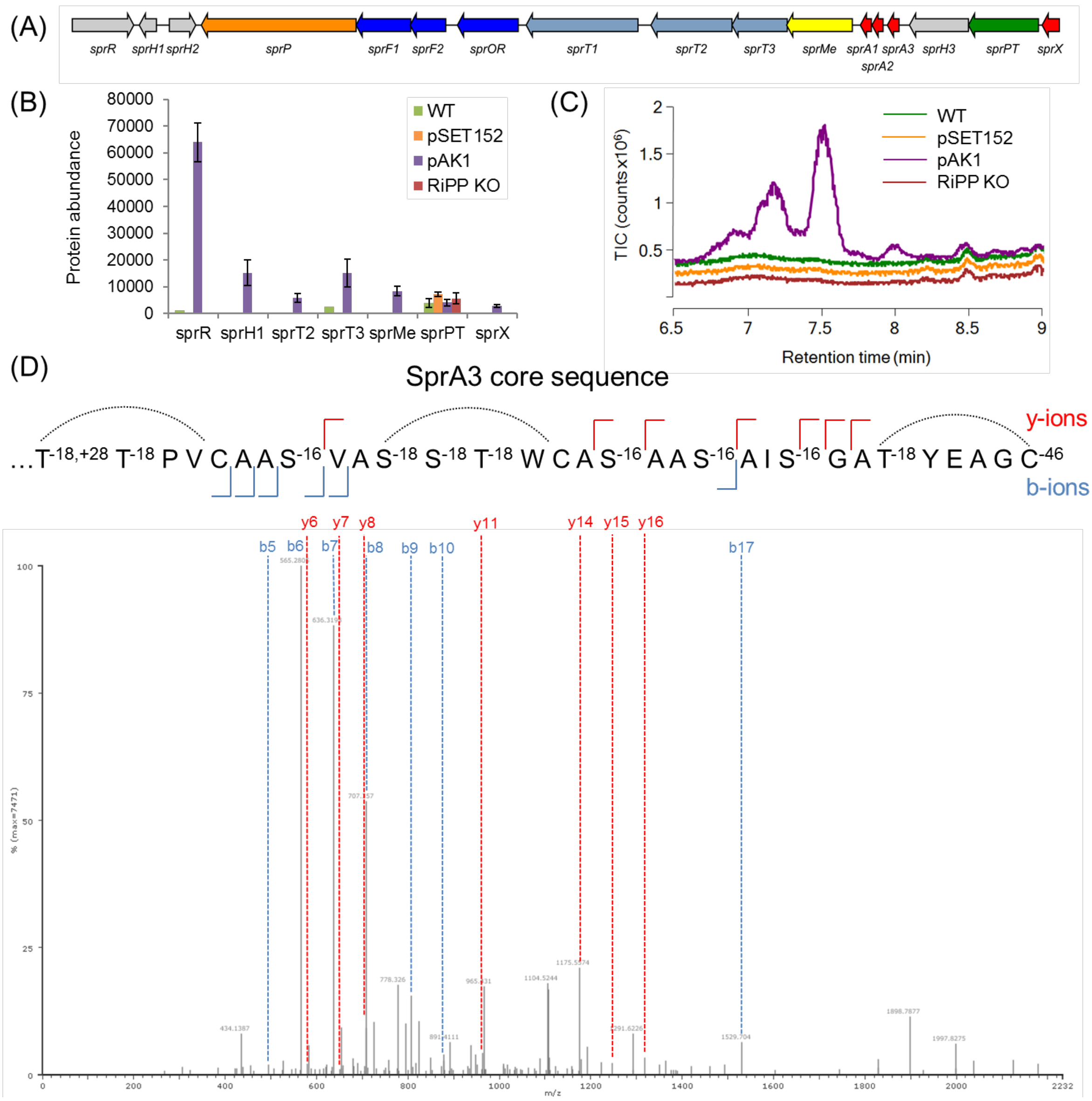

Figure 3. The *Streptomyces pristinaespiralis RiPP* (spr) gene cluster produces a highly modified RiPP. A) The *spr* gene cluster encodes three putative precursors, three transporters, a peptidase and an assortment of modifying enzymes (see Table 1). B) Protein abundance of the products of the *spr* gene cluster in *S. pristinaespiralis* ATCC 25468 and derived strains. Strains were grown in NMMP and samples were taken after 2 and 7 days. Enhanced expression of the regulator (from construct pAK1) resulted in the partial activation of the gene cluster. Genes that could not be detected are not illustrated. C) Chromatogram of crude extracts from strains grown under the same conditions as under A), samples after 7 days. Several peaks were detected in the extract from the strain with expression construct pAK1 between 7 and 8 minutes. C) b and y ions detected from one of the predominant peaks found in the crude extract (corresponding to monoisotopic mass of 2703.235 Da). The fragmentation pattern could be matched to the sprA3 precursor.

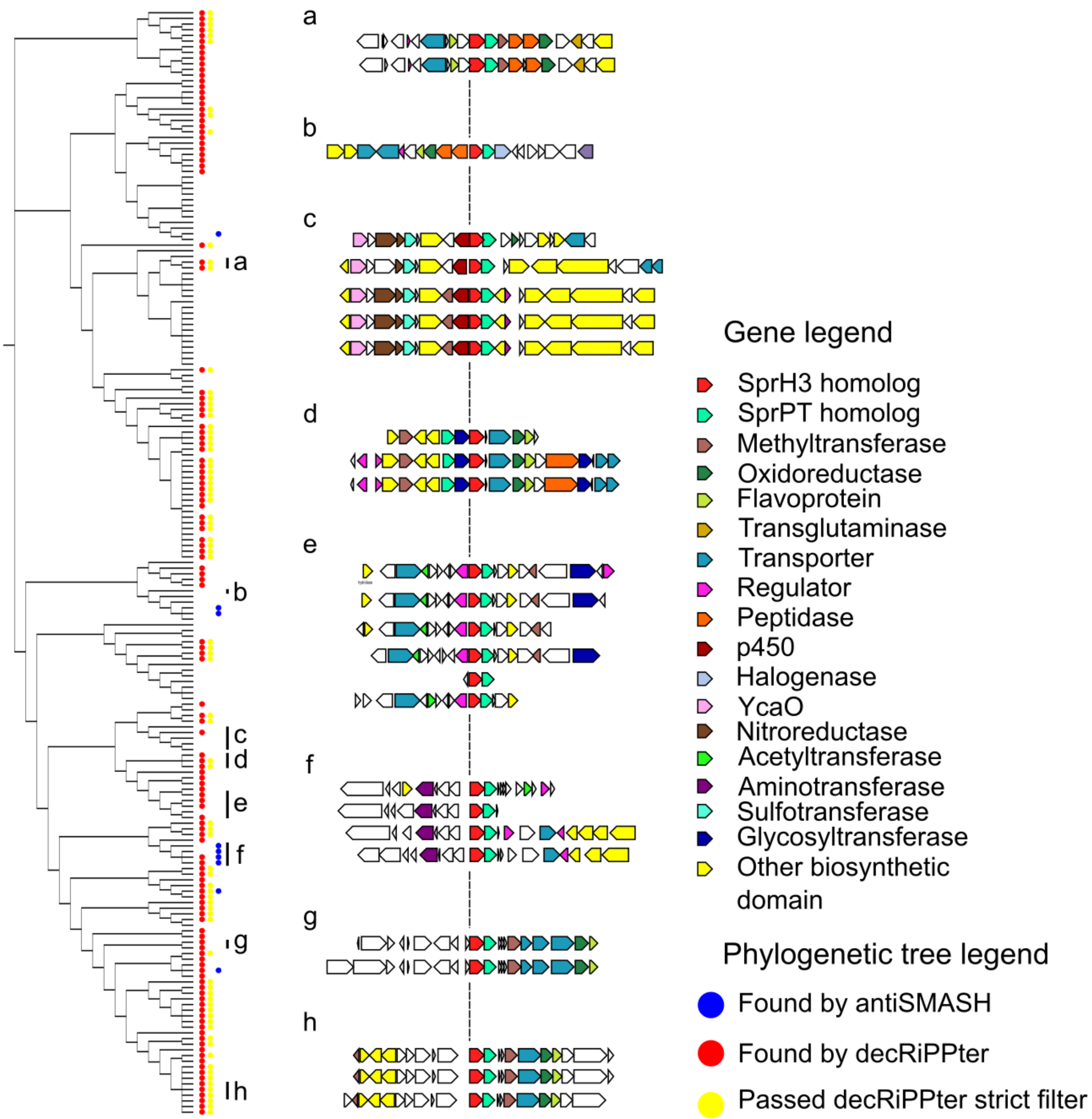

Figure 4. Orthologs of *sprPT* and *sprH3* cooccur in a wide variety of genetic contexts. (Left side) Phylogenetic tree of gene clusters containing homologs of *sprPT* and *sprH3*, visualized by CORASON^{76}. A red dot indicates that the genes were present in a gene cluster found by decRiPPter, a yellow dot that it passed the strict filter (see Table 1 for details). A blue dot indicates overlap with a BGC identified by antiSMASH. (Right side) Several gene clusters with varying genetic contexts are displayed. Group (g) represents the query gene cluster. The genetic context varies, while the gene pair itself is conserved. Color indicates predicted enzymatic activity of the gene products as described in the legend.

Table 1. Increasing the strictness of the filter used on the found gene clusters results in a higher saturation of RiPP BGCs.

Filter	**Filter details**	**Number of detected gene clusters**	**Number of detected gene clusters overlapping with antiSMASH RiPP BGCs**	**Percentage of detected gene clusters overlapping with RiPP BGCs**
None	-	718268	5908	0.8%
Mild	Gene cluster COG score: <= 0.25 In the gene cluster: • >= 3 genes • >= 2 biosynthetic genes In or around the gene cluster: • >= 1 transporter	21419	1678	7.8%
Strict	Gene cluster COG score: <= 0.10 In the gene cluster: • >= 3 genes • >= 2 biosynthetic genes In or around the gene cluster: • >= 1 transporter • >= 1 regulator • >= 1 peptidase	2471	357	14.4%

587 **Table 2. Annotation of the *Streptomyces pristinaespiralis* RiPP (*spr*) gene cluster.**

Gene name	Accession	NCBI Annotation	Protein domains found	Proposed function
sprR	ALC22061.1	LuxR family transcriptional regulator		Cluster-specific regulator
sprH1	ALC22062.1	hypothetical protein		Unknown
sprH2	ALC22063.1	hypothetical protein		Unknown
sprP	ALC22064.1	Peptidase M16 domain-containing protein	PF00675 Insulinase PF05193 Peptidase M16 inactive domain	RiPP maturation protease
sprPT1	ALC22065.1	Flavoprotein	PF01636 Phosphotransferase	Cysteine decarboxylation
sprF	ALC22066.1	Flavoprotein	PF02441 Flavoprotein	Cysteine decarboxylation
sprOR	ALC22067.1	5,10-methylene tetrahydromethanopterin reductase	PF00291 Luciferase-like monooxygenase	Reduction of dehydroalanine and dehydrobutyric acid
sprT1	ALC22068.1	ABC transporter ATP-binding protein	PF00005 ABC transporter PF00664 ABC transporter transmembrane region	Transport
sprT2	ALC22069.1	ABC transporter	PF12698 ABC-2 family transporter protein	Transport
sprT3	ALC22070.1	ABC transporter ATP-binding protein	PF00005 ABC transporter PF13732 Domain of unknown function (DUF4162)	Transport
sprMe	ALC22071.1	carminomycin 4-O-methyltransferase	PF00891 O-methyltransferase domain	N-terminal methylation
sprA1	ALC22072.1	hypothetical protein		RiPP precursor
sprA2	ALC22073.1	hypothetical protein		RiPP precursor
sprA3	ALC22074.1	hypothetical protein		RiPP precursor
sprH3	ALC22075.1	hypothetical protein	PF17914 HopA1 effector protein family	Dehydration/cyclization
sprPT2	ALC22076.1	hypothetical protein	PF01636 Phosphotransferase	Dehydration/cyclization
sprX	ALC22077.1	hypothetical protein		Unknown

Table 3. Co-occurrence of genes found in the *spr* gene cluster with homologs of *sprPT* in the analyzed 1,295 *Streptomyces* strains.

Gene name	Co-occurrence with *sprPT* (percentage)
sprH3	99.49
sprMe	20
sprT1	35.38
sprT2	12.31
sprT3	12.82
sprOR	64.62
sprF1	39.5
sprF2	68.72
sprP	38.5
sprH1	9.0
sprH2	2.0
sprR	28.5
sprA1	1.03
sprA2	1.03
sprA3	16.92

References

1. Cooper MA, Shlaes D. Fix the antibiotics pipeline. *Nature*. 2011;472(7341):32. doi:10.1038/472032a

2. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: Confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov*. 2007;6(1):29-40. doi:10.1038/nrd2201

3. Davies J. Origins and evolution of antibiotic resistance. *Microbiologia*. 1996;12(1):9-16. doi:10.1128/mmbr.00016-10

4. Lewis K. Platforms for antibiotic discovery. *Nat Rev Drug Discov*. 2013;12(5):371-387. doi:10.1038/nrd3975

5. Kolter R, Van Wezel GP. Goodbye to brute force in antibiotic discovery? *Nat Microbiol*. 2016;1(2):1-2. doi:10.1038/nmicrobiol.2015.20

6. Barka EA, Vatsa P, Sanchez L, et al. Taxonomy, Physiology, and Natural Products of Actinobacteria. *Microbiol Mol Biol Rev*. 2016;80(1):1-43. doi:10.1128/mmbr.00019-15

7. Bérdy J. Thoughts and facts about antibiotics: Where we are now and where we are heading. *J Antibiot (Tokyo)*. 2012;65(8):385-395. doi:10.1038/ja.2012.27

8. Bentley SD, Chater KF, Cerdeño-Tárraga AM, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*. 2002;417(6885):141-147. doi:10.1038/417141a

9. van der Aart LT, Nouioui I, Kloosterman A, et al. Polyphasic classification of the gifted natural product producer streptomyces roseifaciens sp. Nov. *Int J Syst Evol Microbiol*. 2019;69(4):899-908. doi:10.1099/ijsem.0.003215

10. Rutledge PJ, Challis GL. Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat Rev Microbiol*. 2015;13(8):509-523. doi:10.1038/nrmicro3496

11. van der Meij A, Worsley SF, Hutchings MI, van Wezel GP. Chemical ecology of antibiotic production by actinomycetes. *FEMS Microbiol Rev*. 2017. doi:10.1093/femsre/fux005

12. Medema MH, Fischbach M a. Computational approaches to natural product discovery. *Nat Chem Biol*. 2015;11:639-648.

13. Blin K, Shaw S, Steinke K, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*. 2019;47(W1):W81-W87. doi:10.1093/nar/gkz310

14. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res*. 2017;45(W1):W49-W54. doi:10.1093/nar/gkx320

15. van Heel AJ, de Jong A, Song C, Viel JH, Kok J, Kuipers OP. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res*. 2018;46(W1):W278-W281. doi:10.1093/nar/gky383

16. Cimermancic P, Medema MH, Claesen J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*. 2014;158(2):412-421. doi:10.1016/j.cell.2014.06.034

17. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biol Evol*. 2016;8(6):1906-1916. doi:10.1093/gbe/evw125

18. Séelem-Mojica N, Aguilar C, Gutiéerrez-García K, Martínez-Guerrero CE, Barona-Gómez F. Evomining reveals the origin and fate of natural product biosynthetic enzymes. *Microb Genomics*. 2019;5(12). doi:10.1099/mgen.0.000260

19. Arnison PG, Bibb MJ, Bierbaum G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep*. 2013;30(1):108-160. doi:10.1039/C2NP20085F

20. Oman TJ, Van Der Donk WA. Follow the leader: The use of leader peptides to guide natural product biosynthesis. *Nat Chem Biol*. 2010. doi:10.1038/nchembio.286

21. Freeman MF, Gurgui C, Helf MJ, et al. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science (80-)*. 2012;338(6105):387-390. doi:10.1126/science.1226121

22. Ogasawara Y, Kawata J, Noike M, Satoh Y, Furihata K, Dairi T. Exploring Peptide Ligase Orthologs in Actinobacteria�Discovery of Pseudopeptide Natural Products, Ketomemicins. 2016. doi:10.1021/acschembio.6b00046

23. Noike M, Matsui T, Ooya K, et al. A peptide ligase and the ribosome cooperate to synthesize the peptide pheganomycin. *Nat Chem Biol*. 2015;11(1):71-76. doi:10.1038/nchembio.1697

24. Morinaka BI, Lakis E, Verest M, et al. Natural noncanonical protein splicing yields products with diverse β-amino acid residues. *Science*. 2018;359(6377):779-782. doi:10.1126/science.aao0157

25. Caruso A, Bushin LB, Clark KA, Martinie RJ, Seyedsayamdost MR. A Radical Approach to Enzymatic #-Thioether Bond Formation A Radical Approach to Enzymatic β-Thioether Bond Formation. 2018. doi:10.1021/jacs.8b11060

26. Hudson GA, Burkhart BJ, DiCaprio AJ, et al. 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*. May 2019:jacs.9b01519. doi:10.1021/jacs.9b01519

27. Tietz JI, Schwalen CJ, Patel PS, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat Chem Biol*. 2017;13(5):470-478. doi:10.1038/nchembio.2319

28. Schwalen CJ, Hudson GA, Kille B, Mitchell DA. Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics. *J Am Chem Soc*. 2018;140(30):9494-9501. doi:10.1021/jacs.8b03896

29. Santos-Aberturas J, Chandra G, Frattaruolo L, et al. Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool. *bioRxiv*. December 2018:494286. doi:10.1101/494286

30. Santos ELC de los. NeuRiPP: Neural network identification of RiPP precursor peptides. *bioRxiv*. May 2019:616060. doi:10.1101/616060

31. Merwin NJ, Mousa WK, Dejong CA, et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc Natl Acad Sci U S A*. 2020;117(1):371-380. doi:10.1073/pnas.1901493116

32. Kautsar SA, Blin K, Shaw S, et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res*. 2020;48(D1):D454-D458. doi:10.1093/nar/gkz882

33. Mitchell A, Chang HY, Daugherty L, et al. The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res*. 2015;43(D1):D213-D221. doi:10.1093/nar/gku1243

34. Cimermancic P, Medema MH, Claesen J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*. 2014;158(2):412-421. doi:10.1016/j.cell.2014.06.034

35. El-Gebali S, Mistry J, Bateman A, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47(D1):D427-D432. doi:10.1093/nar/gky995

36. Haft DH. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. 2001;29(1):41-43. doi:10.1093/nar/29.1.41

37. Choudoir M, Pepe-Ranney C, Buckley D. Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in Streptomyces. *Antibiotics*. 2018;7(1):12. doi:10.3390/antibiotics7010012

38. Xu L, Ye K-X, Dai W-H, Sun C, Xu L-H, Han B-N. Comparative Genomic Insights into Secondary Metabolism Biosynthetic Gene Cluster Distributions of Marine Streptomyces. *Mar Drugs*. 2019;17(9):498. doi:10.3390/md17090498

39. Amos GCA, Awakawa T, Tuttle RN, et al. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. doi:10.1073/pnas.1714381115

40. Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. *PLoS Comput Biol*. 2014;10(12). doi:10.1371/journal.pcbi.1004016

41. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59-60. doi:10.1038/nmeth.3176

42. Wolf YI, Koonin E V. A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes. *Genome Biol Evol*. 2012;4(12):1286-1294. doi:10.1093/gbe/evs100

43. Dalquen DA, Dessimoz C. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biol Evol*. 2013;5(10):1800-1806. doi:10.1093/gbe/evt132

44. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J Matrix Anal Appl*. 2008;30(1):121-141. doi:10.1137/040608635

45. Kriventseva E V., Kuznetsov D, Tegenfeldt F, et al. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 2019;47(D1):D807-D811. doi:10.1093/nar/gky1053

46. Waterhouse RM, Seppey M, Simao FA, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 2018;35(3):543-548. doi:10.1093/molbev/msx319

47. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for

visualization of relationships across diverse protein superfamilies. *PLoS One*. 2009;4(2). doi:10.1371/journal.pone.0004345

48. Medema MH, Kottmann R, Yilmaz P, et al. Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol*. 2015;11(9):625-631. doi:10.1038/nchembio.1890

49. Blin K, Wolf T, Chevrette MG, et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*. 2017;(1):1-6. doi:10.1093/nar/gkx319

50. Kersten RD, Yang YL, Xu Y, et al. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol*. 2011;7(11):794-802. doi:10.1038/nchembio.684

51. Onaka H, Nakaho M, Hayashi K, Igarashi Y, Furumai T. Cloning and characterization of the goadsporin biosynthetic gene cluster from Streptomyces sp. TP-A0584. *Microbiology*. 2005;151(12):3923-3933. doi:10.1099/mic.0.28420-0

52. Kelly WL, Pan L, Li C. Thiostrepton Biosynthesis: Prototype for a New Family of Bacteriocins. doi:10.1021/ja807890a

53. Gomez-Escribano JP, Song L, Bibb MJ, Challis GL. Posttranslational β-methylation and macrolactamidination in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics. *Chem Sci*. 2012;3(12):3522-3525. doi:10.1039/c2sc21183a

54. Zhang Q, Doroghazi JR, Zhao X, Walker MC, Van Der Donk WA. Expanded Natural Product Diversity Revealed by Analysis of Lanthipeptide-Like Gene Clusters in Actinobacteria. 2015. doi:10.1128/AEM.00635-15

55. Horie A, Tomita T, Saiki A, et al. Discovery of proteinaceous N-modification in lysine biosynthesis of Thermus thermophilus. *Nat Chem Biol*. 2009;5(9):673-679. doi:10.1038/nchembio.198

56. Fawaz M V, Topper M, Firestine SM. The ATP-Grasp Enzymes. *Bioorg Chem*. 2011;39(5-6):185-191. doi:10.1016/j.bioorg.2011.08.004

57. Ziemert N, Ishida K, Weiz A, Hertweck C, Dittmann E. Exploiting the natural diversity of microviridin gene clusters for discovery of novel tricyclic depsipeptides. *Appl Environ Microbiol*. 2010;76(11):3568-3574. doi:10.1128/AEM.02858-09

58. Van Der Palen CJNM, Reijnders WNM, De Vries S, Duine JA, Rob ;, Van Spanning JM. *MauE and MauD Proteins Are Essential in Methylamine Metabolism of Paracoccus Denitrificans*.

59. Jacobi A, Rossmann R, Böck A. The hyp operon gene products are required for the maturation of catalytically active hydrogenase isoenzymes in Escherichia coli. *Arch Microbiol*. 1992;158(6):444-451. doi:10.1007/BF00276307

60. Mast Y, Weber T, Gölz M, et al. Characterization of the “pristinamycin supercluster” of Streptomyces pristinaespiralism bt_213 192..206. doi:10.1111/j.1751-7915.2010.00213.x

61. Folcher M, Morris RP, Dale G, Salah-Bey-Hocini K, Viollier PH, Thompsonţ CJ. A Transcriptional Regulator of a Pristinamycin Resistance Gene in Streptomyces coelicolor*. 2000. doi:10.1074/jbc.M007690200

62. Garneau S, Martin NI, Vederas JC. Two-peptide bacteriocins produced by lactic acid bacteria.

Biochimie. 2002;84(5-6):577-592. doi:10.1016/S0300-9084(02)01414-1

63. Sit CS, Yoganathan S, Vederas JC. Biosynthesis of Aminovinyl-Cysteine-Containing Peptides and Its Application in the Production of Potential Drug Candidates. *Acc Chem Res*. 2011;44(4):261-268. doi:10.1021/ar1001395

64. Clausen M, Lamb CJ, Megnet R, Doerner PW. PAD1 encodes phenylacrylic acid decarboxylase which confers resistance to cinnamic acid in Saccharomyces cerevisiae. *Gene*. 1994;142(1):107-112. doi:10.1016/0378-1119(94)90363-8

65. Tang J, Lu J, Luo Q, Wang H. Discovery and biosynthesis of thioviridamide-like compounds. *Chinese Chem Lett*. 2018;29:1022-1028. doi:10.1016/j.cclet.2018.05.004

66. Burkhart BJ, Schwalen CJ, Mann G, Naismith JH, Mitchell DA. YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function. *Chem Rev*. 2017;117(8):5389-5456. doi:10.1021/acs.chemrev.6b00623

67. Vara J, Lewandowska-Skarbek M, Wang YG, Donadio S, Hutchinson CR. Cloning of genes governing the deoxysugar portion of the erythromycin biosynthesis pathway in Saccharopolyspora erythraea (Streptomyces erythreus). *J Bacteriol*. 1989;171:5872-5881.

68. Bierman M, Logan R, O'Brien K, Seno ET, Rao RN, Schoner BE. Plasmid cloning vectors for the conjugal transfer of DNA from Escherichia coli to Streptomyces spp. *Gene*. 1992;116:43-49. doi:10.1016/0378-1119(92)90627-2

69. Fedoryshyn M, Welle E, Bechthold A, Luzhetskyy A. Functional expression of the Cre recombinase in actinomycetes. *Appl Microbiol Biotechnol*. 2008;78(6):1065-1070. doi:10.1007/s00253-008-1382-9

70. Larson JL, Hershberger CL. The minimal replicon of a streptomycete plasmid produces an ultrahigh level of plasmid DNA. *Plasmid*. 1986;15:199-209. doi:10.1016/0147-619X(86)90038-7

71. Yang X, Van Der Donk WA. Post-translational Introduction of D-Alanine into Ribosomally Synthesized Peptides by the Dehydroalanine Reductase NpnJ. 2015. doi:10.1021/jacs.5b05207

72. Cox CL, Tietz JI, Sokolowski K, Melby JO, Doroghazi JR, Mitchell DA. Nucleophilic 1,4-additions for natural product discovery. *ACS Chem Biol*. 2014;9(9):2014-2022. doi:10.1021/cb500324n

73. Zhao X, Van Der Donk WA. Structural Characterization and Bioactivity Analysis of the Two-Component Lantibiotic Flv System from a Ruminant Bacterium. *Cell Chem Biol*. 2016;23(2):246-256. doi:10.1016/j.chembiol.2015.11.014

74. Ross AC, Liu H, Pattabiraman VR, Vederas JC. Synthesis of the lantibiotic lactocin S using peptide cyclizations on solid phase. *J Am Chem Soc*. 2010;132(2):462-463. doi:10.1021/ja9095945

75. Frattaruolo L, Lacret R, Cappello AR, Truman AW. A Genomics-Based Approach Identifies a Thioviridamide-Like Compound with Selective Anticancer Activity. 2017. doi:10.1021/acschembio.7b00677

76. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, et al. A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *bioRxiv*. October 2018:445270. doi:10.1101/445270

77. Cotter PD, O'Connor PM, Draper LA, et al. Posttranslational conversion of L-serines to D-alanines is vital for optimal production and activity of the lantibiotic lacticin 3147. *Proc Natl Acad Sci U S A*. 2005;102(51):18584-18589. doi:10.1073/pnas.0509371102