Exploring the chemical space of post-translationally modified peptides in Streptomyces with machine learning

Kloosterman, A.M.

Exploring the chemical space of post-translationally modified peptides in Streptomyces *with machine learning*

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 12 mei 2021
klokke 10.00 uur

door

Alexander Merlijn Kloosterman

geboren te Leiden, Nederland

in 1990

Promotor: Prof. Dr. G.P. van Wezel

Copromotor: Dr. M.H. Medema

Promotiecommissie: Prof. Dr. A.H. Meijer (Universiteit Leiden)

Prof. Dr. A. Briegel (Universiteit Leiden)

Prof. Dr. O.P. Kuipers (Rijksuniversiteit Groningen)

Dr. T. Abeel (Technische Universiteit Delft)

Funding statement: This work was supported by Grant 731.014.206 (Syngenopep, TKI Chemie) from the Dutch Research Council (NWO).

To my friends and family

Contents

1

Omics-based strategies to discover novel classes of RiPP natural products

Alexander M. Kloosterman

Marnix H. Medema

Gilles P. van Wezel

The work described in this chapter is part of the following publication:

Kloosterman, et al., *Omics-based strategies to discover novel classes of RiPP natural products.* Curr. Opin. Biotechnol., 2020. **6**, p: 60-67.

1 The discovery of antibiotics and the rise of resistance

Every living organism has metabolic pathways to catabolize and produce organic molecules, collectively called natural products. A crude distinction made in 1891 by A. Kossel separates these molecules into two classes: those that are directly essential for growth, development and reproduction of the organism belong to primary metabolism, while the remainder belong to secondary metabolism [1]. Secondary metabolites are highly versatile and cover a large chemical space. Even closely related strains may produce a different repertoire of secondary metabolites [2-5]. While mostly not required for the growth of an organism in a pure culture, secondary metabolites are thought to confer advantages in the natural environment of the producing strains. They have been found to act as means of communication with other species, signal cellular differentiation and scavenge metal ions. Most importantly, many secondary metabolites act as competitive weapons against other species and can be used as anti-bacterial agents in the clinic [2].

Interest in secondary metabolites for medical applications arguably started in 1928, with the discovery of penicillin by Alexander Flemming [6]. Penicillin, first isolated from the fungus *Penicillium notatum*, has strong antimicrobial properties, making it a highly promising candidate for use as a medicine against bacterial infections. It would take another ten years before penicillin was isolated in a pure form and further investigated as a potential drug. When it appeared that the drug had low toxicity and proved suitable for human consumption, its use would become standard practice in the clinic [7].

The potential of secondary metabolites to be of practical use led to a surge of investigations towards the discovery of more of these molecules. Selman A. Waksman would follow up on the discovery of penicillin by screening soil samples for strains harboring biological activities. From these efforts, several more antibiotics were isolated, including actinomycin D, neomycin and streptomycin, all produced by filamentous bacteria belonging to the genus of Actinobacteria [8-10]. Similar screening methods were later executed on a much larger scale by the pharmaceutical industry. Between the 1940's and the 1970's, more than 20 different classes of antibiotics were discovered by high-

throughput screening (HTS) efforts, many of which are still being used in the clinic today [11]. Thousands of compounds with antibiotic activity are currently known, the majority of which were isolated from Actinobacteria [12]. By changing the parameters of the screening, additional secondary metabolites with different useful functions were identified, which could be used as insecticides, hypertension relievers or immunosuppressants [13]. Overall, this period is often considered a golden age of antibiotic discovery and biotechnology.

However, that golden age has since then been declining. A major problem is the rise of resistance against our current repertoire of antibiotics [14]. Although antibiotic resistance is not an unnatural phenomenon [14-16], it is generally accepted that the human over-, under- and misuse of antibiotics is the main cause for the spread of resistance. Genes conferring resistance, e.g. by encoding a copy of a household gene that is insensitive to the antibiotics, are thought to quickly swap between bacteria by horizontal gene transfer (HGT) events. As a result, infectious diseases involving multidrug-resistant (MDR) pathogenic strains, are one the rise, most notably those involving the six ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa* and *Enterobacter* species), and MDR *Mycobacterium tuberculosis* (TB) [17]. Several variants of TB have been identified which are extensively-drug-resistant (XDR) and even total-drug-resistant (TDR), setting us back to the time before antibiotics were discovered [18-20]. Resistance to antibiotics is considered a major threat to our health by the WHO, with casualty projections in 2050 exceeding those of any other disease [21].

At the same time, traditional methods for drug discovery have been on the decline. HTS approaches toward drug discovery suffer greatly from high rediscovery rates of known compounds, with massive screening efforts only finding a handful of potential leads [22-24]. These approaches have become unreliable investments for the pharmaceutical industry, further decreasing the overall output [25]. It appears that much of the low-hanging fruit that could be discovered by HTS has been found, and now, new methods are necessary to keep discovering novel antimicrobials.

1 Genome mining for antibiotics

The advent of next generation sequencing (NGS) in the past two decades has made sequencing of entire genomes feasible and this has given the discovery of natural products new momentum. Due to newly developed techniques like Illumina, Nanopore and PacBio sequencing, it is now cheaper than ever to analyze the genomes of the bacteria and fungi that produce natural products. Conveniently, the genes encoding the enzymes capable of synthesizing natural products are clustered together into biosynthetic gene clusters (BGCs). By using biosynthesis genes characterized from known BGCs as queries, new, homologous BGCs can quickly be identified [26]. The first study investigating genome sequences of *Streptomyces coelicolor*, a model organism of the prolific antibiotic producers, the streptomycetes, discovered more than 20 BGCs, while the strain was known to only produce three compounds [27]. It quickly became clear that only a fraction of identified BGCs could be linked to a natural product. At time of writing, more than a million BGCs can be detected in publicly available genomic and metagenomic sequences [28], while only about 2,000 compounds have been directly linked to a BGC [29]. The BGCs without a known product, called cryptic BGCs, represent a vast reservoir of novel natural products that could be clinically relevant.

The existence of large amounts of uncharacterized BGCs raises the question about why their products have not been previously detected. A reasonable explanation is that secondary metabolite production is a costly process for a microorganism. Production levels must therefore be finely tuned in order to gain a competitive edge while not wasting resources, likely as a response to chemical signals in the environment [30, 31]. In Actinobacteria, this finetuning is accomplished by extensive regulatory networks that govern natural product production [30, 32]. A large number of two-component regulators and as many as 60 different sigma factors per strain make an organism capable of responding to many signals and scenario's.

In an attempt to exploit this, strategies have been developed to activate silent BGCs work by giving cultures specific signals they might encounter in their natural environment. These strategies include the use of molecular elicitors in

high-throughput elicitor screening (HiTES), screening with various carbon sources, applying stress conditions like starvation, and co-culturing strains together [33, 34]. Besides the novel natural products that are isolated in these studies, they also provide insight into what role the secondary metabolite might fulfill, knowing under what conditions it is activated. Alternatively, silent BGCs can be activated by genetically refactoring the BGC. Genes and operons can be rearranged and promoters can be replaced with strong, constitutive promoters to find optimal production conditions [35]. These efforts are generally more labor-intensive than using general chemical signals, requiring a large amount of genetic engineering to isolate the product or products of a single BGC. However, they do allow one to target a single BGC at a time, rather than evoke a more general response from an organism with potentially unwanted side effects. Combination of both approaches are required at different stages of investigation to completely understand the roles and products of novel BGCs.

The tradeoff between confidence and novelty in genome mining

Natural products are divided into classes or families based on their chemical makeup and biochemical origins. These products are built up from primary metabolites, such as amino acids and acetyl- or malonyl moieties. Natural products belonging to the same class share a common biosynthetic logic and homologous enzymes carrying out the reactions. For example, two major classes, called the non-ribosomal peptides (NRPs) and the type I polyketides (PKs), are synthesized by assembly-line machineries [36]. These are large enzyme complexes, that can be divided into modular units, each of which attaches one precursor molecule to a growing chain. The diversity of the resulting secondary metabolites is achieved either by using a large variety of precursors, or by applying additional tailoring to the products, such as glycosylation or cyclization. Other classes, like terpenes and type II PKs, rely on other enzymes to convert precursors into the final product, but these enzymes are still encoded by conserved genes [37, 38].

The classification of natural products provides a suitable framework for the purpose of identifying BGCs from genomic information. For each class of natural products, a set of enzymes can usually be identified that determines a family of natural products, and which can be used as “bait” to identify more

BGCs of that family. Tools like antiSMASH [39] and PRISM [40] use rule-based identification of these BGCs by targeting specific genes conserved among specific natural product classes. The surrounding region of hits found is the scanned for genes encoding additional tailoring enzymes, transporters, regulators, and immunity proteins, thereby identifying the BGC.

While these methods for genome mining have identified large amounts of BGCs, only BGCs with some relation to previously characterized ones will be detected. Methods extrapolating from known BGCs may give high-confidence output, but depending on the exact ruleset used, the BGCs detected will lack novelty [26]. Given the large diversity of natural product classes, it is tempting to speculate about the existence of completely novel classes and chemical scaffolds. To identify BGCs of these classes using only bioinformatics is a difficult challenge, as there should be at least some criteria to identify them. One tool has been described primarily for this purpose, called ClusterFinder [41]. ClusterFinder uses two collections of protein domains: one with those that are frequently associated with BGCs, and one with those that are present in other parts of the genome. It then uses a Hidden Markov Model (HMM) to identify regions of the genome that are enriched in domains indicating BGCs. Combined with large comparative genomics, a novel class of BGCs was identified and characterized. In general, using less restricted search criteria will increase the amount of candidate BGCs found, such as was the case with ClusterFinder. Among these may be more novel BGCs, but at the same time the amount of false positives will increase. This tradeoff between confidence and novelty is one that everyone undertaking of genome mining must consider carefully [26].

RiPPs form a diverse group of peptidic natural products

The ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important and diverse group of natural products, produced by all three branches of life. The unifying theme among all RiPPs is their biosynthetic logic: a precursor gene is translated into a precursor peptide, usually no longer than 100 amino acids. The precursor peptide is then extensively modified by a set of RiPP Tailoring Enzymes (RTE). RTEs usually recognize the peptide by binding a recognition sequence located outside the region that is modified.

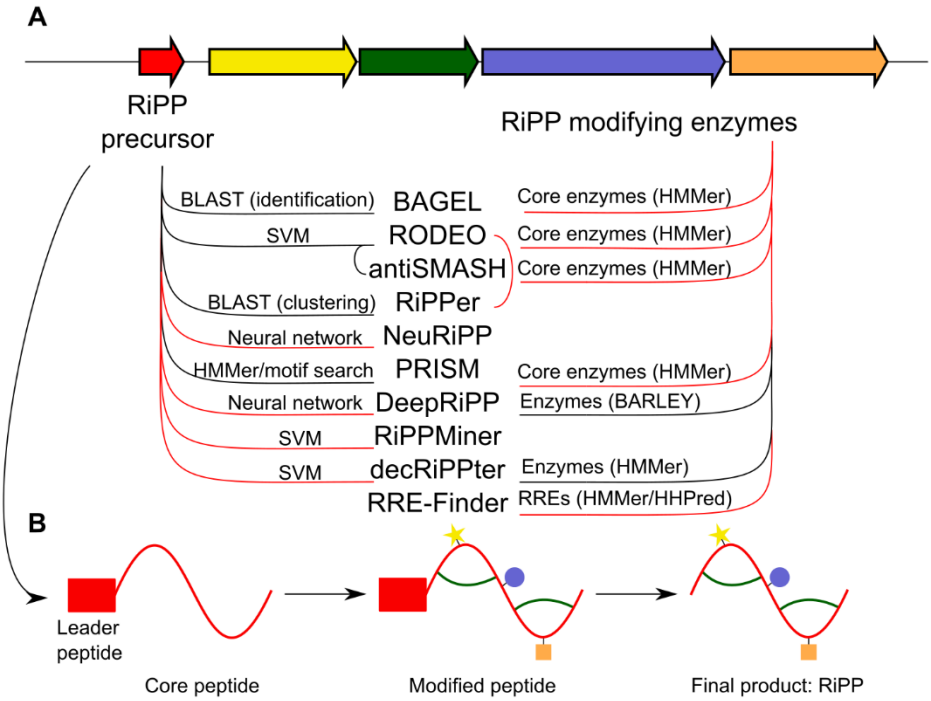


Figure 1. The conserved features of RiPP BGCs provides a framework for genome mining tools. A) A RiPP BGC typically has a gene for a precursor peptide (red) in an operon-like arrangement with genes for RTEs. These genes are used as targets for genome mining tools (see also Table 2). A red line indicates that the detection method is the primary detection method, while black lines indicate additional annotation. B) After translation, a precursor peptide is modified by RTEs, which use the leader peptide as a handle for peptide recognition. After modifications are applied, the leader peptide is cleaved off, resulting in the final product, the RiPP.

This recognition sequence is called the leader peptide if N-terminal, or follower peptide if C-terminal. Afterwards, leader and follower peptides are cleaved off, resulting in the mature RiPP [42-44] (Figure 1).

RiPPs are classified into subclasses or families, all of which share the biosynthetic logic with wildly different results depending on the precursor sequence and the enzymes involved. Every subclass of RiPPs typically has one characteristic modification. For example, all lasso peptides have at least a single crosslink, forming a small loop through which the amino acid chain is threaded [45]. Thiopeptides are all macrocyclized as the result of a 4+2 cycloaddition

between two dehydrated serine residues [46]. Lanthipeptides contain a thioether bridge between a non-C-terminal cysteine and a dehydrated serine or threonine residue [47]. Asides from core modifications, RiPPs may be tailored with several accessory modifications like disulfide bridges, acetylation, methylation and glycosylation. The combination of different precursor sequences and different possible combinations of modifications applied to them creates a wide diversity of different natural product (Figure 2).

As a rapidly expanding class of natural products, RiPPs represent an excellent candidate for further exploration to discover novel chemical scaffolds and antimicrobial leads. The last comprehensive review from 2013 [42] lists more than 20 different RiPP classes. Since then, several dozen RiPPs have been identified with modifications and genetic markers that set them apart from known RiPPs and thus form new RiPP families. An updated review published in 2020 expands this list to more than 40 different candidates [48] (Table 1). Given the number of novel RiPP classes discovered in the past decade, it is likely that many more exist. After all, it takes only a few evolutionary steps for a RiPP BGC to take shape. While the BGCs encoding the assembly lines of NRPS and PKS can be more than 100 kbp long, a RiPP BGC only needs to encode a small precursor and a single modifying enzyme for it to be a RiPP, and some are no larger than that [49]. RTEs could arise out of primary metabolism enzymes and protein modification enzymes with relatively few mutations and genetic arrangements. These BGCs would not need to be directly homologous to known RiPP classes, and would therefore be missed by the current methods of RiPP genome mining.

The diversity of RiPP BGCs reflects the diversity on the chemical level, and as such, there is no single method that effectively identifies all RiPP BGCs. Rather, new genome mining tools are still being developed and new approaches are still being experimented with. Uniquely, the small precursor genes provide a handhold suitable for all RiPP families. In the following section, we review the available tools and the way they differ in their approaches (Figure 1, Table 2), and highlight a few concentrated genome mining efforts has dramatically expanded the number of members of known RiPP families.

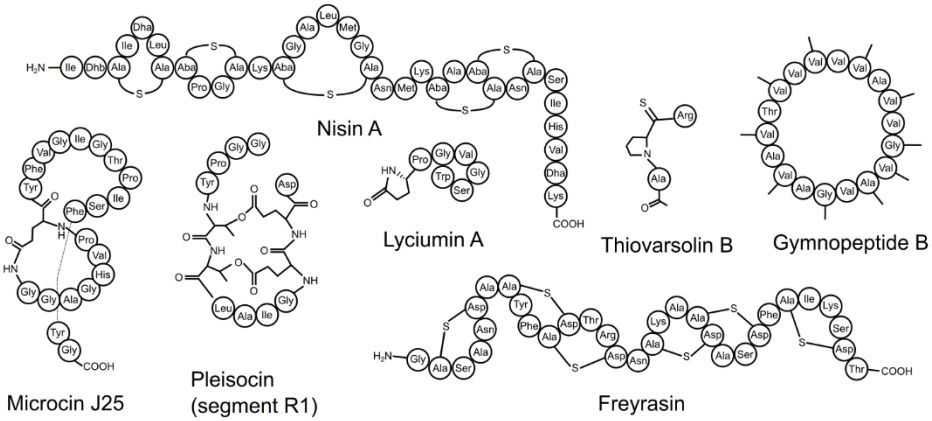


Figure 2. Examples of the rich chemical diversity of RiPPs. RiPP precursors are highly diverse in sequence and can be modified in many ways. Several old and new examples are shown here: nisin A (lanthipeptide) [50], lyciumin A (lyciumin) [51], thiovarsolin B (thioamitide) [52], gymnopeptide B (borosin) [51], microcin J25 (lasso peptide) [53], plesiocin segment R1 (omega-ester containing peptide/graspetide) [54] and freyrasin (ranthipeptide) [55].

Bioinformatic tools for homology-based genome mining of known RiPP families

A widely used approach for RiPP genome mining is to target the modifying enzymes encoded by the RiPP BGC. Most RiPP families contain conserved genes that encode the enzymes responsible the post-translational modification (PTM) characteristic of that RiPP family. The earliest methods of RiPP genome mining used simple BLAST or PSI-BLAST [56, 57] searches to identify homologs of genes encoding RTEs as a starting point for novel RiPP BGCs. These searches lead to the discovery of novel RiPPs with similar modifications as known RiPPs, like haloduracin, trichamide and capistruin [58-61]. BAGEL [62, 63] developed in 2006, was the first tool for automated genome mining of RiPPs. Since then, the more general algorithms antiSMASH [39] and PRISM [40, 64] have been developed, which allow genome mining for BGCs of any class of natural products, including RiPPs. All of these tools use profile Hidden Markov Models (pHMMs [65]) rather than BLAST queries, which are built from protein alignments of many members of a protein domain family. RiPP BGCs are identified based on manually set rules determining which pHMMs detected in

Table 1. All currently classified RiPP families. Adapted from [48].

Class	Example	Class-defining PTM(s)
Amatoxins/phallotoxins	Phalloidin	N-to-C cyclization, Cys-Trp crosslink
Amidinotides	Pheganomycin	Amidino amino acid containing peptides (ATP-grasp)
Atropitides	Tryptorubin	Aromatic amino acids crosslinked to give a noncanonical atropisomer
Autoinducing peptides	AIP-I	Cyclic ester or thioester
Bacterial head-to-tail cyclized peptides	Enterocin AS-48	N-to-C cyclization (DUF95 & ATP-grasp)
Borosins	Omphalotin	Amide backbone N-methylation (N-MT), N-to-C cyclization (POP)
Bottromycins	Bottromycin A1	Macrolactamidine (YcaO)
ComX	ComX168	Indole cyclization and prenylation
Conopeptides	Conantokin G	Peptides produced by cone snails
Crocagins	Crocagin A	Indole-backbone cyclization
Cyanobactins	Patellamides	N-terminal proteolysis (PatA protease)
Cyclotides	Kalata B1	N-to-C cyclization, disulfide(s) (AEP)
Dikaritins	Ustiloxin	Tyr-Xxx ether crosslink (UstY)
Epipeptides	YydF	D-amino acids (rSAM)
Glycocins	Sublancin 168	S, O, or N-glycosylation of Ser/Cys
Graspetides	Microviridin J	Macrolactones/lactams (ATP-grasp)
Lanthipeptides	Nisin	(Methyl)lanthionine, labionin
Lasso peptides	Microcin J25	Macrolactam with threaded C-terminal tail (Asn synthetase homolog)
Linaridins	Cypemycin	Dhb, no lanthionines
Linear azol(in)e containing peptides (LAPs)	Microcin B17	Cys, Ser, or Thr derived azol(in)es (YcaO)
Lipolanthines	Microvionin	C-terminal labionin/avionin containing peptide and N-terminal FAS/PKS segment
Lyciumins	Lyciuman A	Pyroglutamate, Trp-Gly crosslink
Methanobactins	Methanobactin	Oxazolones (DUF692)
Microcin C	Microcin C	Aminoacyl adenylate or cytidylate with a phosphoramidate linkage (ubiquitin E1 homolog)
Mycofactocin	Mycofactocin	Val-Tyr crosslink (rSAM)
Orbitides	Cyclolinopeptide A	N-to-C cyclization; no disulfides
Pantocins	Pantocin A	Glu-Glu crosslink (PaaA)
Pearlins	Thiaglutamate	aa-tRNA derived (PEARL)
Proteusins	Polytheonamide	Nitrile hydratase LP
Pyrroloquinoline quinones	PQQ	Glu-Tyr crosslink (rSAM)
Ranthipeptides	Freyrasin	Sulfur-to-non-Cα thioether crosslink (rSAM)
Rotapeptides	TQQ	Oxygen-to-α-carbon crosslink
Ryptides	RRR	Arg-Tyr crosslink (rSAM)
Sactipeptides	Subtilosin	Thioether crosslink to alpha-carbon (rSAM)
Spliceotides	PlpA	β-amino acids (rSAM)
Streptides	Streptide	Trp-Lys crosslink
Sulfatyrotides	RaxX	Tyrosine sulfation
Thioamitides	Thioviridamide	Backbone thioamide (YcaO)
Thiopeptides	Thiostrepton	[4+2] Cycloaddition of two Dha
Thyroid hormones	Triiodothyronin	Triiodothyronin

close proximity of one another form a RiPP BGC. For example, type I lanthipeptide BGCs can be identified by targeting the protein domains present in the modifying enzymes LanB (PF04738/PF14028) and LanC (PF05147), both of which should be be found encoded by genes located near one another on the genome.

These methods excel at the detection of known RiPP families, for which the RTEs responsible for the hallmark modifications have been identified. Completely novel RiPP families which lack these modifications cannot be detected, however. Nevertheless, these BGCs may still specify RiPPs that are novel because they encode different precursor. Examples of studies investigating these are numerous, and only a handful are mentioned here. For example, antiSMASH-based genome mining led to the discovery of streptocollin, a type IV lanthipeptide [66]. A study investigating the RiPP BGCs of 629 actinobacterial genomes using BAGEL3 detected 477 different RiPP BGCs[67]. Most of these contained unique precursor peptides (e.g. lanthipeptides: 276 out of 301 unique, lasso peptides: 62 out of 67 unique, LAPs: 43 out of 48 unique). A more thorough investigation into only lanthipeptide-like BGCs in Actinobacteria detected 1,163 in 830 genomes. These were further grouped into 100 gene cluster families (GCFs) based on sequence and RTEs encoded. Interestingly, several GCFs encoded RTEs not previously associated with lanthipeptide BGCs, like O-methyltransferases, NRPSs and PKSs [68].

Although detection of RTEs is relatively straightforward, the challenge in the automated detection of RiPP BGCs lies in the correct annotation of the genes encoding precursor peptides. Gene finding algorithms such as Glimmer [69, 70] and Prodigal [71] frequently miss the open reading frames (ORFs) that encode precursor peptides, as they can be as small as 15 nucleotides [49]. BAGEL4, the latest of version of BAGEL, takes additional steps to increase the number of precursor genes detected [62]. In a genomic area that contains genes encoding RTEs, all intergenic small ORFS (<= 72 nt) are extracted, translated and BLASTed against a database of the core sections of known RiPP precursor peptides. This method provides a more detailed annotation of precursor genes. However, since detection is based on known core peptides, completely novel precursor peptides will not be detected by this method.

A more sophisticated approach for precursor detection is taken by RODEO [45, 55, 72-74]. RODEO allows a user to analyze the genomic context of any gene matching a query domain on NCBI. Given its accession number, genes in the context of a query gene are annotated with Pfam and TIGRFAM [75, 76]. The tool was first used to mine genomes for lasso peptide BGCs, using a rule-based system based on detected protein domains. To better detect precursor genes, a machine learning classifier called a Support Vector Machine (SVM) was trained to distinguish between lasso precursor peptides and other peptides. This SVM was trained on several hundreds of features, such as frequency of specific amino acids or amino acid pairs, charge and hydrophobicity. The prediction of this SVM was combined with heuristic scoring of a given small ORF to effectively detect precursor genes. The same model for precursor detection was integrated into antiSMASH, as of version 4.0 [77].

The prerequisite of both a precursor peptide and a specific protein domain has been used to mine for thiopeptides [72], sactipeptides and ranthipeptides [55], lasso peptides [45], lanthipeptides[73] and linaridins [74]. These genome mining efforts have expanded the list of candidate BGCs belonging to each family, and led to the discovery of novel RiPPs, such as citrulassin. In theory, the same process could be applied to any RiPP family, as long as sufficient precursor sequences are available to train an SVM. This method therefore lends itself mostly to well-characterized classes. Interestingly, like in the study described above, the BGCs detected contain a wide assortment of different putative modifying enzymes, which occasionally co-occur within the core RTEs (1-25%). These are predicted to encode for e.g. acetyltransferases, glycosyltransferases, FAD oxidoreductases or methyltransferases. The existence of RiPPs with additional tailoring is not without precedence, exemplified by reports of acetylated lasso peptides [78], glycosylated lanthipeptides [79] and lipidated lanthipeptides [80]. Characterization of these secondary tailoring enzymes provides interesting opportunities to further chart the chemical landscape covered by RiPPs. In addition, given that many RTEs recognize via the leader peptide, these enzymes may be capable of modifying other RiPPs as well, allowing one to further tweak their properties with synthetic biology [81].

1 Explorative domain-based genome mining expands and defines novel RiPP families

The rule-based genome mining used by high-confidence RiPP genome mining tools described above is an effective way to expand known RiPP families, for which a conserved set of genes has been identified. However, for many newly discovered RiPP families, sometimes only a single example BGC is known. Highly homologous BGCs can easily be identified, but more interesting is perhaps the identification of a minimal set of genes that is required for a specific modification unique to the RiPP. Identification of these in novel contexts could lead to the discovery of novel RiPPs that belong to the same or related families. In the following section, we will describe several studies aimed at genome mining of novel RiPP families as well as the discovery of related RiPP families by shared modifications.

Discovery and expansion of omega-ester peptides

The first member of the omega-ester peptides of RiPPs was microviridin, a cytotoxic RiPP with three intramolecular omega-ester or omega-amide crosslinks, which was isolated in 2008 [82]. While initial studies focused on identifying microviridins from the cyanobacterial genus *Mycrocystis* [83], the characterization of two homologous BGCs from *Plesiocystis pacifica* and *Bacillus thuringiensis serovar huazhongensis* led to the identification of plesiocin and the thuringinin group, respectively [54, 84]. Like microviridin, these RiPPs also contained omega-ester and omega-amide crosslinks, although the number of crosslinks and the overall topology of the products were different. As all BGCs encoded a homologous ATP-grasp ligase, these proteins could be used as a query for genome mining of novel BGCs of the same type [85]. This search resulted in 5,276 homologous proteins. Inspection of context of the encoding genes for possible precursor peptides resulted in the identification of 12 groups of new omega-ester containing peptides. This is a sizable increase in the number of candidate BGCs of this family, especially considering that only four ATP-grasp ligases were used as a query. However, the authors note that ~3,200 protein hits could not be assigned a specific precursor, which could be false positives.

Table 2. Tools available for RiPP genome mining.

Name	BGC identification target	Method description	Reference
antiSMASH	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with RODEO's SVMs.	Blin *et al*. [39]
BAGEL	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with BLAST and a known precursor database.	Van Heel *et al*. [62]
RiPP-PRISM	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identifies precursor peptides with HMMer and a motif search.	Sknnider *et al*. [40, 64]
RODEO	Core enzymes	Identifies RiPP BGCs with core enzymes per class. Identification of precursor peptides with SVMs.	Tietz *et al* [45], Schwalen *et al* [72], Hudson *et al* [55], DiCaprio *et al* [86], Walker *et al* [73], Georgiou *et al* [74].
RiPPer	Any enzyme	Identifies RiPP BGCs with any query enzyme. Prioritizes candidate precursor peptides with prodigal-short and BLAST-based clustering.	Santos-Aberturas *et al*. [52]
RiPPMiner	Precursor peptides	Identifies and classifies precursors with a single SVM.	Agrawal *et al*. [87]
NeuRiPP	Precursor peptides	Identifies precursors with a neural network.	De Los Santos. [88]

Table 2 (continued).

DeepRiPP	Precursor peptides	Identifies and classifies precursors and BGCs with a neural network (NLPPrecursor). Predicts products and estimates novelty based on genetic context and known modifications (BARLEY). Compares metabolomics and matches MS/MS spectra to predicted products (CLAMS).	Merwin *et al*. [89]
DEREPLICATOR	NA	Clusters peptide natural products based on MS/MS spectra.	Mohimani *et al*. [90]
VarQuest	NA	Matches peptide natural products to their variants with unknown modifications based on MS/MS spectra.	Gurevich *et al*. [91]
MetaMiner	Core enzymes	Identifies RiPP BGCs with antiSMASH. Predicts products based on genetic context and known modifications. Matches predicted products to MS/MS spectra.	Cao *et al*. [92]

Novel thioamidated RiPPs found by a bait-based approach combined with precursor clustering

The above example highlights how the combination of a putative precursor and a single RTE of interest as a query allows identification of new types of RiPP BGCs. RiPPer was developed to generalize this procedure for any type of RTE [52]. The search starts with a query RTE, which is used find the genes encoding their homologs within a given database. To identify possible precursor genes, the surrounding region (+- 8 kbp) of each hit is reannotated with an adapted version of the genefinding software Prodigal called prodigal-short. The adapted version has a lower cut-off point for the minimum size of a gene (60 nt instead of 90), allowing it to more effectively identify RiPP precursors. All short genes

(between 60 and 360 nt) are scored by the prodigal score, which is increased if it is on the same strand as the query RTE. This approach does not take into account the sequence of the precursors, but was still able to detect 94.1% and 96.7% of two test sets of precursor peptides from microviridin and lasso peptide genome mining studies [45, 83]. However, because multiple candidate precursor peptides are reported per BGC, the total number of precursor peptides identified by this method was several times higher than the training set. To increase the specificity of detected precursors, the authors clustered the precursor peptides detected based on sequence similarity. Large groups of conserved peptides are more likely to be encoded by real ORFs, and indeed, the largest group of peptides was found to overlap with previously identified precursor peptides. Peptides encoded by spurious ORFs are less likely show significant similarity to one another, and therefore small groups of precursor peptides can be discarded as false positives.

The authors used the *tfua* gene as a query RTE, which encodes a protein thought to be involved in the formation of thioamidated RiPPs, like thioviridamide [93, 94]. The nearby candidate precursors were clustered, which resulted in thirty networks, two of which were encoded by thioviridamide-like BGCs. Of each of the networks, the genomic context of the genes encoding these precursor peptides for each of the networks was manually analyzed. Based on this, the authors highlighted 12 different candidate RiPP families. Experimental characterization of one of these families led to the discovery of the thiovarsiolins. These RiPPs share the small thioamidated amino acids backbones, but otherwise share no similarity with thioviridamide-like RiPPs. While a single RTE like the one encoded by the *tfua* gene may not be a reliable marker, in conjunction with a group of predicted precursor peptides, it leads to many promising and yet to be uncovered RiPPs.

Radical SAM enzymes as a versatile RiPP marker

Genes that are shared between different RiPP families make interesting targets for the discovery of novel RiPP families. Examples include the cyclase domain, which is shared between all different lanthipeptides subtypes [47], and the YcaO protein, which is involved in the maturation of bottromycins, LAPs, thiopeptides and thioamitides, and can catalyze three different reactions [95]. A protein that

is involved in the maturation of a wide number of RiPPs is the radical S-adenosyl methionine (rSAM) enzyme. The rSAM enzyme uses a radical S-adenosyl methionine (SAM) intermediate to catalyze a range of different reactions [96], ranging from simple methyltransferase reactions in bottromycins [97] to conversions of L-D stereochemistry in proteusins [98] to crosslinks in sactipeptides [99]. Radical SAMs can be involved in primary metabolism as well as secondary metabolism, and are highly diverse, which makes them somewhat unreliable for RiPP genome mining [100]. In conjunction with a precursor peptide, however, or by targeting a specific clade more closely associated with RiPP metabolism, many more interesting RiPP BGCs could be uncovered.

rSAMs themselves are as diverse as their modifications, but pHMMs have been developed for rSAMs in specific niches, such as those responsible for the maturation of sporulation killing factor A (TIGR04403) [101]. Identification of a specific subclade of rSAMs and using this as a query has previously led to the discovery of mycofactocin [102]. The authors in this study identify a subclade of without any assigned function, and use partial phylogenetic profiling to discover protein families associated with this subclade. A group of proteins was found strongly associated with this particular clade, which led to the discovery of the new RiPP mycofactocin. Expanded genome mining using the rSAMs associated with sactipeptides as a query, in conjunction with SVM-based precursor identification, has led to the discovery of ranthipeptides [55]. In addition, rSAMs were identified in a RiPP BGC in *Pleurocapsa sp. PCC 7319*, which was found to specify α-keto-β-amino acid-containing RiPPs, although the exact genome mining method was not reported [103].

Last of all, by building a query for the detection of genes encoding rSAMs and quorum-sensing type regulators, a study reported the presence of these enzymes in many genomes [104]. The rules for detection were based on the BGC of streptide [105]. The production of streptide is regulated by quorum sensing, and its BGC contains two genes encoding a two-component system that mediate this regulation. A search in Streptococcal genomes revealed that the three genes encoding the regulatory system and a rSAM appeared to co-occur frequently. One of these BGCs was experimentally characterized, leading to the discovery of WGK, rotapeptides and ryptides [104, 106-108].

1

The RiPP recognition element as a guide for novel RiPP family discovery

The core element that is associated with most RiPP families is the RiPP Recognition Element (RRE). RREs were first discovered as short domains of roughly 100 amino acids that showed high structural similarity to PqqD [109]. Experimental characterization showed that the element could bind precursor peptides, and in many cases was required for RiPP maturation. Thorough analysis of all known RiPP classes revealed that the element is present in roughly half of all RiPP classes discovered to date. This domain may either be present in a small, stand-alone protein, or be fused to another enzymatic domain. In many cases, the domain is essential for RiPP maturation, even for stand-alone RREs, suggesting they act as a guide for other modifying enzymes to aid in precursor peptide recognition. In some cases, though, the domain is vestigial [110]. This similarity was discovered with HHPred [111], an algorithm for the comparison of pHMMs and secondary structure predictions made by PSIPRED [112]. As of now the only known method for the reliable detection of RREs is via HHPred, which is a time-consuming algorithm taking several minutes per query. Nevertheless, as this element promises to be highly specific towards RiPPs, but is still independent of any specific RiPP family, it would make an excellent target for RiPP genome mining.

Shared enzymology between RiPPs and non-RiPPs leads to discovery of 3-thiaglutamate

In a few unusual cases, genome mining for RiPP BGCs can lead to the discovery of non-RiPP BGCs. For example, a recent genome mining effort aimed at finding lanthipeptide-like gene clusters in Actinobacteria identified several genes for LanB, without a nearby gene encoding LanC [68]. LanB is involved in the dehydration of serine and threonine residues in type I lanthipeptides. This reaction takes place in two steps, where first the hydroxyl group is glutamylated using tRNA-glutamate as the donor, and then eliminated by a separate protein domain [47]. Strikingly, these LanB homologs did not contain an elimination domain. Characterization of the BGC showed that the precursor peptide is used catalytically to produce 3-thiaglutamate [113]. A cysteine was attached to the C-terminus of a precursor-like peptide by the LanB homolog, converted to 3-thiaglutamate and then excised. The precursor peptide could still be used as a

recognition site for the other enzymes, three of which even contained RREs, but was otherwise not consumed for the production. The LanB homologue was renamed as a peptide aminoacyl-tRNA ligases, and these products were collectively called pearlins.

Precursor-centric genome mining of RiPP BGCs

The one thing that all RiPP BGCs have in common is the gene encoding a precursor peptide, and this is a vital element in all RiPP genome mining strategies. Usually, the precursor gene is in operon with the genes encoding the modifying enzymes, so detection of a precursor gene will result in detection of a new BGC. The precursor peptides show a high amount of variability, however, which limits precursor-centric genome mining using similarity-based methods, even within a single RiPP family. In addition, the precursor genes may be very small (<= 100 nt), and therefore missed by automatic gene annotation programs. Nevertheless, identification of the precursor peptide is highly valuable as it greatly speeds up experimental characterization.

The reliability of machine-learning methods to detect encoded precursor peptides could re-invigorate precursor-centric genome mining. In general, any such method should have a low false positive rate in order to be useful for precursor-based genome mining. After all, the number of small open reading frames that are precursor candidates far exceeds the number of expected RiPP precursors. A *Streptomyces* genome will likely contain only 1 to 5 RiPP BGCs [67]. Assuming that on average, a small ORF can be found between each pair of genes, there will be 8,000 small ORFs to analyze. A RiPP classification method with a low false positive discovery rate of 1% and a perfect true positive discovery rate of 100% will detect all positive hits, but also roughly 80 negative hits, outweighing the positive hits 40 to 1. Therefore, these models should either have a very low false positive rate or be integrated into larger pipelines in order to be useful for novel RiPP discoveries. This proves a difficult challenge, especially considering that there are relatively few positive examples of RiPP precursors compared to the number in the negative training set. Besides RODEO, three other tools use different machine-learning models to detect and classify precursor peptides: RiPPMINER, NeuRiPP and DeepRiPP.

Like RODEO, RiPPMINER uses a trained SVM to distinguish precursor peptides from other peptides, although unlike RODEO, a single SVM is used for all RiPP classes [87]. The predicted precursor peptides are then classified to their respective RiPP class using a multi-class SVM. The tool can identify well-characterized RiPP families such as lanthipeptides, lasso peptides and linaridins, by training on a manually curated training set of more than 500 RiPP precursors. The precision and sensitivity for identification were 0.93 and 0.90, respectively. While these values are fairly high, the tool might still report a fair number of false positives when used as a starting point for genome mining for the reasons stated above. Nevertheless, RiPPMINER was used in a pipeline with the ClusterFinder algorithm [41] and transcriptome data analysis to detect novel candidate RiPP BGCs in the fungus *Trichoderma* spp [114] showing that tools with relatively low sensitivity can find novel results when used in conjunction with other datasets.

NeuRiPP [88] uses a neural network rather than an SVM for precursor peptide classification. The model trained here takes the raw sequence as input, instead of calculated features like amino acid frequency or hydrophobicity. Several network architectures were tested, of which the parallel convoluted neural network (CNN) performed best. The tool separated precursor peptides from non-precursor peptides with an accuracy of 99.84% on the entire training set. In addition, it was capable of detecting bottromycin precursor peptides, despite the fact that these were lacking from the training data. This suggests that the network was capable of identifying precursor-specific features from the raw sequences across different RiPP families. The author suggested integration of the tool into RiPPer [52] as an effective means to filter precursor peptides in a class-independent manner. As a proof of concept, the precursor peptide networks associated with the *tfua* gene, also identified by RiPPer (see above) were analyzed. Of the 12 peptide networks prioritized by the authors, 8 were identified as precursor peptides by NeuRiPP, despite the fact to no precursor peptides of these RiPP classes were present in the training data.

DeepRiPP is an assembly of three modules and the first tool described that fully integrates precursor-centric mining with comparative genomics and metabolomics tools [89]. The first module, NLPPrecursor, uses a Universal

Language Model Fine-Tuning (ULMFiT) neural network to detect encoded precursor peptides. This is a neural network architecture used for language processing that has shown to be highly effective in building models from training sets with low amount of data. It further classifies detected precursor peptides into specific RiPP families. BARLEY, the second part of the pipeline, then tries to estimate all possible RiPPs that may be produced from a detected precursor peptide, using known modifying enzymes in the surrounding genomic context, and a machine-learning model to estimate the cleavage site. All possible final products are compared to each other and to a database of known RiPPs and the distance between all products is calculated, either using the genomic predictions of the final products of the known products. In this way, known products can be easily dereplicated, and BGCs can be identified whose predicted products are distant from known RiPP BGCs, increasing the odds that new RiPP variants will be discovered. In the final step, extracts of all the strains analyzed made under various growth conditions are analyzed by LCMS. CLAMS, the third module of the pipeline, then tries to find correlations between the absence and/or presence of predicted RiPP products with detected LCMS peaks. The fragmentation data of these peaks are also matched to possible peptide fragments calculated from the precursors. By mining 65,421 bacterial genomes, 19,498 new possible RiPP products were identified. The authors then extract 463 of these strains under various conditions, creating a metabolomics database of 10,498 extracts. In these extracts, three new RiPPs could be identified with CLAMS, belonging to the lanthipeptide, lasso peptide and thiopeptide RiPP families, respectively.

In summary, machine-learning methods provide an excellent way to accurately predict precursor peptides. These methods can supplement RiPP genome mining of known classes, as is done in RODEO and antiSMASH. More importantly, precursor-centric genome mining can lead to the discovery of novel RiPP BGCs, without needing to first identify characteristic modifying enzymes. Since precursor-centric genome mining carries an inherent risk for a high amount of false positives, the results of these classifiers need to be carefully analysed. Indeed, these methods proved most successful when combined with more extensive data analysis, such as analysis of genomic context, comparative genomics, transcriptomics and metabolomics.

1 Integration with omics in larger pipelines

As illustrated by DeepRiPP described above, the integration of metabolomics data could accelerate the identification of RiPPs. In contrast to normal proteins, however, RiPPs contain modified amino acids and are rarely linear. For known RiPP classes, the modifications can be predicted based on genomic information. Predicted peptide fragments containing these modifications can be matched to the spectra with tools like DEREPLICATOR [90] (recently updated with NPS [115]) and CLAMS (available within the DeepRiPP pipeline [89]). DeepRiPP is perhaps the most integrative pipeline for RiPP discovery. Besides structure prediction based on the identification of known modifications, it also combines comparative genomics with comparative metabolomics, to prioritize peaks whose presence/absence matches that of the BGCs of interest. However, it could even be extended further, by also considering transcriptomics and proteomics data. Elicitors should therefore be added to activate the expression of cryptic BGCs, whereby comparative metabolomics combined with transcriptomics or proteomics will allow linkage of BGC expression profiles to changes in metabolites. This will allow scientists not only to observe more metabolites than under one specific growth condition, but also to predict which metabolites are produced by which BGCs.

A major challenge for automated MS/MS analysis that remains is dealing with new modifications. VarQuest [91], an extension of DEREPLICATOR, can identify peptide variants based on known peptides, even if these variants contain unknown modifications. MetaMiner [92] combines genomics and metabolomics to predict precursor modifications and find associated spectra, which can contain unknown modifications. Completely *de novo* identification of novel RiPPs with only unknown modification has yet to be explored by tools like these, but represents a sizable computational challenge. Even so, just matching a small sequence of unmodified amino acids to part of a candidate novel RiPP precursor is a valuable addition to more explorative RiPP searches. Identified, novel precursors could then be fed back to the training data of the precursor classifiers, creating an iterative process in which the classifiers will become increasingly specific and tuned toward a larger variety of RiPP classes.

Outline of the thesis: towards the detection of completely novel RiPP subclasses

The diversity in tools described above highlights the challenges in RiPP genome mining. Traditional RTE-based approaches do an excellent job at increasing the number of members of a RiPP family, as long as a well-defined set of enzymes characteristic of that class is known. Increasingly well-polished methods for the identification of RiPP precursors make an excellent supplement to these methods, providing additional information to properly identify the final product. Integration with metabolomics further streamlines RiPP identification, and can unite metabolomic and genomic information.

Most of the tools described above, however, do not focus on the discovery of completely novel RiPP subclasses. The discovery of these is a difficult challenge – after all, if no modifying enzymes can be used as queries, BGCs cannot be discovered with methods that target specific domains. Even though RiPPer gives a user more freedom in this regard, it still relies on the selection of a query domain, which biases the results. However, one feature is always present in almost all RiPP subclasses: each RiPP BGC should encode a precursor, and contain at least one modifying enzyme. These domain-independent features could be exploited to mine RiPPs in a less restricted manner, and lead to the discovery of new RiPP subclasses.

Detection of precursors in a class-independent manner could be accomplished with machine-learning-based classifiers. The high confidence of these classifiers has already led to precursor-based genome mining, but mostly of known RiPP subclasses. Interestingly, NeuRiPP was capable of predicting some RiPP precursors of RiPP classes for which it had not been trained. Apparently, this neural network is capable of identifying some property of combination of properties that distinguishes precursors of any class from other peptides. If this set of properties is shared among not-yet-discovered RiPPs, it is possible that precursor-based genome mining could lead to the identification of completely novel RiPPs, as no restrictions would be placed onto the genomic contexts in terms of known RTEs.

In **Chapter 2**, we describe a novel tool for the detection of RiPP Recognition Elements (RREs) – the domain that is shared among the most different bacterial RiPP classes. Specific profile Hidden Markov Models have been designed for each of the different types of RRE. This allows for high-confidence detection of RREs of known classes in precision mode. A second mode, called exploratory mode, is based on HHPred and can detect more distantly related RREs, at the cost of computational power and more false positives. These methods allow the detection of novel RRE-enzyme fusions, that can lead to the discovery of novel RiPP subclasses.

In **Chapter 3**, we describe an innovative tool for the identification of novel RiPPs, called decRiPPter (Data-driven Explorative Class-independent RiPP TrackER). This tool utilizes an SVM-based RiPP precursor classifier, which is independent of RiPP subclass, and can therefore be used to identify novel RiPPs. Instead of focusing on the amino acid sequence, decRiPPter examines the genomic contexts of encoded precursor peptides for possible RTEs, associated with RiPPs or otherwise. The results have been combined across many genomes to form candidate RiPP families. The work underlines the power of artificial intelligence approaches for the discovery of new candidate bioactive molecules.

Chapter 4 describes the application of decRiPPter for the identification of a novel class of lanthipeptides. BGCs of this family are widespread among Actinobacteria and Firmicutes, but so far their function was unknown. Experimental characterization of a gene cluster from *Streptomyces pristinaespiralis* revealed that it indeed specifies a novel RiPP, that we called pristinin A3. Pristinin A3 contains many modifications also found in other types of lanthipeptides. Lanthipeptides are further classified by their modifying enzymes. Since the modifying enzymes involved in the generation of this family of RiPPs are novel, we classified this RiPP as a new lanthipeptide subclass, called class V. The complex two-dimensional structure of pristinin A3 was elucidated by mass spectrometry and NMR.

In **Chapter 5,** a different type of RiPP BGC is characterized. This BGC shows distant similarity to known RiPP BGCs of different classes, as it contains genes encoding a radical SAM enzyme and an ATP-grasp ligase. Still, the presence of these genes alone do not place it clearly in any known RiPP subclass.

In addition, two well-conserved genes encoding predicted precursors show a unique motif that is repeated multiple times. A detailed bioinformatic description is given explaining the homologies of this BGC and it's relation to other known RiPP BGCs. In addition, experimental work is presented describing the activation of the BGC and the analysis of chemical extracts aimed at identifying the final product.

In **Chapter 6**, the results are summarized and reviewed in a general discussion. The explorative approach taken towards RiPP genome mining, and the use of machine learning classifiers for this purpose, are reviewed. the challenges encountered in this thesis are described and possible solutions are proposed. Further extenstions for the decRiPPter pipeline are outlined, which could futher help future efforts in class-independent RiPP genome mining. Also, the RiPP BGCs that were studied in this work are further discussed, including their possible classifications with regard to currently accepted schemes.

1

2

RRE-Finder: a genome-mining tool for class-independent RiPP discovery

Alexander M. Kloosterman*
Kyle E. Shelton*
Gilles P. van Wezel
Marnix H. Medema
Douglas A. Mitchell

* These authors contributed equally to this work.

The work described in this chapter is published as:

Kloosterman, et al., *RRE-Finder: a genome-mining tool for class-independent RiPP discovery.* mSystems, 2020. **5**(5).

Abstract

Many ribosomally synthesized and posttranslationally modified peptide classes (RiPPs) are reliant on a domain called the RiPP recognition element (RRE). The RRE binds specifically to a precursor peptide and directs the posttranslational modification enzymes to their substrates. Given its prevalence across various types of RiPP biosynthetic gene clusters (BGCs), the RRE could theoretically be used as a bioinformatic handle to identify novel classes of RiPPs. In addition, due to the high affinity and specificity of most RRE-precursor peptide complexes, a thorough understanding of the RRE domain could be exploited for biotechnological applications. However, sequence divergence of RREs across RiPP classes has precluded automated identification based solely on sequence similarity. Here, we introduce RRE-Finder, a new tool for identifying RRE domains with high sensitivity. RRE-Finder can be used in precision mode to confidently identify RREs in a class-specific manner or in exploratory mode to assist in the discovery of novel RiPP classes. RRE-Finder operating in precision mode on the UniProtKB protein database retrieved ~25,000 high-confidence RREs spanning all characterized RRE-dependent RiPP classes, as well as several yet-uncharacterized RiPP classes that require future experimental confirmation. Finally, RRE-Finder was used in precision mode to explore a possible evolutionary origin of the RRE domain. The results suggest RREs originated from a co-opted DNA-binding transcriptional regulator domain. Altogether, RRE-Finder provides a powerful new method to probe RiPP biosynthetic diversity and delivers a rich data set of RRE sequences that will provide a foundation for deeper biochemical studies into this intriguing and versatile protein domain.

Introduction

As of late 2019, nearly one-quarter of a million prokaryotic genomes were publicly available in the National Center for Biotechnology Information (NCBI) genome databases [116]. This vast genomic resource has accelerated the pace of natural product discovery, with a recent surge of interest pertaining to the ribosomally synthesized and posttranslationally modified peptides (RiPPs) [42]. RiPP biosynthesis starts with the ribosomal synthesis of a linear precursor peptide. The genes for RiPP precursor peptides are often short, hypervariable in sequence, and composed of two parts—an N-terminal leader region and a C-terminal core region. With a few notable exceptions, the precursor peptide is genetically encoded adjacent to one or more genes encoding proteins that bind with high specificity and affinity to the leader region of the precursor. This interaction facilitates subsequent posttranslational modification of the core residues . After modification is complete, the leader region is enzymatically removed and the mature RiPP product is exported from the producing organism [117] (Figure 1). The exact nature of the posttranslational modifications is used to categorize RiPPs into individual classes, of which nearly 40 have been reported [42]. For example, lanthionine linkages define the lanthipeptide class, while oxazol(in)e and thiazol(in)e heterocycles define the linear azol(in)e-containing peptide (LAP) class [118, 119].

Many RiPP biosynthetic proteins recognize and bind their cognate precursor peptide through a domain known as the RiPP recognition element (RRE) [109]. The RRE consists of a conserved secondary structure of three N-terminal alpha helices followed by a three-stranded beta sheet. The precursor peptide binds in a cleft between the third alpha helix (α3) and the third beta strand (β3), forming an ordered, four-stranded, antiparallel beta sheet (Figure S1). RRE domains can exist either as discretely encoded proteins (<100 residues) or as fusions to a larger protein domain [99, 109, 120-122] . In cases where a RiPP biosynthetic gene cluster (BGC) encodes a discrete RRE protein, this protein binds the leader peptide and serves as a scaffold for recruiting the necessary modifying enzymes. All characterized RREs share structural similarity to PqqD, which is a protein involved in synthesis of pyrroloquinoline quinone (PQQ), a

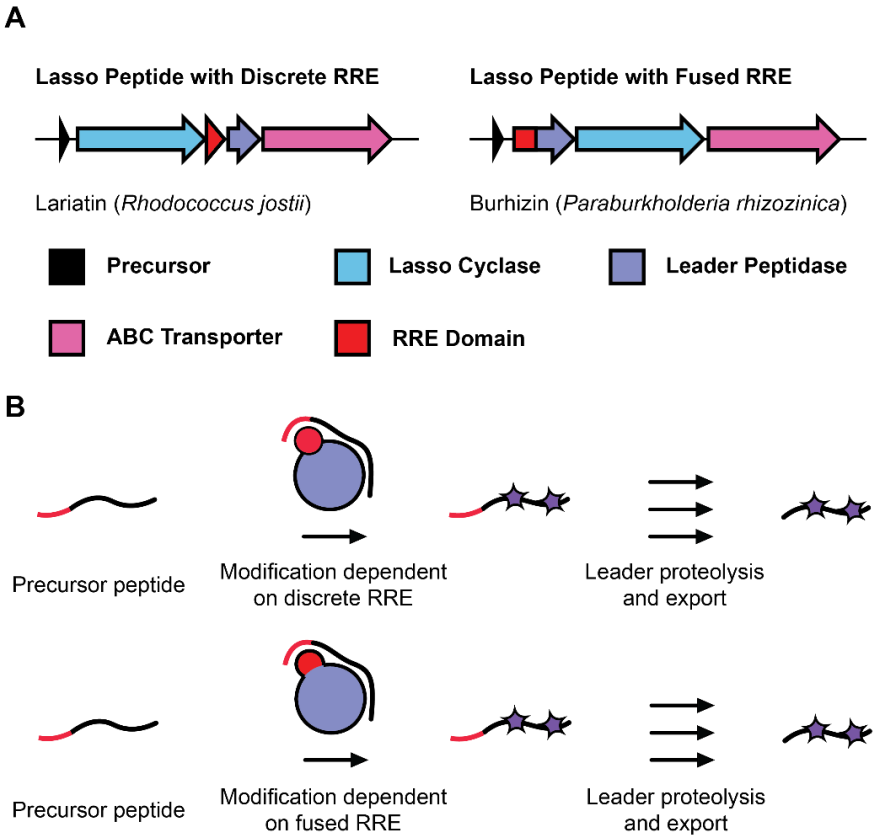


Figure 1. RRE-dependent RiPP biosynthesis. (A) RiPP BGCs contain one or more short precursor peptide(s); their genes often lie adjacent to those for the modifying enzymes, leader peptidases and proteins for immunity/export (often ABC transporters). RRE domains are found as discrete polypeptides or fused to larger biosynthetic proteins. (B) Modifying proteins bind the leader region of the precursor peptide using RRE domains. Post-translational modifications are then installed on the core region of the precursor peptide.

redox cofactor produced by many prokaryotes [123]. Thus, the existence of a PqqD-like protein encoded near regulators, enzymes, and transporters is strongly indicative of an RRE-dependent RiPP BGC. The prevalence of PqqD-like proteins in RiPP BGCs led to the discovery of the RRE domain and its conservation across RiPP classes in 2015 [109]. Before this, the importance of leader peptide recognition was established in the biosynthesis of a few RiPPs, such as nisin (lanthipeptide) and streptolysin S (LAP) [124, 125]. In addition, an

RRE-containing protein from microcin C7 biosynthesis (MccB) was cocrystallized with its cognate leader peptide in 2009, but owing to RRE sequence divergence, it was not appreciated at the time that other RiPP classes employ a similar domain [126].

Consistent with the rapid expansion of characterized RiPP BGCs, a diverse collection of modifications and enzymatic domains are found among the ~40 known RiPP classes. However, the lack of a common genetic feature remains a major obstacle in the bioinformatic detection of novel RiPP classes. The fact that RRE domains are prevalent in prokaryotic RiPP BGCs provides an opportunity. Of the ~30 known RiPP classes produced by prokaryotes, over 50% contain an identifiable RRE domain (Table S1 and Table S2). Considering that the RRE domain appears to be the most conserved class-independent feature in RiPP BGCs, it theoretically could be used as an imperfect but useful bioinformatic handle to expand known RiPP sequence-function space by identifying new RRE-dependent RiPP classes.

The strategy outlined above is complicated by the sequence diversity of the RRE domain [99, 109, 122, 123]. For example, if a pairwise sequence alignment method (e.g., NCBI BLAST [127]) is used to compare RRE domains from two unrelated RiPP classes, sequence similarity will frequently not be detected, particularly in cases where the RRE domain is fused to a larger protein. The most appropriate Pfam [128] model (a family of proteins sharing sequence similarity) for defining the RRE domain is PF05402, which extensively covers bona fide PqqD proteins from PQQ-producing BGCs. PF05402 incompletely retrieves RRE-containing proteins from only a few other RiPP classes (e.g., lasso peptides and sactipeptides), and indeed, most RREs from other RiPP classes have no representation in this Pfam [129-131] (Figure S2). These results underscore the inability of a single bioinformatic model to capture the breadth of RRE sequence diversity. Owing to the fact that RREs share considerable structural similarity, HHpred [111] is a more sensitive algorithm for detecting RRE domains. HHpred detects remote protein homology by aligning profile hidden Markov models (pHMMs; a model that defines amino acid frequency for a protein family) and comparing their (predicted) secondary structures. RREs were originally detected using this method by analyzing several RiPP-modifying

enzymes, which showed consistent homology to PqqD [109]. However, HHpred requires generation of a multiple sequence alignment (MSA) and secondary structure prediction using PSIPRED [112]. These steps require several minutes of computing time per protein query, rendering the process unattractive for larger data sets and precluding global analyses of RRE diversity. In this work, we report a customized tool that permits the rapid and accurate detection of RREs in known and potentially novel RiPP classes with the principal goal of directing natural product hunters to the most fruitful areas of the RiPP sequence-function space.

Results and Discussion

Development of RRE-Finder

This work presents RRE-Finder, a new tool for mining RRE domains from microbial genomes. RRE-Finder has two modes of operation (Figure 2). The first is precision mode, which employs a set of 35 custom pHMMs designed to detect RRE domains in a class-dependent manner (Figure S3 and Table S3). The precision-mode pHMMs are primarily based on known RiPP classes—in most cases, representative RRE-containing proteins from these classes have been verified to bind their cognate precursor peptide through biophysical experiments, such as X-ray crystallography or fluorescence polarization binding assays (Table S2). The second mode, exploratory mode, uses a truncated version of the HHpred [111] pipeline with a custom database of detected RREs. Depending on the end user's objective, RRE-Finder can be used in precision mode to accurately predict the presence of an RRE domain as well as the likely RiPP class in which the precursor peptide belongs. Alternatively, in exploratory mode, the user can retrieve a wider array of putative RRE-containing proteins to assist in the discovery of novel RRE-dependent RiPP classes. RRE-Finder accelerates the process of identifying RRE domains by several orders of magnitude compared to HHpred. Precision mode, for instance, can analyze >5,000 protein sequences per second (Table S4). In addition to 29 core models based on known RiPP classes, precision mode includes 6 auxiliary models based on high-confidence, novel RiPP classes. We justified the inclusion of these models based on repeated observation of RRE domains within RiPP-like genomic contexts across multiple prokaryotic species. The 35 pHMMs that comprise precision mode are provided in Data Set S2 (available at https://figshare.com/articles/Dataset_S2_HMM_files/12030651).

In general, for RiPP classes where an extensive survey of the bioinformatic space has been performed (e.g., lasso peptides [45, 86], sactipeptides and ranthipeptides [55], and thiopeptides [72]), custom pHMMs were built by first visualizing sequence space through use of a sequence similarity network (SSN) for all RRE-containing proteins in the data set [132]. SSN visualization using Cytoscape [133] facilitated selection of the most diverse and nonredundant subset of RRE primary sequences for seed sequence alignment.

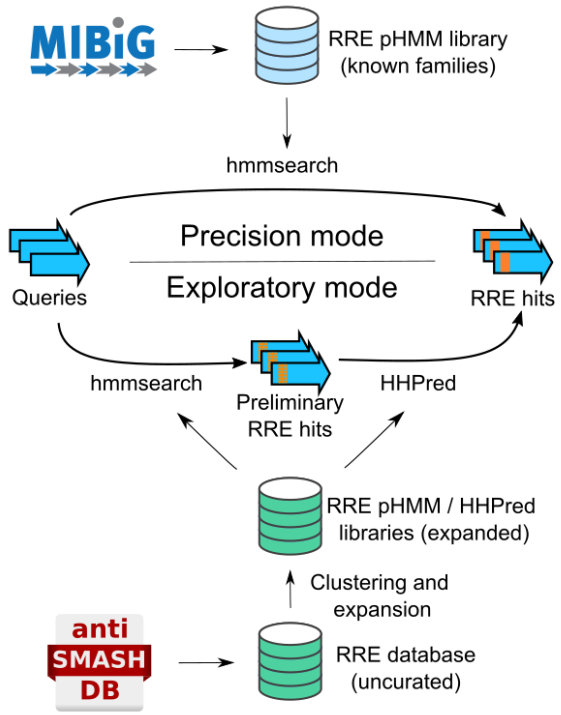


Figure 2. RRE-Finder employs two modes for RRE detection. Precision mode (top) of RRE-Finder uses a set of pHMMs to accurately predict RREs. These pHMMs are based on characterized RRE domains for individual RiPP classes, either from published datasets or from the MIBiG database. Exploratory mode uses a combination of pHMMs and a truncated HHpred pipeline (including secondary structure prediction) to facilitate the identification of divergent RRE sequences (albeit with a higher false-positive rate).

In cases where a published data set was available for a given RiPP class, model prediction accuracy was gauged by using hmmscan (from the HMMER3 suite [134]) on the relevant data set using bit scores of 15, 25, and 35 (referred to here as tolerant, moderate, and stringent cutoffs). A given pHMM was considered acceptable if >95% of RRE-containing proteins within the data set were retrieved by the model at a bit score of 25 (Table S5). In cases where a deep bioinformatic profiling of a RiPP class had not been previously published or where a mature natural product is not known (i.e., clusters predicted by the auxiliary models), seed alignment input sequences were gathered using PSI-

BLAST [57] to find diverse homologous sequences to a representative sequence from each given class. The generated pHMMs were considered valid if an hmmsearch of the UniProtKB database [135] with a bit score cutoff of 25 gave only hits within BGCs with architectures similar to those of the target class. In addition, characterized data sets of RiPP proteins (e.g., lanthipeptides [68, 73], lasso peptides [45, 86], and sactipeptides [55]) were used to test auxiliary models using hmmscan analysis. Models giving few or no hits were considered to have acceptably low false-positive rates.

Exploratory mode, on the other hand, was built for the detection of RRE domains with greater sequence divergence from those detected by precision mode. For this mode, we employed a variation of the HHpred pipeline to detect structural similarity to RRE domains. HHpred uses a clustered UniProt database (uniclust30) [136], which comprises a small, representative set of all UniProt protein sequence diversity. Query proteins are compared to the uniclust30 database to generate a representative protein family for the query, and the consensus sequence of this representative protein family is compared to those of other protein families. This search also incorporates comparison of (predicted) secondary structures. As such, HHpred can detect distantly related sequences and overlap in secondary structures between a query protein and the UniProt database. However, the vast search space used far exceeds what is necessary if the goal is to detect RRE domains.

To accelerate the HHpred pipeline for RRE detection, we first built a smaller, more specialized HHpred database, consisting of ∼2,400 diverse RRE sequences. These sequences were gathered by retrieving 5,000 RiPP BGCs from the antiSMASH database [137] using HHpred. Rather than manually curating the retrieved RREs in a class-specific manner, as was done for precision mode, all detected RREs were indiscriminately included. The only manual curation carried out was the removal of helix-turn-helix-containing proteins and other transcriptional regulators. While these proteins may display structural similarity to RREs, they are not involved in RiPP biosynthesis and therefore were excluded from the data set. The selected RREs were supplemented with 7 RREs from LAP BGCs and an RRE from a proteusin BGC, as no BGCs from these RiPP classes were present in the antiSMASH database.

The collection of ~2,400 RREs was used to build databases for two filtering steps. For the first filter, all RREs were clustered into representative protein families with MMSeqs2 [138], resulting in 377 RRE families. These RRE families were further enriched by querying each family against the uniclust30 database using HHblits, an iterative search tool from HHpred [139]. For each of the 558 resulting RRE families, custom pHMMs were constructed, allowing an initial filtering step with hmmsearch [134]. The second filtering step functions in a manner similar to that of HHpred. However, rather than using the uniclust30 database to retrieve a protein family for a query, we employed a smaller, custom HHpred database consisting of the ~2,400 RRE sequences retrieved from the antiSMASH database and their related protein families retrieved by HHblits. When this custom database is used, only protein queries that are homologous to one of the 377 clustered RRE families will return results. For queries lacking homology, no protein family would be found in the database, effectively filtering out such sequences. Finally, exploratory mode compares the family of proteins homologous to a query protein to three RRE structures in the Protein Data Bank (PDB entries 5V1T, 5SXY, and 3G2B). Any proteins showing homology to these models are output as putative RRE domains. In all, by employing a small, custom library of RRE sequences, exploratory mode significantly accelerates detection of RREs relative to the standard HHpred pipeline.

Model validation against the MIBiG database.

As an initial test of accuracy, RRE-Finder was evaluated in precision and exploratory modes against the MIBiG database [140]. This database contains characterized BGCs for ~2,000 natural products, including polyketides, nonribosomal peptides, and RiPPs. All proteins within the MIBiG set (version 1.4) of RiPP (n = 242) and non-RiPP BGCs (n = 1,575) were analyzed by RRE-Finder at tolerant, moderate, and stringent bit scores (Figure 3).

In general, both precision and exploratory modes accurately predicted the presence of RRE domains in >90% of the RRE-dependent RiPP BGCs. Taken together, both modes retrieved 93% (115/122) of RRE-containing proteins found by HHpred (Table S6). With increasing bit score stringency, the number of RRE sequences retrieved decreased in both RiPP and non-RiPP BGCs, as expected (Figure 3). At all bit score cutoffs, exploratory mode predicted more

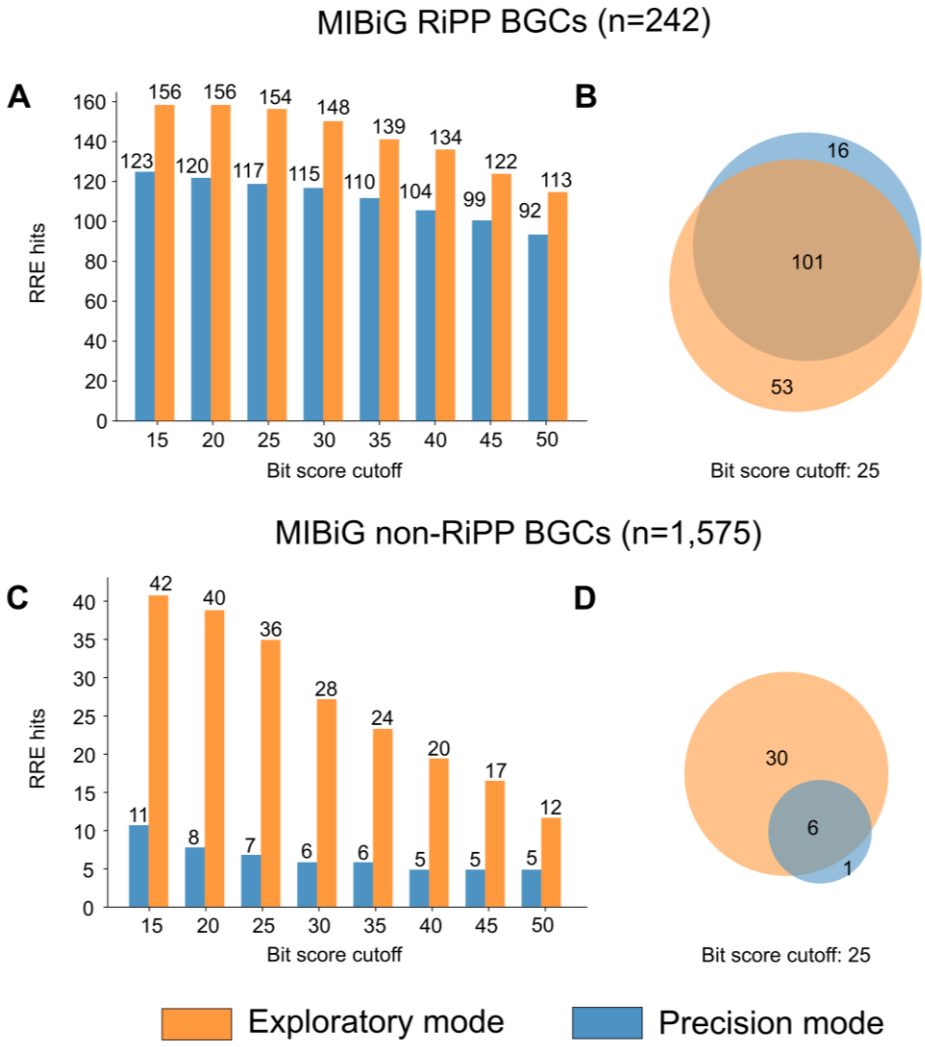


Figure 3. MIBiG validation of RRE-Finder. Both modes were used to retrieve RRE-containing proteins in 242 RiPP BGCs (A and B) and 1,575 non-RiPP BGCs (C and D) from the MIBiG database. With increasing bit score stringency, the number of RRE detected decreased in both types of BGCs (A and C). At a bit score of 25, exploratory mode of RRE-Finder detects most of the RREs found by precision mode in RiPP BGCs (B), as well as several other RREs. However, the number of RREs detected in non-RiPP BGCs is lower for precision mode compared to exploratory mode (D).

RRE domains in RiPP BGCs (higher true-positive rate than precision mode), while precision mode retrieved fewer proteins from non-RiPP BGCs (lower false-positive rate than exploratory mode). After further analysis, we chose a bit score cutoff of 25 as a compromise between precision and recall. At this cutoff, most of the RREs found within the MIBiG set by precision mode were also found by exploratory mode (101/117) (Figure 3). Only the RREs of linear azol(in)e-containing peptides (LAPs) [118] and streptides [104] proved more difficult to detect by exploratory mode (Table S6). The inability of exploratory mode and HHpred to reliably predict LAP RRE domains may reflect a large diversity of leader peptide recognition sequences within this class that is better captured by the five distinct LAP models used by precision mode.

By contrast, precision mode detected only 66% (101/154) of the RREs retrieved by exploratory mode. A notable number (n = 17) of the RRE-containing proteins not detected by precision mode were those contained in LanB-like proteins, which are found in certain lanthipeptide and thiopeptide BGCs. It has been shown that the LanB RRE domain found in thiopeptide BGCs is possibly vestigial, as the cognate leader peptide is not required for catalytic processing [110]. Exploratory mode also detected several (n = 14) RREs fused to dehydrogenase enzymes present in cyanobactin, LAP, and thiopeptide BGCs, which were not detected by precision mode. These RREs may also be vestigial; thus, precision mode does not include models for identifying these RRE-like domains. HHpred analysis similarly does not detect many of these potentially inactive RREs; thus, exploratory mode provides the best coverage of functional and vestigial RRE domains in this instance. We note that some of the RREs detected by exploratory mode, such as those from the thioamide-containing RiPP and pheganomycin pathways, are presumed to be functional but have yet to be experimentally validated (Table S6).

While exploratory mode detects a greater number of RREs, it also displays a higher false-positive rate (e.g., proteins retrieved from known non-RiPP BGCs). The false positives primarily consisted of helix-turn-helix domains and proteins with homology to known RRE-containing proteins that occur in non-RiPP contexts, such as radical S-adenosylmethionine (rSAM) enzymes (Table S7). Many DNA-binding regulators possess a helix-turn-helix domain,

which are structurally homologous to RRE domains (Figure S4). Indeed, most RRE domains analyzed by HHpred show homology to known DNA-binding domains and regulatory elements (e.g., PDB entries 3DEE, 2G9W, and 2OBP). Because regulatory proteins are not known to bind or modify RiPP precursor peptides, RRE-Finder includes an option to filter results that correspond to such domains.

RRE-Finder operating in either mode retrieved LanB-like proteins within polyketide BGCs. There is precedence for the assimilation of RiPP-modifying enzymes into polyketide pathways [68], although the RRE domain within these proteins may be vestigial (Figure S5, Table S7). Thus, retrieval of proteins outside canonical RiPP BGCs may not always constitute a false positive. Further biochemical validation is required to confirm or refute a functional RRE in these instances.

Finally, some pHMMs employed by precision mode were generated largely using RRE sequences from the MIBiG database. In these cases, validation against MIBiG alone is not sufficient to confirm or refute whether these models exhibit appropriate recall and precision. As an orthogonal means of precision mode validation, we ran hmmscan on ~5,000 RiPP BGCs from the antiSMASH database used to generate the exploratory-mode database [137]. As previously stated, these BGCs primarily belong to the lanthipeptide, thiopeptide, LAP, sactipeptide, and lasso peptide classes. Because this collection of BGCs includes RRE-dependent and RRE-independent RiPPs (e.g., class II to IV lanthipeptides) [141], there are BGCs anticipated to not be retrieved by precision mode. These clusters were purposely included in the analysis as a negative control. All proteins within the 5,000 BGCs were scanned by precision mode at tolerant, moderate, and stringent bit scores. The percentages of scanned BGCs predicted by precision mode to contain an RRE were 90%, 87%, and 83%, respectively. The 10% of BGCs not predicted to contain an RRE by precision mode were manually examined, with the majority belonging to RiPP classes that are RRE independent. Some BGCs also contained regulatory elements that represent false positives by HHpred; these proteins were appropriately not retrieved by precision mode. Thus, precision mode accurately predicts the presence of RREs in an unbiased collection of BGCs and appropriately omits RRE-independent RiPP clusters.

Defining the scope of RRE-dependent RiPP BGCs

Next, we profiled the extent to which the RRE domain is present within sequenced genomes by mining the entire UniProtKB database [135]. Using hmmsearch at a bit score threshold of 25, precision mode retrieved ~25,000 proteins (~13,000 nonredundant sequences) (Figure 4). A parallel search using exploratory mode with regulators filtered out yielded ~35,000 nonredundant RRE-containing proteins, almost completely encompassing the proteins retrieved by precision mode. As expected, the numbers of proteins retrieved by precision mode is larger than has been previously reported for virtually all RiPP classes, owing to on-going genome sequencing. For example, the thiopeptide precision model is the top-scoring model for more than 600 of the retrieved UniProtKB proteins, an ~25% increase from the most recent bioinformatic survey of thiopeptide BGCs [72]. In other cases, the number of retrieved proteins for a given model is misleading. For example, the precision mode model for discretely encoded lasso peptide RREs is the top-scoring model for almost 8,000 of the retrieved proteins. However, subsequent analysis revealed that only ~4,000 of these sequences co-occur with the requisite leader peptidase and lasso cyclase. This number is more consistent with the most recent lasso peptide survey, which reported ~3,000 lasso peptide BGCs [86, 142]. Proteins retrieved by the discrete lasso peptide model often co-occur with other common RiPP enzymes, such as rSAM enzymes which represent ~300 of the false positives. Thus, we caution that the number of proteins retrieved by any given model should not be equated to the number of BGCs specific to a particular RiPP class without analysis of the local genomic neighborhood. Full information on proteins retrieved by precision mode is available in Data Set S3 (https://figshare.com/articles/Dataset_S3_RRE_domains/12568193).

Figure 4 shows the number of retrieved proteins at tolerant, moderate, and stringent bit score cutoffs, as a measure of precision model specificity. Notably, due to partial model overlap in closely related RiPP classes (e.g., PQQs/lasso peptides and LAPs/thiopeptides/cyanobactins), the overall numbers of retrieved proteins for these models do not drastically increase going from moderate to tolerant bit scores. Thus, the majority of “false positives” detected by precision models at lower significance cutoffs represent RRE-dependent RiPP BGCs of a separate RiPP class.

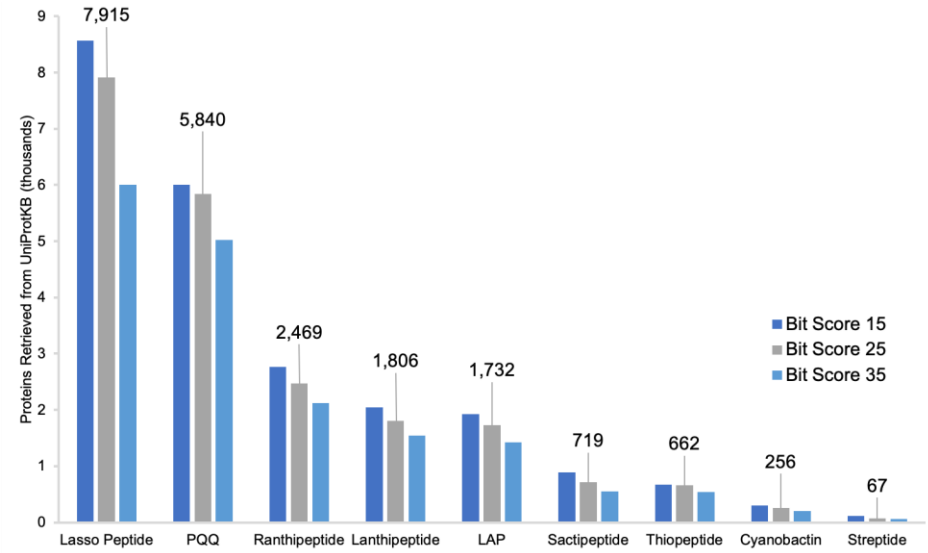


Figure 4. Summary of proteins retrieved from UniProtKB using precision mode. The number of proteins retrieved from the UniProtKB database are summarized for several classes of RiPPs. A scan of the entire UniProtKB database of non-redundant proteins was carried out at three bit scores. In cases where a given UniProt accession was retrieved by more than one precision model (due to partial model redundancy), the protein was only counted toward the model of higher significance. For classes with more than one precision mode HMM (e.g. LAPs and sactipeptides), the numbers presented are the sum of proteins retrieved by each individual model. Full data on proteins detected by each precision mode model is available in Dataset S3 (https://figshare.com/articles/Dataset_S3_RRE_domains/12568193). LAP, linear azol(in)e-containing peptide. PQQ, pyrroloquinoline quinone.

The excised RREs from all proteins identified by precision mode were visualized using a sequence similarity network (SSN) [132]. The SSN confirms known relationships between RREs in separate RiPP classes. For example, discretely encoded lasso peptide RREs (referred to as the B1 or E protein) group separately from RRE-leader peptidase fusions (known as the B2 or B protein), consistent with a different recognition sequence for these two varieties of lasso peptide (Figure 5; Figure S6) [45, 86]. In contrast, the heterocycloanthracins (LAPs) cluster more tightly with thiopeptides than other LAPs. This relationship was expected given that heterocycloanthracin and thiopeptide BGCs feature an RRE domain fused to an ocin-ThiF-like protein (TIGR03693) that delivers the

peptide substrate to the biosynthetic enzymes [118, 143]. In other LAP pathways, the RRE is fused to members of TIGR03882 [109, 118, 143, 144]. Members of TIGR03882 recognize the peptide substrate through the RRE and perform cyclodehydration reactions, whereas these functions are carried out by separate proteins in thiopeptide and heterocycloanthracin clusters

Another method to view RRE relatedness is through model redundancy (Figure S7). In cases where there is overlap in the proteins retrieved by multiple models, the redundancy is reflective of RREs in these classes binding their cognate leader peptides through similar sequence motifs. Similarly, lack of model overlap is indicative of a divergent leader peptide recognition sequence. For example, at a moderate bit score, there is virtually no overlap between the lanthipeptide-associated RRE domains with any other RiPP class, reflective of a unique recognition sequence not yet observed elsewhere [141, 145] (Figure S7). We note that model redundancy, particularly in RiPP BGCs with more than one RRE-containing protein, may suggest a similar recognition sequence on the cognate leader peptide. For example, the 3-thiaglutamate (pearlin RiPP class) BGC contains three proteins predicted to contain an RRE. The precision-mode pHMMs for these proteins display greater redundancy with each other than with any other model. This suggests comparable specificity of these RRE domains, as dictated by the α3 and β3 regions, and that these RREs likely bind the same region of the precursor peptide. However, this hypothesis will require further experimental evaluation.

Evolution of the RRE domain

Sequence similarity between recognition sequences in closely related RiPP classes suggests that the RRE domain emerged once and then diverged to recognize a variety of leader peptides. Because the leader peptide binds as an ordered beta-strand between the α3 helix and β3 strand of the RRE, substitutions of key α3 and β3 residues logically tune the RRE specificity toward the cognate peptide substrate. Analysis of residue-level conservation between RREs of divergent RiPP classes reveals that the α3 and β3 regions exhibit higher levels of residue conservation than the remainder of the domain, presumably due to selective pressure to conserve leader peptide-RRE contacts. This holds true even when closely related RiPP classes, such as LAPs and thiopeptides, are

compared (Table S8). The other regions of the RRE, which are not directly involved in leader binding, show lower levels of conservation.

A representative phylogenetic tree of excised RRE domains retrieved by precision mode (bit score of 25) is consistent with the hypothesis that the RRE domain coevolved with the leader peptide to provide specificity in all RRE-dependent RiPP classes (Figure S8). The tree does not include all proteins retrieved by precision mode; rather, 10% of the proteins contained within each SSN cluster (Figure 5) were included, along with all singletons, to generate a diversity-maximized collection of sequences spanning all RRE-dependent classes. The tree employs a helix-turn-helix DNA-binding protein as an outgroup (PDB entry 3DEE), as this protein scores well in HHpred searches of characterized RRE proteins, such as PqqD and LynD. As previously mentioned, it is plausible that the RRE domain evolved from DNA-binding regulatory elements, given the shared secondary structure and the similar function of these domains to specifically bind a stretch of DNA or a peptide (Figure S4). Unsurprisingly, the diversity-maximized tree shows a subset of the discrete lasso peptide RREs branching directly from the helix-turn-helix outgroup. Although discrete RREs called by this model are dispersed throughout the tree, the subset branching most directly from the outgroup is mostly representative of the false positives discussed previously (proteins not co-occurring with lasso peptide machinery). This may suggest that some of these false positives are DNA-binding proteins more closely related to true RREs (either in RiPP or non-RiPP contexts) and that discrete RREs evolved from these regulators. These proteins could also represent discrete RREs from currently uncharacterized RiPP classes. Furthermore, the tree shows clades of fused RRE domains branching off from discrete RREs as separate events for most RiPP classes. Some fused RRE types (e.g., fused lasso peptide RREs, ranthipeptides, and pantocins) form monophyletic clades branching from parent clades with discrete RREs. Other classes, like the lanthipeptides, are dispersed throughout many clades. This may indicate that fusion of the RRE domain to other domains occurred as separate events, even within some RiPP classes. These data are also consistent with the observed domain architectures, as some classes employ N-terminally fused RRE domains, while others exhibit C-terminal fusions (e.g., proteusins).

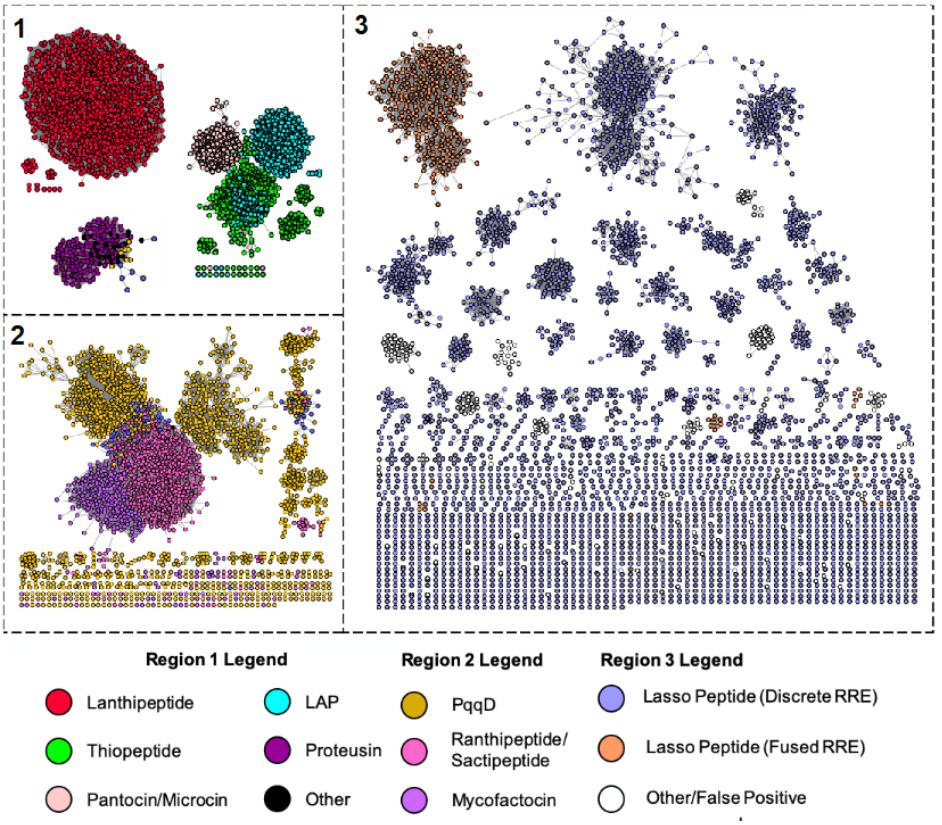


Figure 5. Sequence similarity network of UniProtKB proteins retrieved by precision mode. Shown is a RepNode60 SSN at an alignment score of 22 (sequences with >60% amino acid identity are conflated to a single node and edges represent a BLAST expectation value better than 10^{-22}). Proteins are colored based on the best-fit model by which they were detected. White nodes in region 3 represent proteins that were retrieved by the discrete lasso peptide RRE model but do not co-occur with the requisite leader peptidase and lasso cyclase. The discrete lasso peptide RREs clustering with sactipeptides and ranthipeptides in region 2 are discretely encoded RRE proteins that co-occur with radical SAM enzymes. The SSN was generated using the Enzyme Similarity Tool (https://efi.igb.illinois.edu/efi-est/) [132].

Using RRE-Finder to identify novel RiPP clusters

Theoretically, the sequence space retrieved by exploratory mode and the auxiliary models of precision mode encompasses RRE-containing proteins from yet-undiscovered RiPP classes. To explore this sequence space, divergent clusters mined from UniProtKB were manually examined for novel RiPP contexts. All proteins retrieved were grouped based on their best-fit Pfam

model. Since we expected many regulatory elements or proteins with helix-turn-helix domains among the hits, we filtered these sequences after the first step of the exploratory pipeline, reducing the required computational time.

Among the remaining detected proteins, RRE-Finder reveals several potentially novel RiPP clusters with new gene architectures containing both discrete and fused RRE domains (Figure S9). Included in these clusters are RRE-protein fusions that are not present in known classes, such as RRE-glycosyltransferase fusions and RRE-glutathione *S*-transferase fusions (Figure S10, Table S9 and Table S10). Of the nine potential RiPP BGCs shown in Figure S10, four encode rSAM enzymes, which are found across several RiPP classes [55]. The presence of rSAM enzymes in conjunction with predicted RREs is suggestive of a RiPP BGC. However, of the nine BGCs, only three contained probable precursor peptides (small genes of <150 amino acids, co-occurring with the RRE-containing protein), while four other BGCs contained precursor candidates predicted by RODEO. Therefore, manual curation of potentially novel BGCs found by RRE-Finder is strongly recommended. An overall sequence similarity network of the UniProtKB proteins accessed by exploratory mode is provided in Figure S9.

To date, almost no RiPP classes have been discovered using solely a bioinformatic approach. The mycofactocin class was initially predicted through a bioinformatic study on then-uncharacterized rSAM enzymes [102]. In addition, the ranthipeptide class was defined solely using bioinformatics (as SCIFF [for “six cysteines in forty-five residues”] peptides) [146]; however, this class was incorrectly assumed to be part of the existing sactipeptide class [55]. In other cases, bioinformatics analyses have been used to expand diversity within known RiPP classes; for example, the streptide class has been expanded to include enzymes that diverge from the class-defining Lys-Trp cross-linking enzymes [104, 105]. Also, one new RiPP class—the α-keto β-amino acid-containing peptides—and one RiPP-like class—the pearlins—were discovered through bioinformatic means [103, 147]. These classes, however, were discovered through first identifying a divergent member of a known RiPP biosynthetic enzyme, rather than through a truly unbiased bioinformatic discovery. We expect that RRE-Finder will enable such discoveries.

RRE-Finder incorporation into antiSMASH and RODEO

To encourage the use of RRE-Finder, the algorithm has been made publicly available as a command-line tool for Linux operating systems at https://github.com/Alexamk/RREFinder. Protein queries can be supplied in FASTA or GenBank format. The tool is also capable of analyzing and updating antiSMASH and DeepBGC output files [148]. Precision mode of RRE-Finder will be incorporated into the next release of antiSMASH. We further have incorporated the precision mode of RRE-Finder into RODEO [45], a genome-mining tool for RiPP discovery that provides genomic neighborhood visualization and prediction of precursor peptides. Protein-coding sequences within the genetic locus are annotated according to Pfam and TIGRFAM models to identify conserved domains and predict function. With the “include RRE scoring” function enabled, proteins with an identifiable RRE are annotated, along with their E-value significance. Both the command line version of RODEO (https://github.com/the-mitchell-lab/rodeo2) and the user-friendly Web tool version (http://rodeo.scs.illinois.edu) have been upgraded with the capabilities of RRE-Finder precision mode.

Conclusion and final perspectives

RRE-Finder rapidly and accurately detects RRE domains within known and potentially novel RiPP classes. Although not all RiPP classes are RRE dependent, the majority of prokaryotic RiPP classes are, including the largest known classes (i.e., class I lanthipeptides, lasso peptides, and ranthipeptides). RiPP natural products are a prime candidate for pathway engineering, as precursor peptides and their cognate modifying enzymes are all genetically encoded, typically within one BGC. However, efforts to bioinformatically predict RiPP BGCs lag behind those for predicting polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) BGCs, due to a lack of strongly conserved protein domains spanning multiple RiPP classes. Through precision mode of RRE-Finder, we have shown that characterized RiPP classes contain more members than currently reported, although analysis of the genomic neighborhood should be performed to confirm class identity. Precision mode can further be employed, particularly with a tolerant bit score threshold, to predict novel RRE domains, such as those predicted by the auxiliary models. Finally, using RRE-Finder in exploratory mode reveals a set of ~35,000 proteins that are predicted to contain an RRE, suggesting that additional classes of RRE-dependent RiPPs remain to be uncovered.

Materials and Methods

Generation of precision mode models

Precision mode was generated to accurately predict the presence of RRE domains specific to characterized RiPP classes, as well as RRE domains in selected bioinformatically predicted RRE-dependent RiPP clusters. There are 29 models employed by precision mode of RRE-Finder (not including auxiliary models), each specific to a given discrete or fused RRE protein within a characterized RiPP class (see Figure S3 for represented classes). Each precision model consists of a custom profile hidden Markov model (pHMM). To build each pHMM, five to 20 representative sequences were selected from a given RRE class for seed sequence alignment. For several RiPP classes, an extensive bioinformatic survey of biosynthetic gene clusters has been conducted. When available, these data sets were employed to select seed sequences. The data sets included those describing known gene clusters for lanthipeptides [73], lasso peptides [45], thiopeptides [72], cyanobactins [149], bottromycins [150], linear azol(in)e-containing peptides (LAPs, including heterocycloanthracins, plantazolicins, nitrile hydratase-like leader peptides [NHLP]-derived RiPPs, Nif11-derived RiPPs, goadsporins, and cytolysins) [118], pantocins/microcins [151], and radical *S*-adenosylmethionine-derived RiPPs (including sactipeptides, ranthipeptides, quinohemoprotein amine dehydrogenases, and streptides). In these cases, sequence diversity was evaluated by generating a sequence similarity network (SSN) using the Enzyme Function Initiative Enzyme Similarity Tool (EFI-EST) [132] and visualizing the SSN with Cytoscape [133]. Five to 20 sequences (depending on number of clusters in the SSN) were selected from divergent clusters on the SSN.

Bioinformatic data sets were not available for the following RRE-dependent RiPP classes: PQQ [123], proteusins, mycofactocins, trifolitoxins, α-keto β-amino acid-containing peptides, and pearlins. In these cases, a list of homologous sequences to a canonical gene were obtained with position iterative BLAST searching (PSI-BLAST) [57] with three iterations and an E-value cutoff of 0.05 in November 2019 using the GenBank nonredundant protein sequence database. Once a list of homologous sequences was obtained, an SSN was generated in the manner described above, and diverse sequences were selected for seed sequence alignment.

Seed sequences were analyzed for the presence of an RRE domain using the HHpred Web tool (https://toolkit.tuebingen.mpg.de) [111]. A protein was considered to contain an RRE if part or all of the protein matched a PqqD model (either PDB entry 5SXY or 3G2B) with 80% probability or greater. All proteins containing RRE domains were excised *in silico* to contain only the residues matching the relevant PqqD model. Excised RRE sequences were then aligned using MAFFT 7.450 [152]. MAFFT alignments were run using the L-INS-I alignment option. Multiple-sequence alignments were used directly to generate a pHMM using HMMER version 3.3 [134]. Models were built using the hmmbuild function and pressed into binary form using the hmmpress function.

Validation of precision mode models

Precision mode models were validated against the full data sets from which seed sequences were chosen, excluding the sequences which were included in the pHMMs themselves. For each model, the pHMM was run against the full data set for the relevant RiPP class using the hmmscan function of HMMER3.3 [134]. Hmmscan was run with a bit score cutoff of 25 and with all other options set to default. A given model was deemed functional if >95% of RRE-containing protein sequences in a data set were retrieved by the pHMM at this bit score threshold. In cases where this criterion

was not met, sequences not retrieved by the model were used to enrich the original seed sequence alignment and an improved model was generated. In cases where an extensive bioinformatic survey was not available for a certain RiPP class, model accuracy was assessed in two ways: First, the set of homologous proteins generated by PSI-BLAST during model generation was tested against the pHMM using hmmscan with a bit score cutoff of 25. Second, an hmmsearch was performed using the HMMER3.3 Web tool (https://www.ebi.ac.uk/Tools/hmmer/search) against the UniProtKB database. The biosynthetic gene clusters surrounding gene hits were visualized using the RODEO Web tool [45] (http://rodeo.scs.illinois.edu). A model was considered valid if >95% of the proteins retrieved by PSI-BLAST were detected by the model and >90% of proteins retrieved from the UniProtKB database co-occurred with genes belonging to Pfams known to associate with that RiPP class. Finally, all models were tested for false-positive rates. All models were run against a data set of 3,000 protein sequences selected from across the data sets used for generating all precision mode models using hmmscan at a bit score cutoff of 35. Models were considered to have acceptably low false-positive rates if <100 hits for any given model belonged to a divergent RiPP class.

As described above, precision mode models were also validated against a set of ~5,000 proteins from the antiSMASH database. These protein sequences were employed in the generation of exploratory mode and thus were a form of cross-validation between the two modes of RRE-Finder. This data set consists of RRE-containing proteins primarily from the thiopeptide, lasso peptide, lanthipeptide, sactipeptide, and LAP classes. Not all proteins contained within the data set canonically contain RRE domains, particularly those belonging to class II to IV lanthipeptides. All precision-mode models were assessed by hmmscan searches against this data set with bit score cutoffs of 15, 25, and 35 (representing tolerant, moderate, and stringent bit score thresholds).

Generation of exploratory mode

Exploratory mode was generated for the purpose of identifying RRE sequences with higher divergence from RREs in known RiPP classes in a more unbiased manner than precision mode. For exploratory mode, we constructed a truncated version of the HHpred pipeline [111]. In this pipeline, a query sequence is first expanded with HHblits into a multiple sequence alignment (MSA) using a database of interest, in this case the uniclust30 database [136]. The secondary structure of the MSA is predicted using the adds.pl script available in the PSIPRED function of the HHsuite tool [112]. The MSA is then searched with HHsearch against a second database, which consists of three sequences from the Protein Databank (PDB) corresponding to RRE crystal structures (PDB entries 5V1T, 5SXY, and 3G2B). To closely mimic the HHpred pipeline, we used the uniclust30 database for MSA generation (version from August 2018 [https://uniclust.mmseqs.com]). This database contains all sequences from the UniProt database clustered with MMseqs2 [138] at a cutoff of 30% pairwise sequence identity.

For the initial generation of an RRE database, we used the above-mentioned pipeline to search 5,000 RiPP BGCs from the antiSMASH database against the uniclust30 database. Regions showing distant similarity to the reference RRE domains (probability, ≥40%; length, ≥50 residues) were extracted with 15 flanking residues on each side, and the extracted regions were resubmitted to the same pipeline with a higher cutoff to confirm the results (probability, ≥90%; length, ≥50 residues). Additional RRE sequences were added for the LAP, streptide, and proteusin RiPP families, for which no entries were available in the antiSMASH database.

The resulting database of RREs was used to generate a custom HHpred database as described in the documentation of the HHsuite tool, including the addition of secondary structure predictions with PSIPRED. In parallel, all RREs found were clustered with MMSeqs2 using default settings (pairwise identity, ≥80%) and the sequences in each cluster of RREs were aligned using MUSCLE [153]. The resulting alignment was converted into .a3m format using the reformat.pl script available in the HHsuite tool. Each alignment was then further enriched with more homologous sequences from the UniProtKB database by using HHblits with the uniclust30 database with three iterations. Finally, the expanded alignments were converted into pHMMs using HMMER3.3.

In exploratory mode, each query is first subjected to hmmsearch using the pHMMs described above. Queries passing the initial cutoff (see main text) and with minimum alignment length of 50 residues have the relevant regions extracted, including 15 flanking residues on each side. The candidate RRE region is then subjected to the HHpred pipeline described above. In the first step of MSA generation, however, the custom database containing RRE regions is used instead of the uniclust30 database. RRE regions showing homology to the reference RRE domains (length, ≥50 residues; probability, ≥90%) are considered hits.

Reducing false positives

To remove sequences containing transcriptional regulators (a large source of false positives using exploratory mode), we constructed a list of Pfam pHMMs containing a variety of DNA-binding regulators and other helix-turn-helix domains that share structural homology to the RRE domain. Each resulting hit is searched against this database with hmmsearch using the trusted cutoffs of each pHMM. Overlap of a regulator with a retrieved RRE is indicated in the output file. Information on which Pfams were filtered out is available in Data Set S4 (https://figshare.com/articles/Dataset_S4_Pfam_filtering/12568136).

Analysis of the MIBiG database

The pipeline described above was used to analyze all proteins from the MIBiG database (version 1.4), using bit score cutoffs ranging from 15 to 50. The resulting hits were separated into those belonging to RiPP and non-RiPP BGCs. Hits from the RiPP BGCs were additionally clustered per RiPP class. RiPP BGCs containing only precursors were removed.

Analysis of the UniProtKB database

The pipeline described above was used to analyze all proteins from the UniProtKB/TrEMBL database (UniProt release 2019_09). A bit score cutoff of 25 was used for precision mode and the initial filter of exploratory mode. For exploratory mode, proteins identified as likely regulators were removed after the initial hmmsearch step in the exploratory pipeline.

For the discovery of new classes, UniProtKB hits found by both modes of RRE-Finder, in particular using the auxiliary models of precision mode, were annotated with Pfam models (version 32.0) [130]. Several hits containing a Pfam domain that indicated an enzymatic activity were selected, and their genomic neighborhoods were investigated, as well as their overlap with antiSMASH gene clusters. In addition, the presence of RRE domains in these hits was confirmed by submitting to the HHpred Web tool (https://toolkit.tuebingen.mpg.de/tools/hhpred).

For analysis of the UniProtKB database using precision mode, the HMMER3.3 Web tools were used. Each model was individually run through hmmsearch of the UniProtKB database with a bit score cutoff of 25. Retrieved proteins for each model were compiled, and duplicate protein

accessions were removed to determine the exact number of unique proteins detected by each precision model. Information on duplicate hits from two or more precision models were used to determine model overlap and RRE relatedness, as shown in Figure S7.

Generation of sequence similarity networks and a diversity-maximized phylogenetic tree

The unique protein accessions from hmmsearch of the UniProtKB database using precision mode were directly used to generate an SSN using EFI-EST [132] (https://efi.igb.illinois.edu/efi-est/) and visualized with Cytoscape [133]. All sequences were excised to consist of only the RRE domain using a custom script. This script employs hmmsearch to identify the residues of a protein corresponding to the query pHMM and includes only those residues in the FASTA output. All SSNs shown are either a RepNode60 or RepNode80 network, meaning that protein sequences sharing more than 60% or 80% sequence identity are conflated into one node on the network. In general, alignment scores for network visualization were chosen to reflect a cutoff where sequences with >40% sequence identity cluster together. For the networks shown in this work, these alignment scores were 22 and 25 (representative of E-value cutoffs of 10^{-22} and 10^{-25}, respectively).

A diversity-maximized, maximum-likelihood phylogenetic tree was generated by first selecting a smaller subset of the sequences represented on the SSN. All sequences represented by clusters consisting of 1 to 3 nodes were included in the tree. For larger clusters, a random sampling of 10% of the sequences in the cluster was used for tree generation. All sequences were excised to contain only the RRE using the methods described above. The subset of sequences was used to generate a multiple-sequence alignment using MAFFT 7.450 [152]. MAFFT alignments were run using the L-INS-I alignment option. The MSA was transformed into an approximate-maximum-likelihood tree using FastTree 2.1 [154] with the default Jones-Taylor-Thornton (JTT) model. The tree was visualized using the Interactive Tree of Life (iTOL) website (http://itol.embl.de/).

Integration of RRE-Finder into RODEO and antiSMASH

Precision mode models have also been incorporated into both the GitHub and Web tool versions of RODEO 2 (http://rodeo.scs.illinois.edu). Included is an option to score RRE domains, which, if selected, will show which precision-mode models are matched, along with the default Pfam matches. The integration of precision mode is in progress for version 6.0 of antiSMASH, which is currently in the development phase and will be reported elsewhere. In addition, the standalone RRE-Finder tool is available on GitHub (https://github.com/Alexamk/RREFinder) and is capable of detecting RREs in precision mode and exploratory mode directly from antiSMASH and DeepBGC output [148].

Supplementary information for Chapter 2

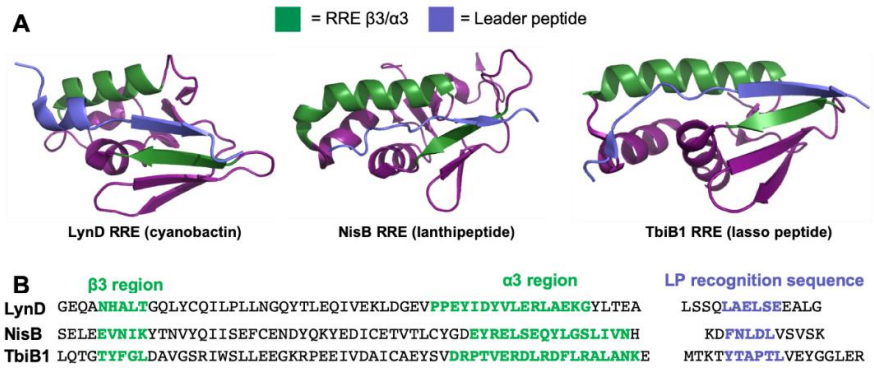


Figure S1. Structural homology of the RRE domain. (A) The crystal structures of three RRE domains (excised for LynD and NisB) are shown from three RiPP classes. The leader peptide is highlighted in blue, while the conserved cleft in the RRE that binds the leader peptide (LP) is highlighted in green. (B) The sequences of each of the three RRE domains shown in A.

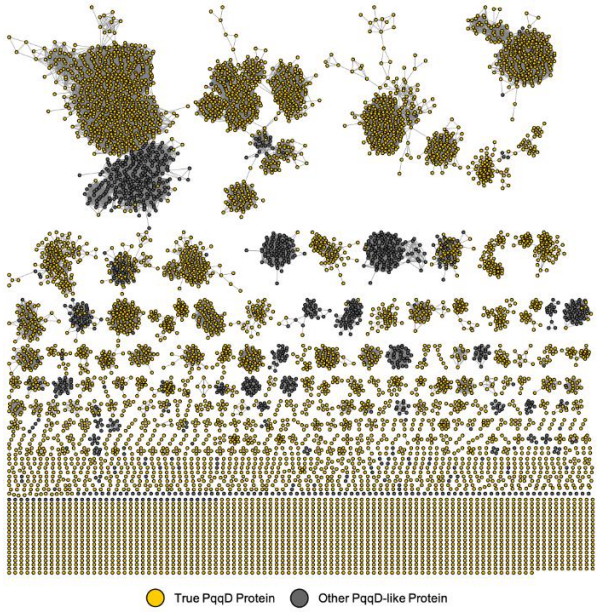


Figure S2. Sequence diversity of the RRE domain. Sequences belonging to PF05402 (PqqD) are represented in the SSN. The network was generated at an alignment score of 25 (E-value = 10^{-25}) and is presented as a RepNode80 (protein sequences with greater than 80% identity are conflated to a single node). Nodes are colored gold if the gene co-occurs within two open-reading frames of a radical SAM enzyme (i.e. a PqqE homolog), indicating that the protein may be a true PQQ biosynthesis protein.

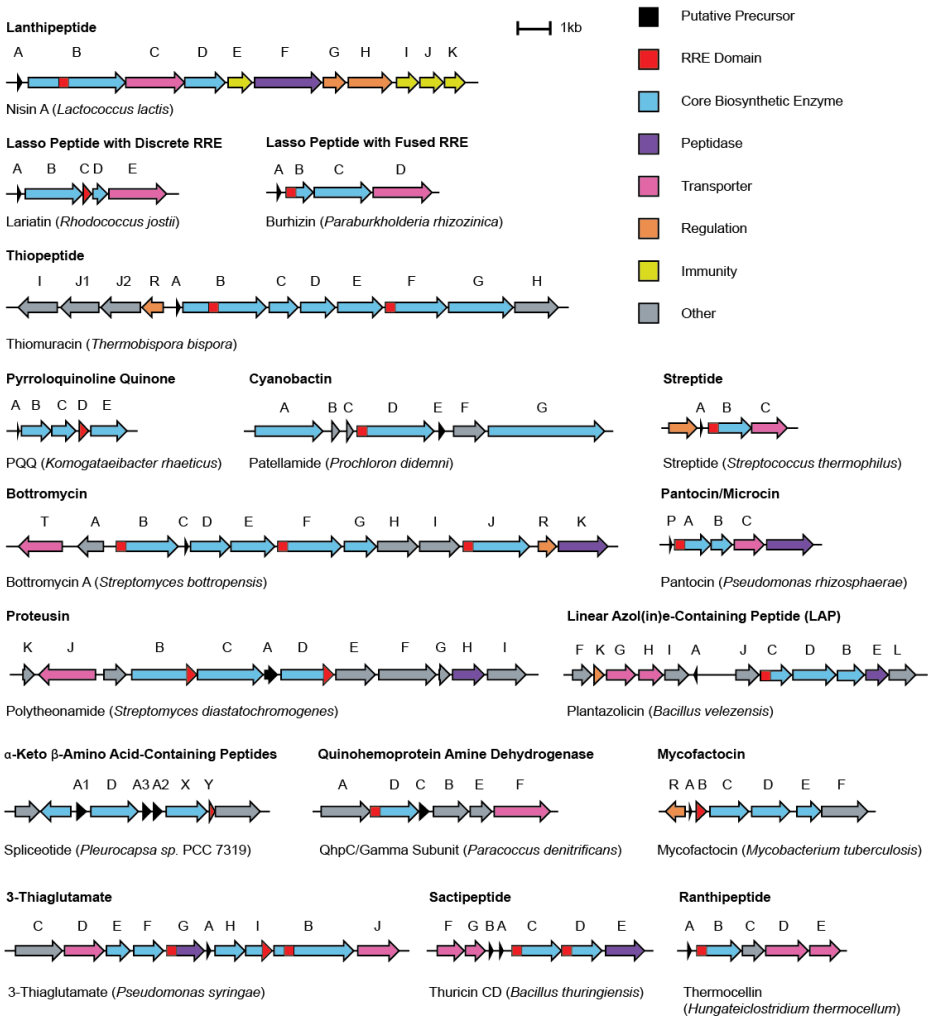


Figure S3. Representative RiPP gene clusters for precision mode models. One representative example is given for each RiPP class represented by one or more precision mode models. See Table S3 for a list of BGCs targeted. The 35 pHMMs comprising precision mode are provided in Data Set S2 (https://figshare.com/articles/Dataset_S2_HMM_files/12030651). The relevant class is shown in bold above the BGC, while the specific product encoded by the cluster is shown below the cluster. RRE domains are highlighted in red. In cases where RRE domains are fused to other domains, the red portion of the open reading frame represents the location of the RRE within the protein. QHNDH, quinohemoprotein amine dehydrogenase; DUF, domain of unknown function; rSAM, radical S-adenosylmethionine.

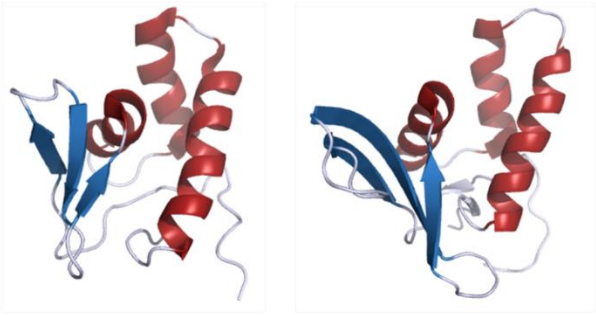


Figure S4. Structural homology of the RRE to DNA-binding proteins. The RRE consists of a conserved secondary structure of three α-helices and three β-strands, highlighted in blue and red in the structures shown. This secondary structure is also present in many regulatory and DNA-binding elements, such as the truncated DNA-binding portion of the Neisseria protein shown. HHpred analysis also shows high structural similarity (>90% probability) between several DNA-binding elements and RRE-containing proteins. Sequence similarity between transcription regulators and RRE domains still remains low, with the two sequences shown sharing only 33% amino acid sequence identity. Thus, it is plausible that RRE domains evolved from transcriptional regulatory proteins.

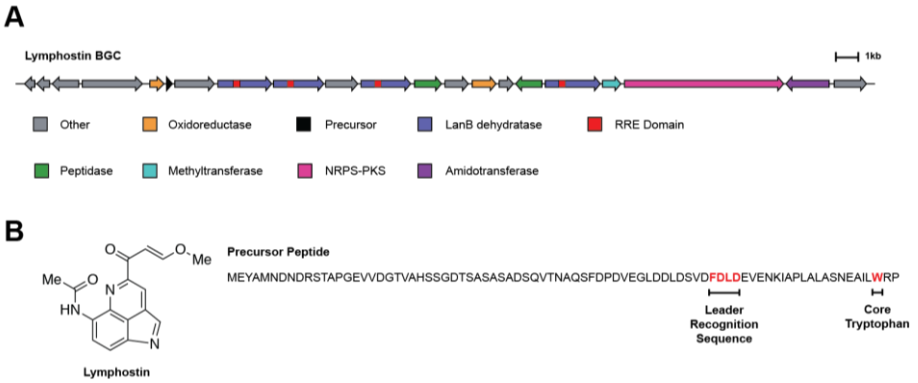


Figure S5. An RRE is detected in a type II PKS BGC by RRE-Finder. (A) RRE-containing proteins found in type II PKS clusters. The lymphostin BGC, a member of the pyrroloquinoline alkaloid class of RiPPs [155]. Many pyrroloquinoline alkaloid (PQA) clusters contain both a PKS-NRPS module and one or more LanB-type enzymes containing internal RRE domains. (B) Structure of lymphostin, a RiPP derived from tryptophan.

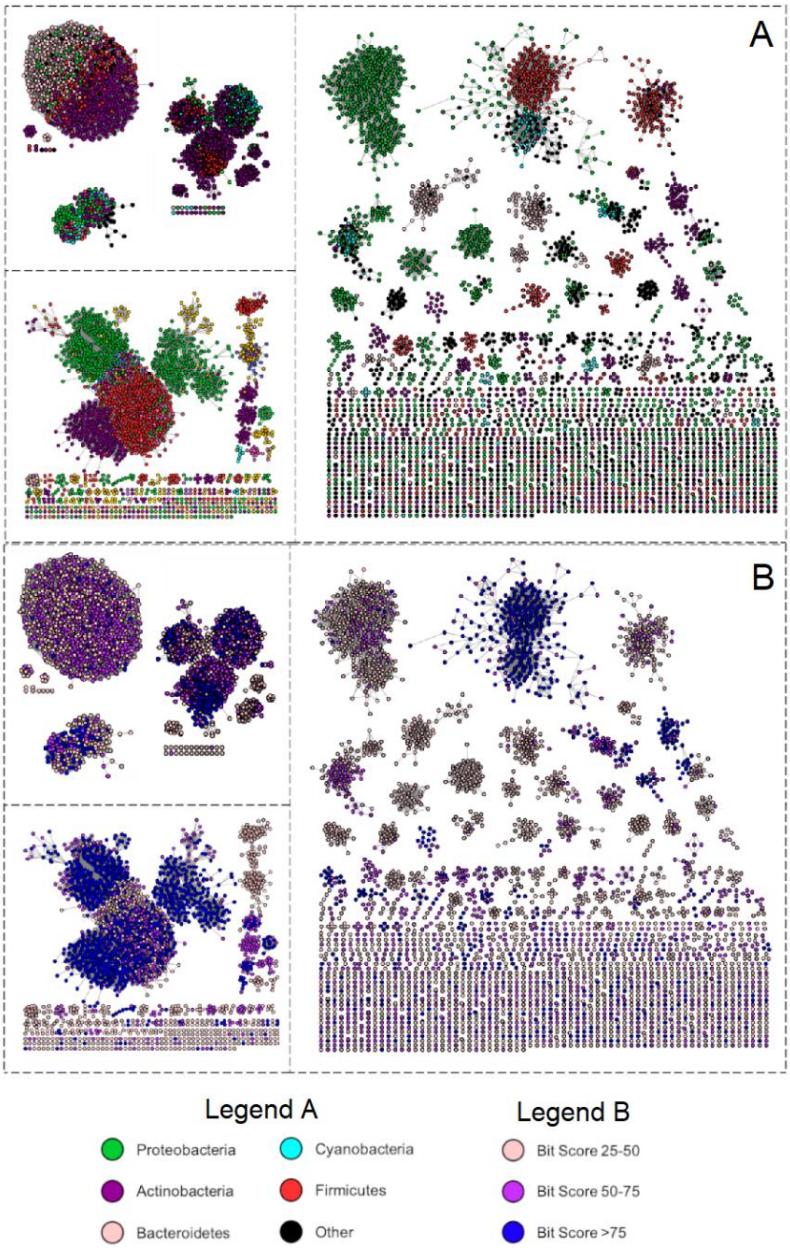


Figure S6. Sequence similarity networks of UniProt hits retrieved by precision mode. (A) Sequence similarity network of retrieved UniProt proteins annotated by taxonomic origin. The SSN is identical to Figure 5 but has been recolored by taxonomy of the producing organism (alignment score of 22 [RepNode60]). (B) The sequence similarity network in panel A was recolored according to the bit score significance of the match to a precision model (alignment score of 22 [RepNode60])**.**

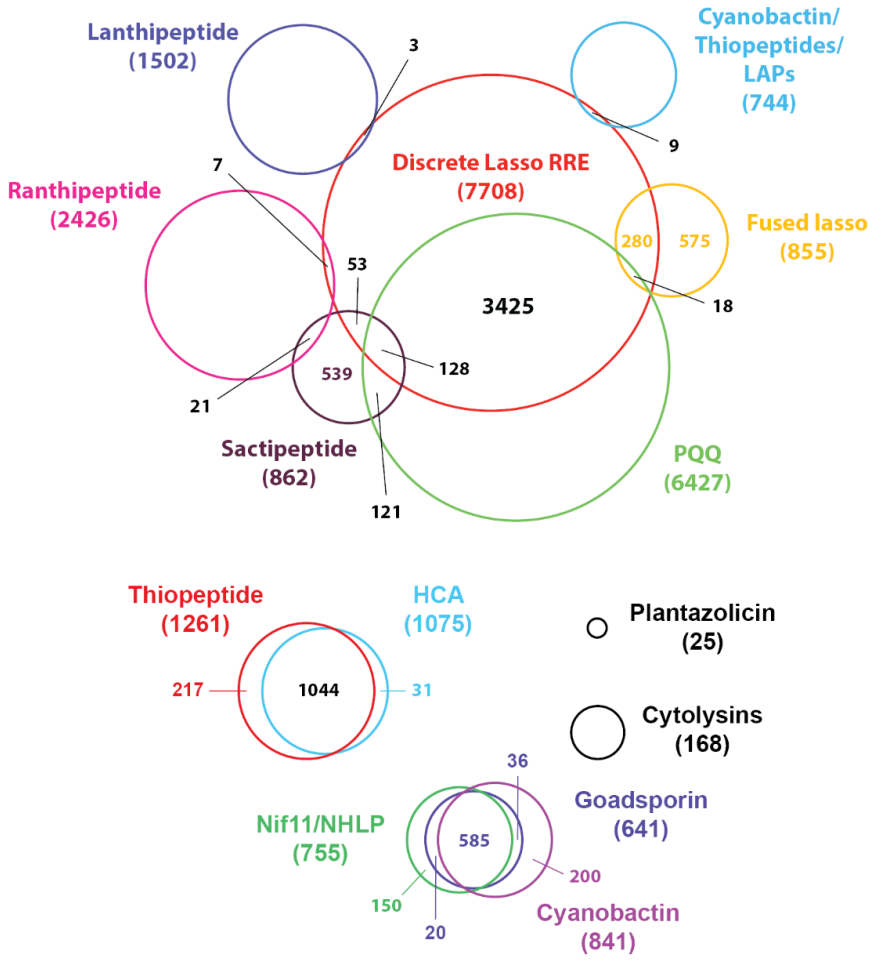


Figure S7. Overlap of retrieved UniProt proteins in the most populous RiPP classes (Top) and YcaO/RRE-dependent RiPP classes (bottom). Individual precision models for each of the shown RiPP classes were employed for hmmsearch of the UniProtKB database at a bit score cutoff of 25. The total number of retrieved sequences for each model is in parentheses. The numbers within circles indicates model redundancy or overlap, owing to the same sequence being retrieved by more than one precision model at a bit score of 25. The discrete lasso peptide RRE model retrieves more proteins than anticipated. For example, many lasso peptide RREs co-occur in clusters with radical *S*-adenosylmethionine enzymes. In addition, there is significant overlap between the RREs of lasso peptides and those from PQQ clusters. For the Ycao/RRE-dependent RiPP classes, model overlap reveals that some numbers of retrieved proteins for precision mode are artificially high. For example, there are only ~500 proteins retrieved by the thiopeptide model that co-occur with canonical thiopeptide modifying enzymes, such as the [4 + 2] cycloaddition enzyme. The other proteins retrieved by this model are heterocycloanthracins, which employ a highly similar leader peptide recognition sequence and RRE domain primary sequence. NHLP, nitrile hydratase-like leader peptide [156]; HCA, heterocycloanthracin.

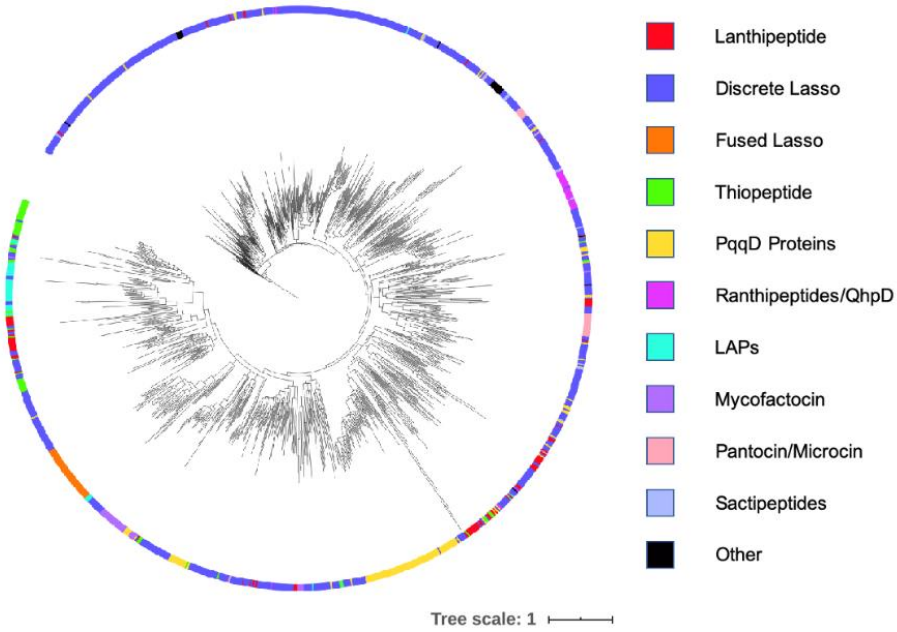


Figure S8. Representative phylogenetic tree for retrieved UniProt proteins. Shown are RRE sequences retrieved by a hmmsearch of the UniProtKB database using precision mode. Arc segments are colored based on the precision mode model matched with the highest bit score. RRE domains from discrete lasso peptide BGCs share the most sequence similarity to non-RiPP regulatory proteins and thus branch most directly from the transcriptional regulator outgroup (appearing at ~10 o'clock on the tree; PDB entry 3DEE).

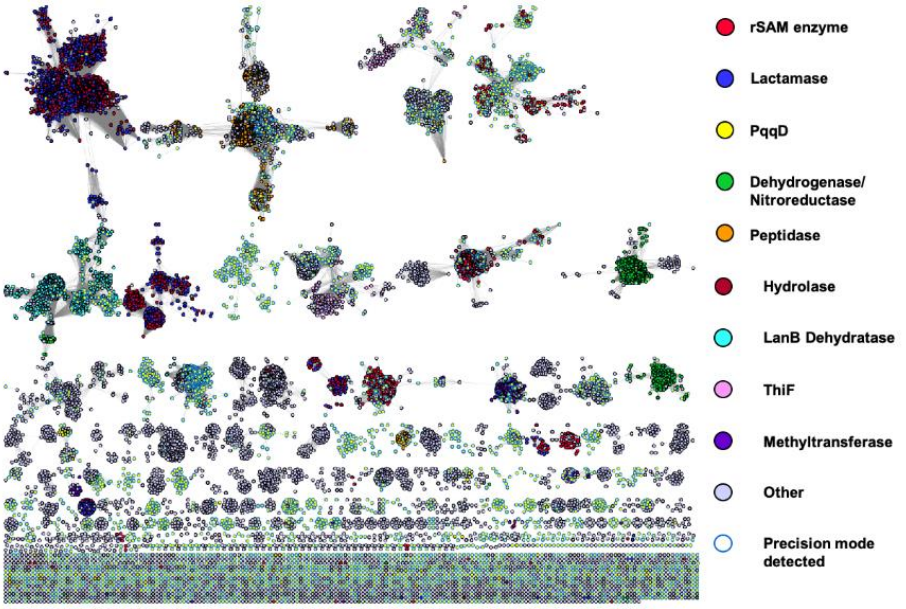


Figure S9. Sequence similarity network of retrieved UniProt proteins by exploratory mode. The UniProtKB database was searched using exploratory mode at a bit score cutoff of 25 (alignment score of 30 [RepNode60]). All proteins retrieved by exploratory mode, not inclusive of proteins retrieved by precision mode at the same bit score cutoff, are visualized on the sequence similarity network. Nodes are colored based on UniProt annotations that were highly represented in the network. Proteins that were retrieved by precision mode at bit score cutoffs under 25 have blue outlines. The network was generated using EFI-EST [132] and visualized with Cytoscape [133].

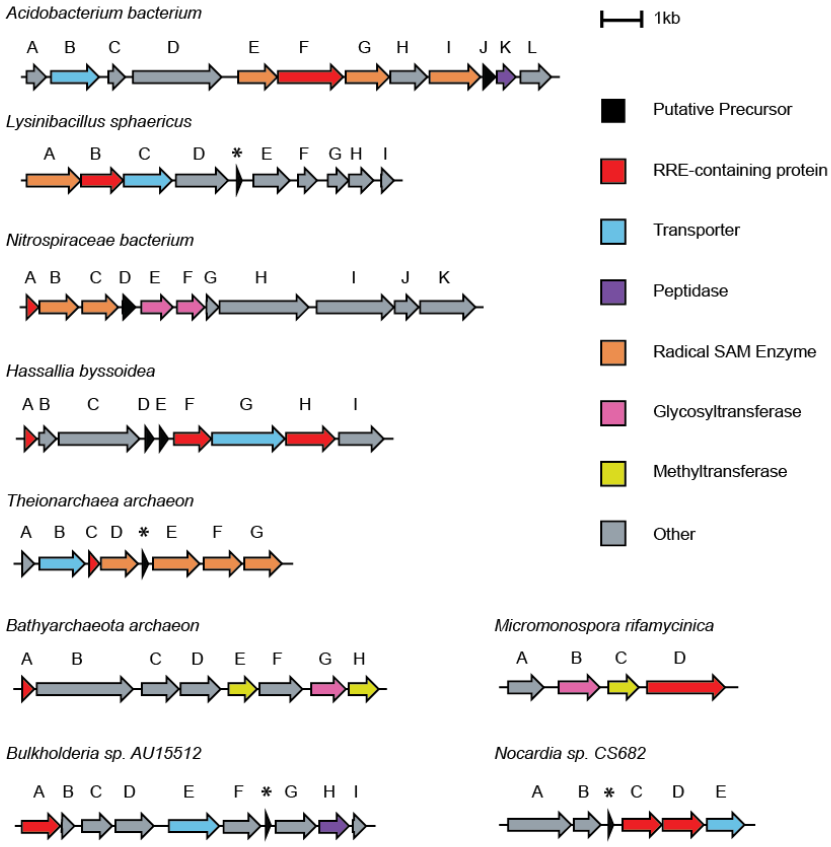


Figure S10. Example RiPP BGCs found by RRE-Finder. Shown are nine BGCs that contain RRE domains in novel contexts. Proteins highlighted in red indicate proteins containing RRE domains as predicted by RRE-Finder. All RRE domain-containing proteins are listed in the accompanying table along with protein accessions. Some of the BGCs shown were mined using exploratory mode of RRE-Finder, while the others were mined using the auxiliary models of precision mode. In cases where a likely precursor peptide was predicted by RODEO but no NCBI accession was found, the precursor is marked with an asterisk.

Table S1. Representative list of prokaryotic RiPP classes containing RREs. RRE domains are present in over 50% of RiPP classes produced by prokaryotes. These classes are listed along with information pertaining to the type of RRE fusion, and an example product. Classes are listed as RRE dependent if at least one protein in the BGC is predicted to contain an RRE by RRE-Finder exploratory mode. Not all of these classes have been confirmed to be RRE dependent by experimental studies. Although there are no general trends as to which classes are RRE dependent, some enzymes—such as rSAM enzymes and cyclodehydratases—commonly co-occur with discrete or fused RRE domains.

Class Name	Example Product	RRE Type	Citation DOI
Lanthipeptides	Nisin A	Fused to LanB protein (class I lanthipeptides only)	10.1038/333276a0
Pyrroloquinoline quinones (PQQ)	PQQ cofactor	Discrete RRE	10.1128/jb.171.1.447-455.1989
Lasso peptides	Microcin J25	Fused to leader peptidase or discrete	10.1128/jb.178.12.3661-3663.1996
LAPs	Microcin B17	Fused to E1-like homolog (and sometimes YcaO as well)	10.1126/science.274.5290.1188
Sactipeptides	Subtilosin	Fused to rSAM sactionine enzyme	10.1128/JB.182.11.3266-3273.2000
Pantocins/ Microcins	Pantocin A	Fused to E1-like homolog	10.1002/anie.200351054
Cyanobactins	Patellamide A	Fused to YcaO domain (azoline-containing cyanobactins only)	10.1038/nchembio.84
Thiopeptides	Thiostrepton	Fused to the F-component of the cyclodehydratase	10.1073/pnas.0900008106
Mycofactocins	Mycofactocin	Discrete RRE	10.1186/1471-2164-12-21
Bottromycins	Bottromycin A1	Fused to rSAM methyltransferase	10.1039/C2SC21190D
Proteusins	Polytheonamide	Fused to rSAM epimerase and rSAM methyltransferase	10.1126/science.1226121
Streptides	Streptide	Fused to rSAM enzyme	10.1038/nchem.2237
Ranthipeptides	Freyrasin	Fused to rSAM enzyme	10.1021/jacs.9b01519
α-Keto β-amino acid-containing peptides	PlpA	Discrete RRE	10.1126/science.aao0157
Rotapeptides	TQQ	Fused to rSAM enzyme	10.1021/jacs.9b05151
Ryptides	RRR	Fused to rSAM enzyme	10.1021/jacs.9b09210

Table S2. Representative RRE domains that have been structurally characterized. LAP, linear azol(in)e-containing peptides; PQQ, pyrroloquinoline quinone.

Protein	RiPP Class	PDB Accession	UniProtKB Accession	Citation DOI
LynD	Cyanobactin	4V1T	A0YXD2	10.1038/nchembio.1841
TruD	Cyanobactin	4BS9	B2KYG8	10.1002/anie.201306302
NisB	Lanthipeptide	5WD9	P20103	10.1038/nature13888
McbB	LAP	6GOS	P23184	10.1016/j.molcel.2018.11.032
TfuB1	Lasso peptide	6JX3	Q47AT5	10.1021/acschembio.9b00348
TbiB1	Lasso peptide	5V1V	D1CIZ5	10.1073/pnas.1908364116
MccB	Microcin	6OM4	Q47506	10.1039/c8sc03173h
PaaA	Pantocin	5FF5	Q9ZAR3	10.1021/jacs.5b13529
PqqD	PQQ	3G2B/5SXY	Q8P6M8	10.1002/prot.22461 10.1021/acs.biochem.7b00247
CteB	Ranthipeptide	5WGG	A3DDW1	10.1021/jacs.7b01283
SkfB	Sactipeptide	6EFN	O31423	10.1074/jbc.RA118.005369
SuiB	Streptide	5V1T	A0A0Z8EWX1	10.1073/pnas.1703663114
TbtB	Thiopeptide	6EC7	D6Y502	10.1073/pnas.1905240116

Table S3. Description of RRE-containing proteins targeted by precision mode. BGCs are illustrated in Figure S3. In cases where one BGC contains more than one protein with an RRE, separate NCBI protein accession identifiers are given for each.

Natural Product	Protein	RRE Type	NCBI Accession
Nisin A	NisB	Fused to lanthipeptide dehydratase	ADJ56353.1
Lariatin	LarC	Discrete	BAL72548.1
Burhizin	BurB	Fused to lasso peptidase	CBW74825.1
Thiomuracin	TbtB	Fused to lanthipeptide dehydratase	ADG87277.1
Thiomuracin	TbtF	Fused to ocin-ThiF protein	ADG87281.1
PQQ	PqqD	Discrete	WP_034930240.1
Patellamide	PatD	Fused to cyclodehydratase	AAY21153.1
Bottromycin	BmbB	Fused to methyltransferase	CCM09442.1
Bottromycin	BmbF	Fused to methyltransferase	CCM09446.1
Bottromycin	BmbJ	Fused to methyltransferase	CCM09450.1
Polytheonamide	PoyB	Fused to methyltransferase	AFS60637.1
Polytheonamide	PoyD	Fused to epimerase	AFS60640.1
Plantazolicin	PznC	Fused to cyclodehydratase	CBJ61638.1
Thuricin CD	TrnC	Fused to rSAM enzyme	AED99784.1
Thuricin CD	TrnD	Fused to rSAM enzyme	AED99785.1
Streptide	SuiB	Fused to rSAM enzyme	ABJ66529.1
Spliceotide	PlpY	Discrete	WP_019503879.1
Pantocin	PaaA	Fused to ThiF protein	WP_043190265.1
Thermocellin	CteB	Fused to rSAM enzyme	WP_003517268.1
Mycofactocin	MftB	Discrete	WP_019735253.1
QHNDH	QhpD	Fused to rSAM enzyme	SDJ52620.1
3-Thiaglutamate	PmaB	Fused to short LanB enzyme	KPW26932.1
3-Thiaglutamate	PmaG	Fused to protease	KPW26903.1
3-Thiaglutamate	PmaI	Fused to DUF	KPW26921.1

Table S4. RRE-Finder computing times. RRE-Finder analysis times compared to HHpred. Both precision and exploratory modes of RRE-Finder significantly decrease analysis times compared to HHpred, the gold standard for detecting RREs. Exploratory mode has longer analysis times than precision mode, due to the detection of distant protein homology. However, exploratory mode is still roughly 3,000 times faster than HHpred analysis. Analysis was carried out on an Intel Xeon E5-4640 at 2.4 GHz, using 4 threads.

Method	Dataset	Entries	Time Required (h)
RRE-Finder (precision)	MIBiG (all)	31,025	0.002
RRE-Finder (exploratory)	MIBiG (all)	31,025	0.2
HHPred	MIBiG (RiPP only)	2,513	54

Table S5. Model validation of precision mode for select RiPP classes. Four populous classes of RiPPs were selected for thorough model validation, using the most recent published data sets of predicted BGCs for sactipeptides, ranthipeptides, lanthipeptides, and thiopeptides [55, 72, 73]. In all cases, the proteins from each data set known to contain RRE domains were queried against the relevant precision model using hmmscan at tolerant (15), moderate (25), and stringent (35) bit score cutoffs. To determine the false-positive rate of the lanthipeptide model, all LanB-type enzymes in the data set belonging to type II to IV lanthipeptide biosynthetic pathways were queried. To determine the false-positive rates of the sactipeptide, thiopeptide, and ranthipeptide models, a neighboring protein to each RRE domain was queried. The neighboring proteins queried were ABC transporters (for sactipeptides/ranthipeptides) and cyclodehydratase enzymes (for thiopeptides).

Dataset	Bit Score			
	15	25	35	Total in Dataset
Lanthipeptide, I *(True Positive)*	1950	1910	1640	2020
Lanthipeptides, II-IV *(False Positive)*	90	20	3	4453
Sactipeptide *(True Positive)*	799	769	690	865
Sactipeptide *(False Positive)*	1	1	0	865
Ranthipeptide *(True Positive)*	2241	2150	1960	2301
Ranthipeptide *(False Positive)*	10	7	4	2301
Thiopeptide F Protein *(True Positive)*	495	492	440	515
Thiopeptide F Protein *(False Positive)*	5	3	2	515

Table S6. Validation of RRE-Finder modes against the MIBiG database. RRE domains predicted by RRE-Finder and HHpred are grouped based on RiPP class. Precision and exploratory mode combined detect almost all of the RRE-containing proteins detected by HHpred (rightmost column). Precision mode readily detects RRE domains in known RiPP classes. Exploratory mode also detects these RREs but additionally retrieves putative RRE domains in thioviridamide-like and pheganomycin BGCs. Some of these RREs were also predicted by HHpred; thus, exploratory mode gives results in these cases similar to those obtained with HHpred. However, exploratory mode only sparingly detects RREs in the LAP and streptide RiPP classes.

RiPP Class	Nr. of BGCs	Protein annotation	Example MIBiG BGC	Example Protein	Total	Precision	Exploratory	HHPred
Lasso peptide	35	Leader peptidase	BGC0000581	McjB	12	8	10	7
	35	PqqD-like	BGC0000575	LarC	23	23	23	23
Lanthipeptide	31	LanC-like	BGC001392	NisC	1	0	1	1
	31	LanB dehydratase	BGC0000535	NisB	30	29	30	27
Thiopeptide	24	Dehydratase	BGC0000613	TpdB	17	0	16	6
	24	Cyclodehydratase	BGC0000613	TpdF	2	2	2	2
	24	Radical SAM	BGC0001753	TbtI	1	0	1	1
	24	ocin_ThiF-like*	BGC0000603	CltD	23	18	17	17
	24	Dehydrogenase	BGC0000613	TpdE	5	0	4	3
Cyanobactin	13	Cyclodehydratase	BGC0000475	PatD	8	8	8	8
	13	Dehydrogenase	BGC0000475	PatG	8	0	8	8
LAP	10	Cyclodehydratase	BGC0000569	PtnD	7	7	1	3
	10	Dehydrogenase	BGC0000565	GodE	2	1	2	2
	10	Hypothetical protein	BGC0000567	TfxC	1	1	1	1
Thioamitide	4	Methyltransferase	BGC0000625	TvaG	4	0	4	1
Sactipeptide	4	Radical SAM	BGC0000600	ThnB	5	4	3	4
Bottromycin	4	Radical SAM	BGC0000468	BmbB	12	12	12	0**
Pheganomycin	1	Radical SAM	BGC0001148	Pgm3	1	0	1	1
Proteusin	1	Radical SAM	BGC0000598	PoyB	1	0	1	1
	1	Radical SAM	BGC0000598	PoyC	1	1	1	1
	1	Radical SAM	BGC0000598	PoyD	1	1	1	1
Plp	1	Radical SAM	BGC0001745	PlpY	1	1	0	1
Streptide	1	Radical SAM	BGC0001209	SuiB	1	1	0	0
Microcin	1	ThiF-like	BGC0000585	MccB	1	1	1	1
3-thiaglutamate	1	LanB dehydratase	BGC0001486	PmaJ	1	1	1	0
	1	DUF	BGC0001486	PmaI	1	1	1	1
	1	Peptidase	BGC0001486	PmaG	1	1	1	1

* 15 of these proteins show weak similarity to the ocin_ThiF_like domain (TIGR03693).

** RREs in radical SAMs encoded by bottromycin BGCs are typically detected by HHpred at a slightly lower probability than was used as the cutoff (~70 to 90%).

Table S7. Exploratory mode false-positives in non-RiPP BGCs. Exploratory mode retrieved a total of 36 proteins in non-RiPP BGCs at a bit score cutoff of 25. Many retrieved proteins were transcriptional regulators or proteins with a helix-turn-helix (HTH) motif. Other false positives included several proteins with sequence homology to RRE-containing proteins in RiPP BGCs. Some BGCs in MIBiG have poorly defined boundaries and thus may contain genes from nearby BGCs. Thus, some false positives shown may be true RRE domains in adjacent RiPP clusters (e.g., MIBiG BGC0000696, contains a neighboring LanB dehydratase and a LanC cyclase).

False-Positive Type	Number of Proteins Retrieved
Transcription Regulators/HTH Domains	8
Associated with Known RiPPs	17
Other	11

Table S8. Conservation of α3 and β3 regions of the RRE. Residue-level conservation was assessed using three metrics on eight precision mode models. The secondary structures principally responsible for binding the leader peptide (the α3 and β3 regions) were assessed separately from the remainder of the RRE domain. The region of the RRE with the greatest conservation per metric is indicated by red text. Individual RiPP classes were scored by selecting 10 divergent RREs from that class and excising the relevant substructure sequence. In some cases, pairs of RiPP classes that have significant mutual evolutionary relatedness were evaluated jointly; in these instances, a total of 20 sequences were used for the calculations (10 from each class). These data reveal a trend of higher conservation in the α3 and β3 regions of the RRE compared to other regions. Perhaps unsurprisingly, α3 displays the greatest conservation across RiPP classes, given that the contact with the leader peptide is primarily through side chain interactions as opposed to the β3 strand (primarily backbone interactions). HCA, heterocycloanthracin.

	Shannon Information Entropy			**ConSurf (0-9)**			**AACon (0-9)**		
	α3 helix	β3 strand	other	α3 helix	β3 strand	other	α3 helix	β3 strand	other
Goadsporin	**0.81**	0.65	0.45	**7**	6	4	**7**	6	4
Cyanobactin	**0.75**	0.59	0.39	**7**	6	3	**7**	6	4
Goadsporin/Cyanobactin	**0.62**	0.54	0.21	**6**	**6**	2	**6**	5	2
Discrete Lasso peptide	**0.43**	0.33	0.23	**4**	3	2	**4**	3	2
Fused Lasso peptide	**0.51**	0.32	0.31	**5**	3	3	**4**	3	2
Discrete/Fused Lasso peptide	**0.27**	0.22	0.13	**3**	**3**	1	**3**	2	1
Thiopeptide	**0.76**	0.72	0.56	**7**	**7**	6	**7**	**7**	5
HCA	**0.82**	0.74	0.58	**8**	7	6	**8**	7	5
Thiopeptide/HCA	**0.71**	0.64	0.49	**7**	6	5	**7**	6	5
Ranthipeptide	**0.68**	0.57	0.42	**7**	6	4	**7**	5	4
QhpD	**0.71**	0.59	0.47	**7**	6	5	**7**	6	5
Ranthipeptide/QhpD	**0.54**	0.43	0.36	**5**	4	4	**5**	4	3

Table S9. RRE-containing proteins in UniProtKB found by exploratory mode. Proteins retrieved by RRE-Finder were grouped based on Pfam/TIGRFAM domain identification. The overlap with precision mode's core models at a bit score threshold of 25 confirms that many known RRE fusions are detected by both modes, such as those containing YcaO and LanB dehydratase domains. Numbers of proteins retrieved by exploratory mode are inclusive of those retrieved by precision mode. Other novel RRE fusions are identified, such as fusions to metallo-β-lactamases, oxidoreductases, and glutathione *S*-transferases. RRE domains are also found in a number of unannotated small proteins, many of which are likely discrete RREs. Among the filtered proteins containing HTH domains (right column), the vast majority were annotated only as regulatory proteins. Notably, 1,869 short proteins (<120 residues) were filtered out during this step. Whether these proteins represent discrete RREs or simply small regulators could not be determined with the available data. Nevertheless, in most cases, no additional domain fusions were annotated among the filtered.

Protein domain categories	Number of hits exploratory (precision - core)	Enzymes overlapping with regulator domain
DNA-binding proteins and/or regulators (filtered)	22,357 (0)	NA
Other (length >= 120 aa)	16,595 (1,094)	20,267
Short proteins (length < 120 aa)	3,341 (952)	1,869
Metallo-β-lactamase	11,320 (1)	7
PqqD	10,994 (9,128)	18
rSAMs / Fe-S-binding domains	3,919 (2,491)	0
LanB dehydratase	3,313 (1,888)	2
Nitroreductase	1,039 (10)	4
YcaO cyclodehydratase	919 (837)	0
Methyltransferases	813 (11)	65
Translglutaminase	644 (552)	0
Ocin-thiF-like	589 (566)	0
Memo proteins	463 (5)	0
Oxidoreductase	104 (0)	0
Tryptophan halogenase	75 (0)	0
Cyclic nucleotide binding domain	67 (4)	19
Tetratricopeptide repeat	66 (4)	18
Peptidase	64 (2)	81
Glycosyltransferase	56 (20)	1
Asparagine synthase	19 (2)	0
Cupin domain	18 (0)	0
LanC cyclase	13 (0)	0
Glutathione-S-transferase	10 (0)	0
Carbamoyltransferase	8 (0)	0

Table S10. Description of RRE-containing proteins found by RRE-Finder. The letters used to identify a gene correspond to those used in the BGCs in Figure S10.

Organism	Gene	RRE Type	NCBI Accession	HHPRED detected
Acidobacterium bacterium	F	Fused to tetratricopeptide domain	OFW29522.1	x
Lysinibacillus sphaericus	B	Fused to glutathione *S*-transferase	WP_069508305.1	x
Nitrospiraceae bacterium	A	Discrete	RPI38387.1	x
Hassallia byssoidea	A	Discrete	KIF30015.1	x
Hassallia byssoidea	F	Fused to glycosyltransferase	KIF29242.1	-
Hassallia byssoidea	H	Fused to phosphoribosyl transferase	KIF29244.1	x
Theioarchaea archaeon	C	Discrete	KYK35486.1	x
Bathyarchaeota archaeon	A	Discrete	OGD46518.1	x
Micromonospora rifamycinica	D	Fused to carbamoyltransferase	WP_067301990.1	x
Bulkholderia sp. AU15512	A	Fused to iron redox enzyme	OXI24931.1	x
Nocardia sp. CS682	C	Fused to heme-oxygenase enzyme	QBS40287.1	x
Nocardia sp. CS682	D	Fused to iron redox enzyme	QBS40286.1	x

2

3

Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers novel candidate RiPP families

Alexander M. Kloosterman
Peter Cimermancic
Michalis Hadjithomas
Mohamed S. Donia
Michael A. Fischbach
Gilles P. van Wezel
Marnix H. Medema

The work described in this chapter is part of the publication:

Kloosterman, et al., *Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics.* PLOS Biol, 2020. **18**(12): e3002016.

Abstract

Microbial natural products constitute a wide variety of chemical compounds, many which can have antibiotic, antiviral or anticancer properties that make them interesting for clinical purposes. Natural product classes include polyketides (PKS), non-ribosomal peptides (NRPS) 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 subclasses. DecRiPPter combines a classifier based on Support Vector Machines (SVMs) 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. The BGCs of these families encode enzyme families not previously associated with RiPP biosynthesis, or precursors with interesting repeating patterns. These results highlight 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 https://decrippter.bioinformatics.nl. 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 [22, 25], means that we again face huge challenges to treat infections by multi-drug resistant bacteria [157]. 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 [23, 24]. 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 [32, 158]. 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 [27, 32, 159]. 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 [26, 30]. 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 [33]. 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 peptides (NRPs) and terpenes [39, 40, 62]. 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 [41] and EvoMining [160, 161], 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 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) [42]. 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 [43]. Despite this common biosynthetic logic, RiPP modifications are highly diverse. The latest comprehensive review categorizes RiPPs into roughly 20 different subclasses [42], such as lanthipeptides, lasso peptides and thiopeptides. Each of these subclasses 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 subclasses and modifications, new RiPP subclasses are still being found. These often carry unusual modifications, such as D-amino acids [98], addition of unnatural amino acids [162, 163], β-amino acids [103], or new variants of thioether crosslinks [55, 106]. 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 subclass of RiPPs comprises its own unique set of enzyme families to post-translationally modify the precursor peptides belonging to that subclass. Hence, while biosynthetic gene clusters (BGCs) for known RiPP subclasses can be identified using conventional genome mining algorithms, a much more elaborate strategy is required to automate the identification of novel RiPP subclasses.

Several methods have made progress in tackling this challenge. ‘Bait-based’ approaches such as RODEO [45, 55, 72-74, 86] and RiPPer [52] identify

RiPP BGCs by looking for homologues of RiPP modifying enzymes of interest, and facilitate identifying the genes encoding these enzymes in novel contexts to find many new RiPP BGCs. A study was also described using a transporter gene as a query that is less dependent on a specific RiPP subclass [164]. However, these methods still require a known query gene 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 [88]. 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 subclasses [89]. In the end, an algorithm for the discovery of BGCs encoding novel RiPP subclasses will need to integrate various sources of information to reliably identify genomic regions that are likely to encode RiPP precursors along with previously undiscovered modifying enzymes.

Here, we present decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER), an integrative algorithm for the discovery of novel subclasses of RiPPs, without requiring prior knowledge of their specific modifications or core enzymatic machinery. DecRiPPter employs a classifier based on Support Vector Machines (SVMs) 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 and Discussion

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 subclasses. Since no criteria could be used that are specific for individual RiPPsub classes, we used three criteria that generally apply to RiPP BGCs: 1) they contain one or more open reading frames (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 subclasses, we added a fourth criterion: 4) they have no direct similarity to BGCs of known classes (Figure 1).

For the first criterion, we trained several SVM classifiers to distinguish between RiPP precursors and other peptides. A collection of 175 known RiPP precursors, gathered from RiPP clusters from the MIBiG repository [29, 140] was used as a positive training set (Table S1). 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 modifying enzymes. Based on these, several SVMs were trained with different parameters and kernel functions, of which the average was taken as a final score (Materials and Methods). 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.

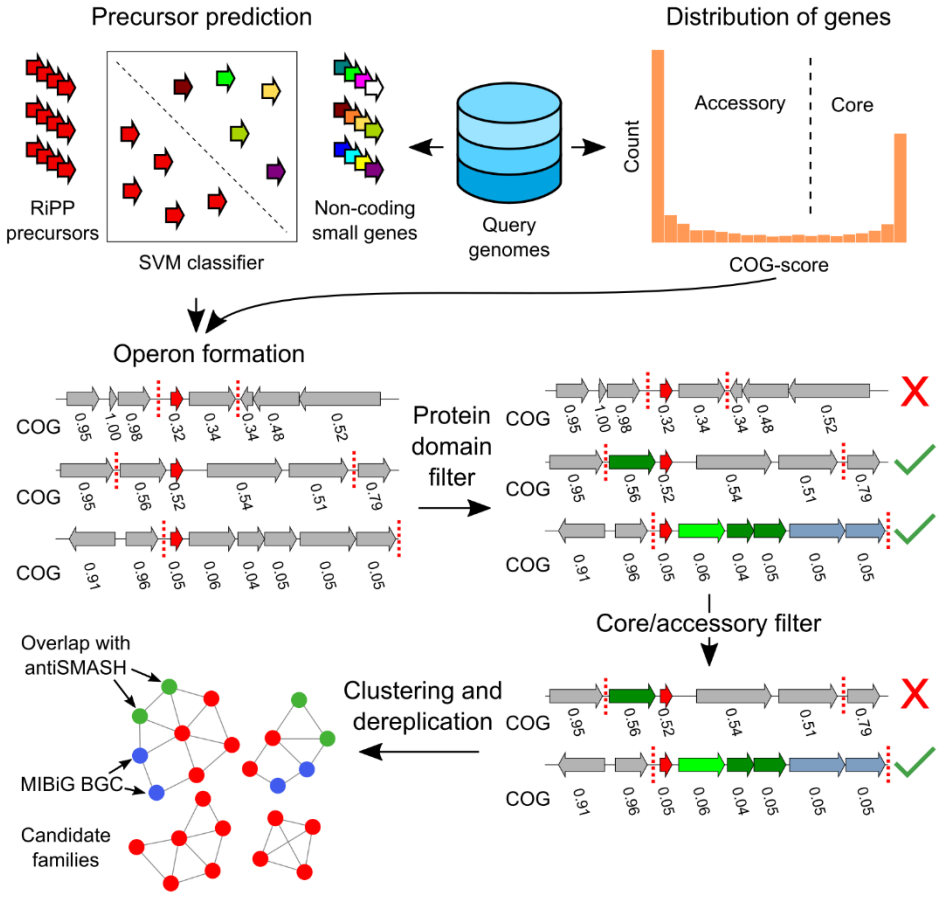


Figure 1. **decRiPPter pipeline for the detection of novel RiPP families.** The SVM classifier is used to identify all candidate RiPP precursors in a given group of genomes, using all predicted proteins smaller than 100 amino acids. The gene clusters formed around the precursors are analyzed for specific 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.

We termed this strategy leave-one-class-out cross-validation. Typically, the classifier was still capable of predicting the subclass that was left out. To validate the classifier, we used it to score precursor hits from the various RiPP mining studies performed using RODEO [45, 55, 72-74, 86]. In general, 66.7% of all precursors identified by RODEO's SVMs were scored as positive by decRiPPter's

classifier (Table S2). This shows that, for known RiPP subclasses, the classifier described here is well capable of detecting the majority of precursor peptides, although it is, unsurprisingly, outperformed by the dedicated, subclass-specific SVMs of RODEO.

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 (Cluster of Orthologous Genes, see description below, Materials and Methods, Figure 2 and Figure S1). These gene clusters were 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 [165]. 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 [41], and manually curated, resulting in a set of 4,131 protein domains. We also constructed Pfam [75] and TIGRFAM [76] 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 (Data S1, available from https://github.com/Alexamk/decRiPPter/tree/master/data/domains/).

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 [3-5]. 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 (Figure 1C and Figure 2).

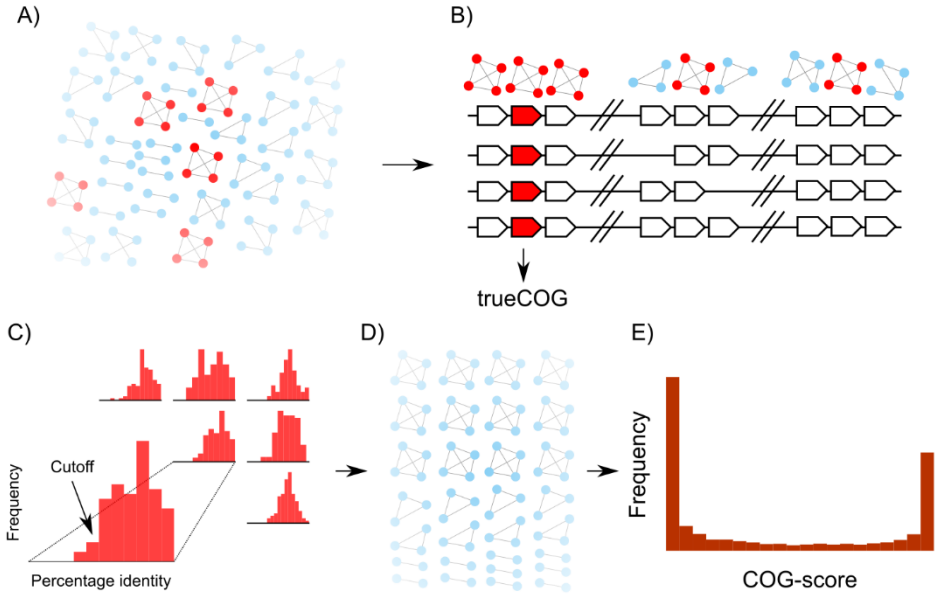


Figure 2. **decRiPPter determines the frequencies of occurrence of genes to calculate the COG score.** In this example, the COG scores of four genomes are calculated. A) All encoded proteins are aligned to find bidirectional best hits (BBHs; edges). All clusters of BBHs conserved across all genomes are displayed as red. If one genome does not contain a homologous gene, or the gene in question is not a BBH with all genes from the cluster from other genomes, it is not considered a conserved group of BBHs. B) If the flanking genes of the clusters of BBHs are also part of clusters of BBHs, the center genes are considered to form a true Cluster of Orthologous Genes (trueCOG). Of the three cases displayed here, only the leftmost group passes this criterion; for the center group, not all genes are conserved, and for the right group, not all genes are BBHs with one another in the flanking groups. C) The distribution of sequence similarities is used to calculate a sequence identity cutoff to use for each pair of genomes. D) All genes are paired using the sequence identity cutoffs determined in the previous step. E) The COG-score is calculated for each gene. Typically, a bimodal distribution can be seen, with many genes either conserved across all genomes, or only present in a single organism.

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 [166]). Therefore, decRiPPter first identifies the distribution of sequence identity values

of protein-coding genes that can confidently be assigned to be orthologs, and uses this distribution to find groups of genes across genomes with ortholog-like mutual similarity. First, a set of high-confidence orthologs, called true conserved orthologous genes (trueCOGs) are identified based on two criteria: 1) they should be bidirectional best hits (BBH) between all genome pairs, and 2) their two flanking genes should also be BBHs between all genome pairs [167]. In other words, decRiPPter looks for sets of three contiguous genes that are highly conserved in both sequence identity and synteny among all analyzed genomes, using DIAMOND [168]. The center genes of these gene triplets are themselves conserved, and have conserved surrounding genes, making it highly likely that they are orthologous to one another. These center genes were therefore considered trueCOGs. While this list of trueCOGs contains high-confidence orthologs, the criteria for orthology set here are strict, and many orthologs are missed by only considering orthologs based on BBHs [169]. We therefore further expanded the list of homologs with ortholog-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. Considering that only sequence identity is used as a cutoff here, these gene pairs are either orthologs or paralogs. The identified gene pairs are then clustered with the Markov Clustering Algorithm (MCL [170, 171]) 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 (Materials and Methods, Figure 2).

To validate our calculations, we analyzed the COG-scores of the highly conserved single-copy BUSCO (Benchmarking set of Universal Single-Copy Orthologs) gene set from OrthoDB [172-174], 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 S2D), 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 S2C). 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 (Table 1). 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 (Materials and Methods).

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 [175], and compared to known RiPP BGCs from MIBiG [29, 140]. In addition, the overlap between predicted RiPP BGCs and gene clusters found by antiSMASH [39, 77] is determined (Figure 1).

decRiPPter identifies 42 candidate novel RiPP subclasses in *Streptomyces*

While RiPPs are found in many different microorganisms, their presence in streptomycetes reflects perhaps the most diverse array of RiPP subclasses within a single genus. Streptomycete*s* produce a broad spectrum of RiPPs, such as lanthipeptides [176], lasso peptides [45], linear azol(in)e-containing peptides (LAPs) [177], thiopeptides [46], thioamide-containing peptides [52] and bottromycins [97, 178, 179]. Their potential as RiPP producers is further highlighted by a recent study showcasing the diversity of lanthipeptide BGCs in *Streptomyces* and other actinobacteria [68]. Even though any genus or set of genomes can be analyzed by the decRiPPter pipeline, we hypothesized streptomycetes to be a likely source of novel RiPP subclasses, 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 (Data S2). 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

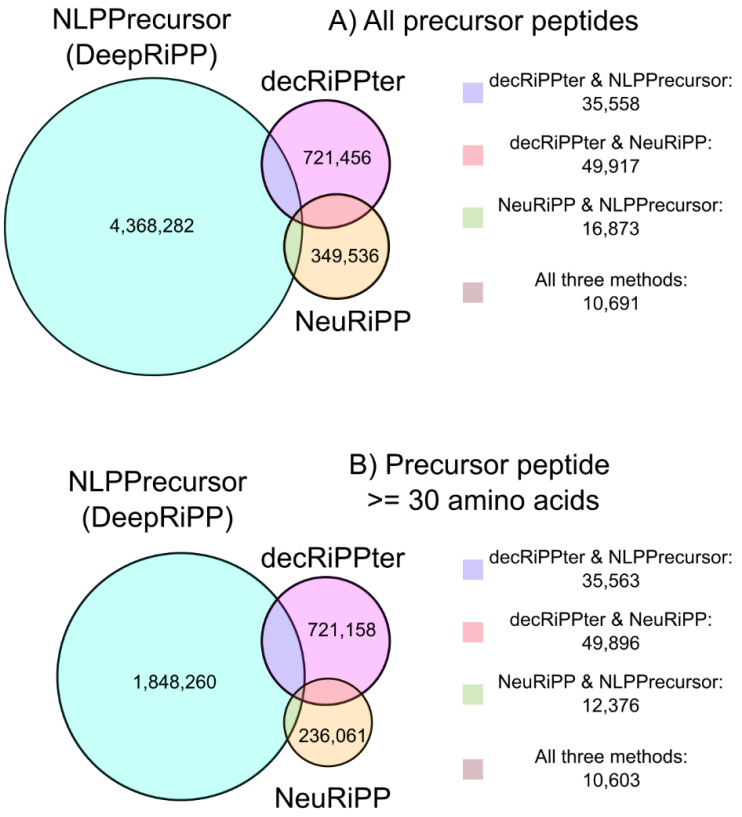


Figure 3. Three machine-learning-based RiPP precursor classifiers give highly different results. All small ORFs from the 1,295 *Streptomyces* genomes were classified by DeepRiPP's NLPPrecursor [89] module, NeuRiPP [88] and decRiPPter. The three tools have only a small overlap (10,691 hits). NLPPrecursor scored six times more hits as positive, and NeuRiPP roughly half when compared to decRiPPter. Many of these hits were very small ORFs (≤ 30 amino acids; (B)), though, while most of decRiPPters predicted precursors were larger than that. The exact accuracy of these tools cannot be determined, as it is unclear which of these hits are false positives, and which are hits in novel RiPP BGCs.

almost all genomes (Figure S3A). We then scored 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-based 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 S2B). 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 %). As a comparison, all ORFs were also analyzed by NLPPrecursor and NeuRiPP (Figure 3) [88, 89], and overlapping hits were removed as was done with decRiPPter's hits. For all three tools, a large number of candidate precursors were hits: NLPPrecursor scored the most ($4.4*10^6$), and NeuRiPP the least ($4.3*10^5$). Surprisingly, the three tools showed little overlap in positive hits ($1,1*10^4$). Considering that NLPPrecursor was parametrized for the detection of precursors of known subclasses and NeuRiPP appeared to be more strict (while our goal was to be more exploratory), we continued with decRiPPter's hits. In principle, the precursor-peptide-finding module of decRiPPter could easily be replaced by, e.g., NeuRiPP in future analyses for which this would be desirable.

We noticed that the majority of the precursor hits of decRiPPter were not found by Prodigal, but were extracted from intergenic regions ($6.6*10^5$ intergenic, $1.6*10^5$ from Prodigal). A GC-plot analysis of 112 hits of both intergenic and Prodigal-detected genes showed that only 5-10% of the intergenic hits showed a GC-plot with clear distinctions between the first, second and third codon position, while the majority of Prodigal-detected genes had the same distinction (Figure S4). These intergenic regions are likely a source of many false positives, and for a more conservative approach one could choose to ignore intergenic hits altogether. Since our aim was to conduct an explorative study to detect novel subclasses, and gene-finding algorithms do frequently miss precursor genes, we chose to continue with all the precursors hits found here.

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. As a comparison, we divided all known RiPP BGCs and all antiSMASH RiPP BGCs found in the analyzed genome sequences into sections containing only adjacent 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 validation sets to fine-tune distance and COG cutoffs for two gene cluster formation methods, which we called the 'simple method' and the 'island method'.

In the 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 S5AB). However, using only distance may cause the gene cluster formation to overshoot into regions not associated with the BGC (e.g. Figure S1). 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 (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 (number of genes: 3.73 ± 3.75 vs 3.44 ± 3.53; Figure S5CDE). 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 S5F).

Overlapping gene clusters were fused, resulting in $7.18*10^5$ gene clusters. To organize the results, all gene clusters were paired to other gene clusters with similar protein domain content (Jaccard index of protein domains; cutoff: 0.5) and containing at least one predicted precursor gene with sequence similarity (NCBI blastp; bitscore cutoff: 30). These cutoffs were shown to distinguish between different RiPP subclasses (Figure S6). 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. The majority of BGCs previously detected by RODEO were also found to overlap (84%, Table S3). Most of the antiSMASH BGCs missed by decRiPPter belonged to the bacteriocin family, which do not necessarily encode a small precursor peptide (Table S3). The remainder of missed BGCs are likely due to precursor genes not being on the same strand as the genes encoding the biosynthetic machinery or due to precursor genes missed by decRiPPter's classifier.

Table 1. Correlation between the strictness of the filter used on the identified gene clusters and the saturation of RiPP BGCs. Genes were considered as being around the gene cluster if within five genes.

Filter	Filter details	Number of detected gene clusters	Gene clusters overlapping antiSMASH RiPP BGCs (percentage)
None	-	718,268	5,908 (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 gene	21,419	1,678 (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 gene • >= 1 regulatory gene • >= 1 peptidase gene	2,471	357 (14.4)

The hits overlapping with antiSMASH constituted only 0.8% of all 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 2). 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 3). 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 in 151 families; Figure 4). While this step eliminated BGCs for hybrids of RiPP and non-RiPP pathways, we felt this filter was necessary to reduce the number of false positives in our dataset, especially considering the rarity of these hybrid BGCs.

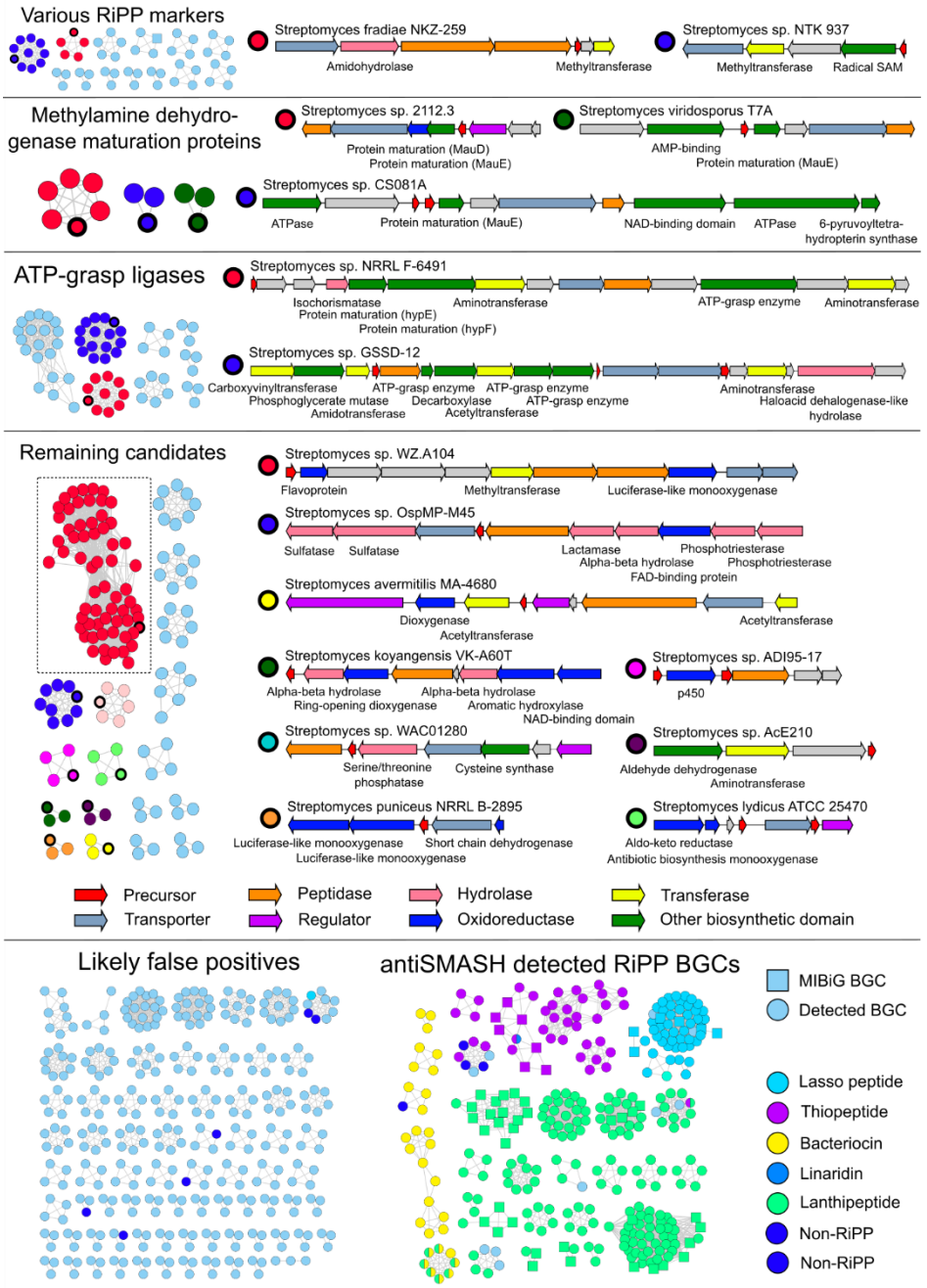
Roughly a third (280) of the remaining gene clusters were members of known families of RiPPs, including lasso peptides, lanthipeptides, thiopeptides, bacteriocins and microcins. In addition, many of the other candidate clusters (54) contained genes common to known RiPP BGCs, such as those encoding YcaO cyclodehydratases and radical SAM-utilizing proteins (Figure 4) 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 [52, 103].

Each remaining family of gene clusters was manually investigated to filter out likely false positives from the candidates. A set of general guidelines followed can be found in the Materials and Methods. 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 [180]), annotations of multiple 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 clusters of more than 50 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 [160]. We therefore only discarded a gene cluster family if multiple clear relations to a known pathway were found. The remaining 42 candidate families, containing were further grouped together into broader families depending on whether a common enzyme was found (Figure 4).

Among our candidate families, 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. 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 [181]. Involvement of ATP-grasp enzymes in RiPP biosynthesis has been reported for microviridin [83] and other omega-ester containing peptides (OEPs) [84], and for pheganomycin [162], 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 are known to be involved in the maturation of of methylamine dehydrogenase, 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 [182]. 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 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 2).

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 [183]. No other homologs of genes in the *hyp* operon were detected, suggesting that these protein-coding genes have adopted a novel function.


Various RiPP markers
Streptomyces fradiae NKZ-259
Amidohydrolase
Methyltransferase
Streptomyces sp. NTK 937
Methyltransferase
Radical SAM
Methylamine dehydrogenase maturation proteins
Streptomyces sp. 2112.3
Protein maturation (MauD)
Protein maturation (MauE)
Streptomyces viridosporus T7A
AMP-binding
Protein maturation (MauE)
Streptomyces sp. CS081A
ATPase
Protein maturation (MauE)
NAD-binding domain
ATPase
6-pyruvoyltetrahydropterin synthase
ATP-grasp ligases
Streptomyces sp. NRRL F-6491
Isochorismatase
Protein maturation (hypE)
Protein maturation (hypF)
Aminotransferase
ATP-grasp enzyme
Aminotransferase
Streptomyces sp. GSSD-12
Carboxyvinyltransferase
Phosphoglycerate mutase
Amidotransferase
ATP-grasp enzyme
Decarboxylase
Acetyltransferase
ATP-grasp enzyme
ATP-grasp enzyme
Aminotransferase
Haloacid dehalogenase-like hydrolase
Remaining candidates
Streptomyces sp. WZ.A104
Flavoprotein
Methyltransferase
Luciferase-like monooxygenase
Streptomyces sp. OspMP-M45
Sulfatase
Sulfatase
Lactamase
Alpha-beta hydrolase
FAD-binding protein
Phosphotriesterase
Phosphotriesterase
Streptomyces avermitilis MA-4680
Dioxygenase
Acetyltransferase
Acetyltransferase
Streptomyces koyangensis VK-A60T
Alpha-beta hydrolase
Ring-opening dioxygenase
Alpha-beta hydrolase
Aromatic hydroxylase
NAD-binding domain
Streptomyces sp. ADI95-17
p450
Streptomyces sp. WAC01280
Serine/threonine phosphatase
Cysteine synthase
Streptomyces sp. AcE210
Aldehyde dehydrogenase
Aminotransferase
Streptomyces puniceus NRRL B-2895
Luciferase-like monooxygenase
Luciferase-like monooxygenase
Short chain dehydrogenase
Streptomyces lydicus ATCC 25470
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

As stated above, 175 gene cluster families, containing a total of 1,036 candidate gene clusters, were left after the strict filter. Of these, 24 families containing 143 gene clusters were removed due to overlap with non-RiPP BGCs. An additional 74 families containing 341 gene clusters were removed by manual curation, making for a total false positive count of 98 families containing 484 gene clusters, just under half of the total (46.7%). A total of 32 families containing 280 gene clusters overlapped with known RiPP BGCs (27.0%), which can be considered true positives. The remaining 272 gene clusters (42 families; 26.3%) are the presented candidates. This means that the actual true positive rate lies between 27.0% and 53.3%, and the false positive rate between 46.7% and 73.0%, depending on the nature of the candidates. For the results from the mild filter, 1,678 gene clusters out of 21,419 were overlapping with known RiPP BGCs (7.8%). How many of the remaining gene clusters (92.2%) are false positives and how many are novel RiPP BGCs can not be determined without a thorough manual examination. From the results of the strict filter, however, it appears there are roughly as many novel RiPP BGCs as there are known ones (272 vs 280). Extending thes ratios to the results of the mild filter would mean that an additional 1,678 gene clusters are novel RiPP BGCs, resulting in an estimated true positive rate of 15.6% and false positive rate of 84.4%. These high false positive rates emphasize that one should interpret the results with caution. However, if even half of the proposed candidates are true RiPP subclasses, this would represent a significant contribution to the total amount discovered.

Figure 4 (opposite page). 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. Each node of the network represent a candidate BGC, while edges represent similarity in both precurs and enzyme domains. The four panels at the top contain families of interest, grouped by common defining characteristics, if present (top panel: 54 gene clusters in 13 families; second panel: 12 gene clusters in 3 families; third panel: 65 gene clusters in 8 families; fourth panel: 141 gene clusters in 18 families). The bottom panel contains the gene clusters marked as likely false positive (left side, 341 gene clusters in 74 families) and the gene clusters overlapping with antiSMASH-detected RiPP BGCs (right side, 280 gene clusters in 33 families). Examples of 15 gene clusters of candidate families are given (nodes with dark circles). Arrow colors indicate enzyme family of the product, and the description of the putative gene products is given below the arrows. The candidate RiPP family represented by the network outlined with a dashed box is discussed further in Chapter 4.

Table 2. Precursor sequences of selected BGCs of candidate RiPP families shown in Figure 4. Serine and threonine residues are marked in green, and cysteine residues are marked in red.

Family	Strain	Precursor sequence
Known RiPP markers	*Streptomyces* sp. NTK 937	MTENTAPEESPEVEAHSAADDAAQAPEQFHDAAEIICGVYDKEIQV
Known RiPP markers	*Streptomyces fradiae* NKZ-259	MPSGMPNDPSTTDGLSRRRVLGTAAAAAVPLPARGAEDAEAKSGPW
Containing MauE	*Streptomyces viridosporus* T7A	MSRALESLSSRLLGLFVPKVEAAASAQACQCFNECWQCARSACCVNT YCGSINCWRSCPGC
Containing MauE	*Streptomyces* sp. CS081A	MARTVGDGSKGCRPSPVSPYGLDQYGDRAASTWGASSATCGVRGEP
		MVKSLSALAGRAFARVLPQETAAAACACPAGSSSWCSGENLYTRFCCS WNCAAKPTCTVTVVYGAC
Containing MauE	*Streptomyces* sp. 2112.3	MFKKLEAVGSALLERLVPRVDASACGTNCWNDCWQCAHSACKVNTC TGALTCLSGNC
ATP-grasp ligases	*Streptomyces* sp. NRRL F-6491	MARAARNLLAITASAALSFLLVQGTGAQEERAFLAGSGQGKVINDLG WG
ATP-grasp ligases	*Streptomyces* sp. GSSD-12	MSSDPSDAAEQGPVGGFITEPLVAAAATTGGCCGEPRSAPEPARSSCC GEPAAEEAPRSCCGEPAAAG
		MADDMIGSGCCETSGNEDVAEDGTECGCACACCD
		MSETSLGNMFWNAAQQPPAATAEEPKKASSCCGPKPEAKAPAEQAA APEKASSCCGPKPAAAAEPEGTPAPKKSSCCG
Other	*Streptomyces* sp. WZ.A104	MQNVTEKDLFDGYTAYTSAEELGLHDGKEAAPAFSPTIPWAIRATIISA RSSQQCAAALGSLAAKTVENKC
Other	*Streptomyces* sp. OspMP-M45	MTEAGLWEEGDAGRRRPLGVPPENWPVPGGRQGMDGQWSGQSS KTIDHPGGAT
Other	*Streptomyces avermitilis* MA-4680	MSSLDKPGRKKWSGPEKWQVILAASSLGVAVVALVGQFAQFL
Other	*Streptomyces koyangensis* VK-A60T	MGDLDEEVAAPGPGRWIRPSSTAGYGWTTSCRTSVFPPASPDSCQAR ETVTWCAWVP
Other	*Streptomyces* sp. ADI95-17	MNSLSEAGCWCHERLKSCPSECKFRVKDGGAVMKFLFLLKDKMTPEK SLKAYAWYHWY
		MCEVCRSSRNPGPWGGCCGDGARLGHGWPVSYYETLLCKSQPHEGL DLGASIGEGFEPTPGDLPAGGQSPHKE
Other	*Streptomyces* sp. WAC01280	MLKGGQLGRFSTNSMNDHREQLGIGPPCLLTFDNAARSSQPSQEAAP CARAES
Other	*Streptomyces* sp. AcE210	MAESPTPEAVAEQPTEVAQPHRLVLLGACGCGSGCGCGCQSGAPCQ CGGCSG
Other	*Streptomyces puniceus* NRRL B-2895	MRTAAAYASGEPPPVAVVKSHGVAFENRVRYVSPVPSTTHAAASAPG SAEGSAPAATA
Other	*Streptomyces lydicus* ATCC 25470	MLWKSCARARCGISIPWNSFEFDHGGTGVVPCVPGVCEFPARDGKEE VT
		MNQGGGEQRGAEVSIRANVGSWLAVRKSPFEAGGSPVSRWEDLPR GVPCPYETGAHQD

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. The pipeline can be run on any set of genomes. We recommend choosing a set of genomes that are sufficiently closely related to share a `core genome` for the COG-score calculations. At the same time, genomes should not be too similar, so that a wide variety of BGCs can be found among them that show variability in their presence/absence pattern across genomes. 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 https://decrippter.bioinformatics.nl.

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 [26]. Tools primarily aimed at accuracy reliably discover large numbers of known natural product BGCs, but are limited by specific genetic markers. On the other hand, while tools aimed at novelty may lead to the discovery of 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 subclasses. 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 [26] is necessary for the discovery of completely novel RiPP subclasses. Additionally, users are able to set their own filters for the identified gene clusters, allowing them to search candidate RiPP subclasses containing specific enzymes or enzyme types within a much more confined search space compared to manual genome browsing. As such, decRiPPter can function as a platform for explorative RiPP genome mining, enabling a large variety of different search strategies to explore further into RiPP chemical space.

Materials and Methods

decRiPPter pipeline

Genome data preparation

As input, decRiPPter uses a set of genomes from species that are part of the same taxonomic group (e.g., genus, family), which it requires for its comparative genomic analyses. decRiPPter downloads genomes from NCBI [184] based on NCBI taxonomic identifiers of species, genera or higher orders of classification. Additional requirements for level of assembly (e.g. "Representative genome") can also be given. decRiPPter can reannotate genomes with prodigal 2.6.3[71], and automatically does so when DNA FASTA files are given as input. In addition, users may analyze their own genomes, in isolation or in conjunction with downloaded genomes.

SVM-based classifier

To predict RiPP precursors, we first collected positively and negatively labeled training data. The positive training data was collected from MIBiG [140] and recent literature, resulting in 175 RiPP precursors across ten subclasses. For the negative training set, we generated a set of 20,000 short non-precursor sequences. Half of these were randomly selected from a set of 35,000 short proteins (<175 amino acids long) from Uniprot (queried June 2014) that were not similar to RiPP precursors based on an NCBI blastp search. The other half were randomly selected from a set of 17,000 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: Escherichia coli, Bacillus subtilis, Streptomyces coelicolor, Bacteroides fragilis, Rhizobium etli, Chloroflexus aurantiacus, Synechococcus sp. PCC 7002, Opitutus terrae, Acidobacterium capsulatum and Pirellula staleyi. For all sequences from both the positive and negative training sets, we computed several physio-chemical properties, such as its length, hydrophobicity, charge, counts of canonical amino-acid residues and classes of amino acids, and highest counts of, e.g., cysteines and serines within contiguous blocks of 20 or 30 amino acids. The method for computing these properties is part of the decRiPPter pipeline, and can be found in the code repository, at https://github.com/Alexamk/decRiPPter/blob/master/lib/features.py. All training data and data collection scripts are available online (https://zenodo.org/record/3834818#.X7JmIOTsbvs)

We then utilized Scikit-Learn implementations of several different supervised machine-learning algorithms. We varied several parameters associated with a given algorithm (e.g., kernel functions, penalty parameters, penalty functions, etc.). Furthermore, we mapped the accuracy as a function of scaling the dataset or changing class weights to take into account the unbalanced dataset (only ~1% of gene clusters in our dataset represent known RiPPs). The RiPP cluster classification accuracy of each combination of scaling, algorithm, and the corresponding set of parameters was evaluated using accuracy and area under receiver operating characteristics (ROC) curve, and leave-one-class-out cross-validation. SVMs with three different kernel functions were trained: two with polynomal kernel function (SVM3: 3rd degree, coef0 of 2.154, kernel coefficient gamma of 2.78*10^-2, regularization parameter C of 0.158; SVM4: 4th degree, coef0 of 2.154, kernel coefficient gamma of 4.64*10^-3, regularization parameter C of 25.119) and one with a radial basis function kernel (SVMr: kernel coefficient gamma of 1*10^-5, regularization parameter C of 6.310*10^5). For each type, one SVM was trained with all training data, while eighteen more were trained by leaving out the sequences of one RiPP subclass from the positive training data at a time. The average of all 57 SVMs was taken as the final SVM score.

COG scores calculation

To calculate the relative frequency of occurrence of each gene, we constructed a pipeline to find all groups of homologous genes (Figure 2). In the first step, protein-coding genes for which orthology can confidently be assigned are grouped into Clusters of Orthologous Groups (COGs). All proteins are aligned to one another using DIAMOND [168], and all bidirectional best hits (BBHs) are identified that share at least 60.0% similarity (Figure 2A). We established two requirements for genes to be confidently annotated as orthologs, based on recent papers [167, 169]: 1) they should constitute BBHs, and 2) their immediate genomic surroundings should be conserved, i.e. the two flanking genes should also be bidirectional best hits between the two genomes. Genes fulfilling these two criteria are paired together, resulting in groups of orthologous genes. Among these groups, decRiPPter then selects those that are completely conserved across all genomes: each group should contain at least one ortholog in each genome, and all orthologs in the group should all fulfill the same requirements for each genome pair. These groups are considered true Clusters of Orthologous Genes (trueCOGs; Figure 2B).

In the second step, a cutoff for protein-coding gene sequence identity is determined for each genome pair, in order to separate orthologs as well as recently evolved paralogs from more distantly related homologs. For any given pair of genomes, the distribution of sequence identities of all gene pairs of their trueCOGs is calculated. The cutoff is then calculated as the average percentage identity, minus three times the standard deviation (Figure 2C). Any two aligned genes with a percentage identity higher than this cutoff are considered to be functionally closely related to one another and paired up. The resulting groups of homologous genes were clustered with the Markov Cluster Algorithm[170, 171] (Figure 2D). From these groups, the relative frequency of occurrence of groups of homologous genes across all query genomes is calculated, called the COG-score (Figure 2E).

In cases when insufficient numbers of trueCOGs (<= 10) could be found in our analyses (because the set of genomes was too diverse, and/or contained too many draft genomes that each miss some of the trueCOGs), the genomes were rearranged into smaller subgroups. We used two general rules to create the groups: 1) Groups should be as large as possible, so that trueCOGs found are conserved across many species, and represent conserved widespread genes. 2) Genomes should be compared to as many other genomes as possible, so as not to introduce bias into the calculation of the COG-score. To fulfil both requirements, partially overlapping subgroups were formed, with the goal of letting each genome be a part of a collection of subgroups that together covered as many of the genomes as possible. To form the subgroups, a pair of genomes with the highest number of trueCOGs was used as a seed, and genomes were added one at a time until the number of trueCOGs dropped below the set cutoff. All the genomes in the group were said to be linked together by this group. The process of group formation was then repeated, starting with genomes for which no group had yet been formed. If all genomes were already part of at least one subgroup, the genomes were selected which were linked to the fewest genomes via the groups they were part of. The process was terminated when adding additional groups did not increase the number of links between genomes for several successive iterations.

Gene cluster formation

In this stage, decRiPPter identifies putative operon-like gene clusters around each candidate precursor peptide-encoding gene, by either of two different methods (Figure S1): In the first method, called the simple method, genes in the same strand orientation as the candidate

precursor peptide-encoding gene are added to the putative gene cluster if the intergenic distance to the previous gene is within a given cutoff. The second method, called the island method, uses both intergenic distance and levels of conservation (COG-score) to determine the gene clusters. First, all genes in the same strand orientation within 750 nucleotides of one another are identified and then grouped into islands. Within islands, genes should be almost directly adjacent (intergenic distance: <= 50 nucleotides). We then fused the islands together using the COG-scores (see above), building on the assumption that genes in a gene cluster should all have similar levels of conservation. Islands were fused together if the average of their COG-scores was within a set range (0.1 plus the sum of the standard deviations of both islands). Not all gene families have similar COG scores when they occur within the gene clusters thus formed; e.g., genes encoding ABC-transporters frequently have close relatives in other biomolecular systems and therefore often have higher COG scores. Hence, to counteract gene cluster formation breaking off prematurely, up to two outlier genes are allowed when fusing islands, if, after adding the outliers, more islands can be added that are within the range for COG-score deviation. Intergenic distances and cutoffs were iteratively finetuned to ensure gene clusters in known RiPP BGCs would be effectively found. Finally, gene clusters that overlap or lie within 50 nucleotides of one another are fused together.

Annotation

For purposes of data exploration (annotation and visualization), each gene cluster is extended to include the 5 flanking genes on either side, and all encoded proteins in the extended gene clusters are annotated with Pfam 31.0 [75] and TIGRFAM [76]. Lists were compiled of all TIGRFAM and Pfam domains associated with either peptidases, transporters, regulators, using a combination of keyword searches on the Pfam and TIGRFAM websites, combined with manual curation. A list of protein domains associated with biosynthetic activity was constructed by linking Pfam domains to E.C. numbers, using InterPro mappings [165]. Biosynthetic TIGRFAM domains were taken directly from the database. Each domain linked to an E.C. number was assumed to have enzymatic activity. The biosynthetic domain list was further expanded with domains used in the ClusterFinder [41] algorithm that were indicative of a biosynthetic gene cluster. The resulting lists are used by decRiPPter to mark proteins either as a regulator, peptidase, transporter or biosynthetic enzyme, in that order, by seeing if any of the identified domains overlapped with the domains in the precompiled lists (Data S1).

Clustering

To cluster the detected gene clusters, the distance between them is calculated in two different ways: 1) amino acid sequences of candidate precursor peptide-encoding genes in the gene clusters are aligned with NCBI BLAST blastp [56] (cutoff: 30 bitscore), and 2) the content of the gene clusters is compared by calculating the Jaccard index of their constituent protein domains (cutoff: 0.5). Gene clusters are paired only if they are paired by both methods. The distance between paired gene clusters is calculated as the average between the Jaccard index and the percentage identity of the aligned precursors. Finally, pairs are clustered using MCL.

Overlap with antiSMASH

Overlap with antiSMASH was determined using antiSMASH 4.0 [77] run in minimal mode.

Availability

The decRiPPter pipeline is available at https://github.com/Alexamk/decRiPPter/. Data from the analysis discussed here is available at https://decrippter.bioinformatics.nl.

Data analysis

Comparison with NeuRiPP and NLPPrecursor

NeuRiPP classifications were performed using the parallel CNN network with the network weights provided by the author [88]. NLPPrecursor was installed and executed with default settings [89]. All open reading frames were analyzed with both methods, and completely overlapping precursor hits on the same frame were removed, as in the decRiPPter pipeline.

Supplementary information for Chapter 3

Data S1. Categorized Pfam and TIGRFAM domains used in decRiPPter pipeline. Available from https://github.com/Alexamk/decRiPPter/tree/master/data/domains/.

Data S2. *Streptomyces* genomes analysed with decRiPPter. Available upon request.

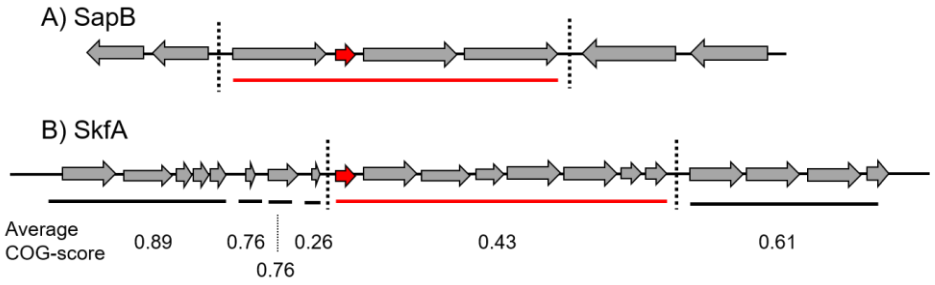


Figure S1. decRiPPter forms putative gene clusters around candidate precursor peptide-encoding genes. Two examples are provided here to illustrate identification of putative gene clusters in decRiPPter. A) In the *sapB* gene cluster, four genes form the main BGC. These four genes are sequential, share the same strand orientation and lie within a small distance of one another (<= 50 nt). They are therefore fused together into a single gene cluster. The flanking genes are on opposite strands, and therefore not considered. B) The *skfA* BGC consists of eight genes sequential genes that share the same strand orientation. However, it is flanked by several other genes that also share the same strand orientation, within relatively short intergenic distances (<= 200 nucleotides). Using the island method, the genes are first fused into six islands, within 50 nucleotides distance of one another (indicated by lines underneath the genes). These islands may then be fused depending on the COG-score, which does not happen here because the difference is too large. The result is that the flanking genes, with a too high COG-score, are not added, and the correct BGC remains.

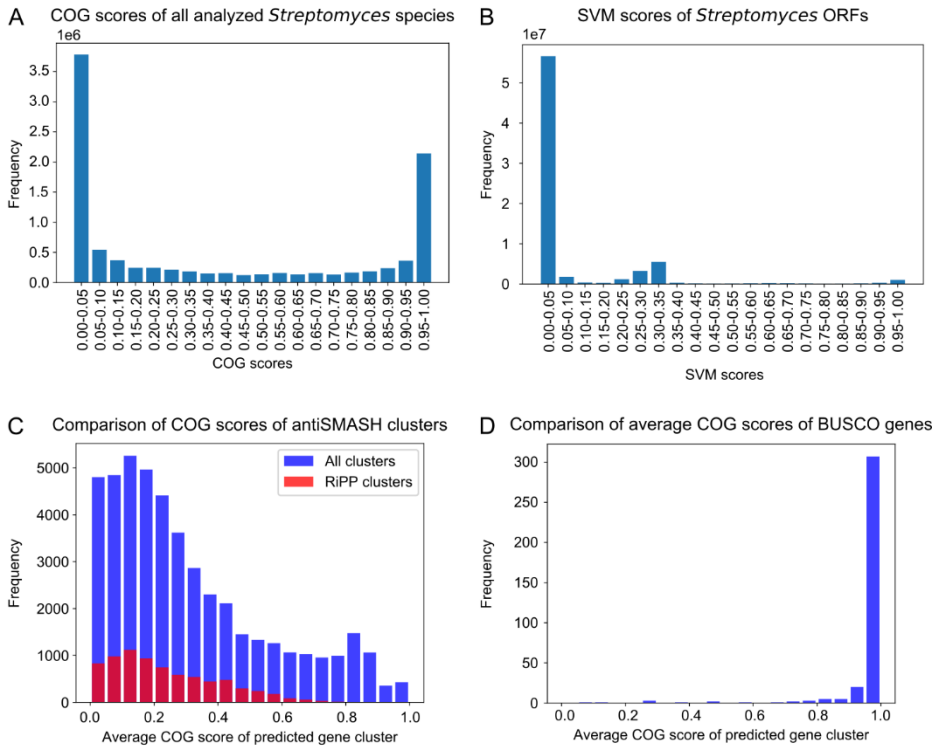


Figure S2. COG and SVM scores in all analyzed 1,295 *Streptomyces* genomes. A) COG scores of all genes in all 1,295 analyzed *Streptomyces* genomes. A high COG score indicates presence of homologs in many different genomes, while a low COG score indicates a more infrequent distribution. COG scores were calculated as described in the methods. B) Distribution of the scores assigned by decRiPPter's SVM-based classifier. A total of 7,1 * 10^7 small ORFs were analyzed. C) Comparison of COG scores of antiSMASH-detected gene clusters. COG scores were averaged over all genes in the predicted gene clusters. COG scores averaged 0.311 +- 0.249 for all gene clusters, and 0.234 +- 0.166 for RiPP gene clusters. D) Comparison of average COG scores of BUSCO genes. The average of each BUSCO [173, 174] gene was calculated for each genome analyzed.

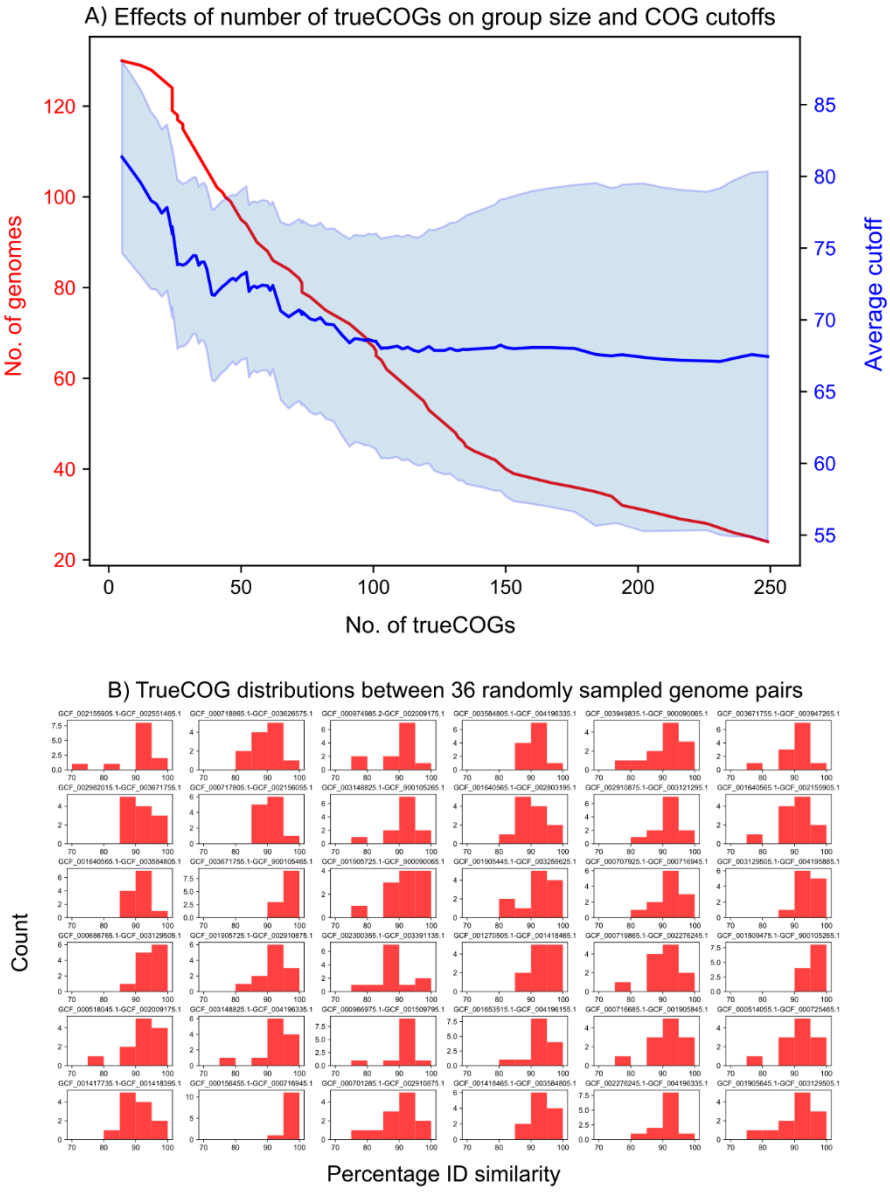


Figure S3. COG-scores calculations depend on genome group size. A) As the minimum number of trueCOGs increases, the number of genomes that can be analyzed together (red line) decreases. In addition, the average COG cutoff (blue line) decreases when more trueCOGs are added, and the spread of COG cutoffs (shaded area; average cutoff +- the standard deviation) increases, suggesting that additional trueCOGs that were added were less conserved and showed higher variability in sequence similarity. B) TrueCOG distribution between 36 randomly sampled genome pairs. Based on these distributions, COG cutoffs were determined.

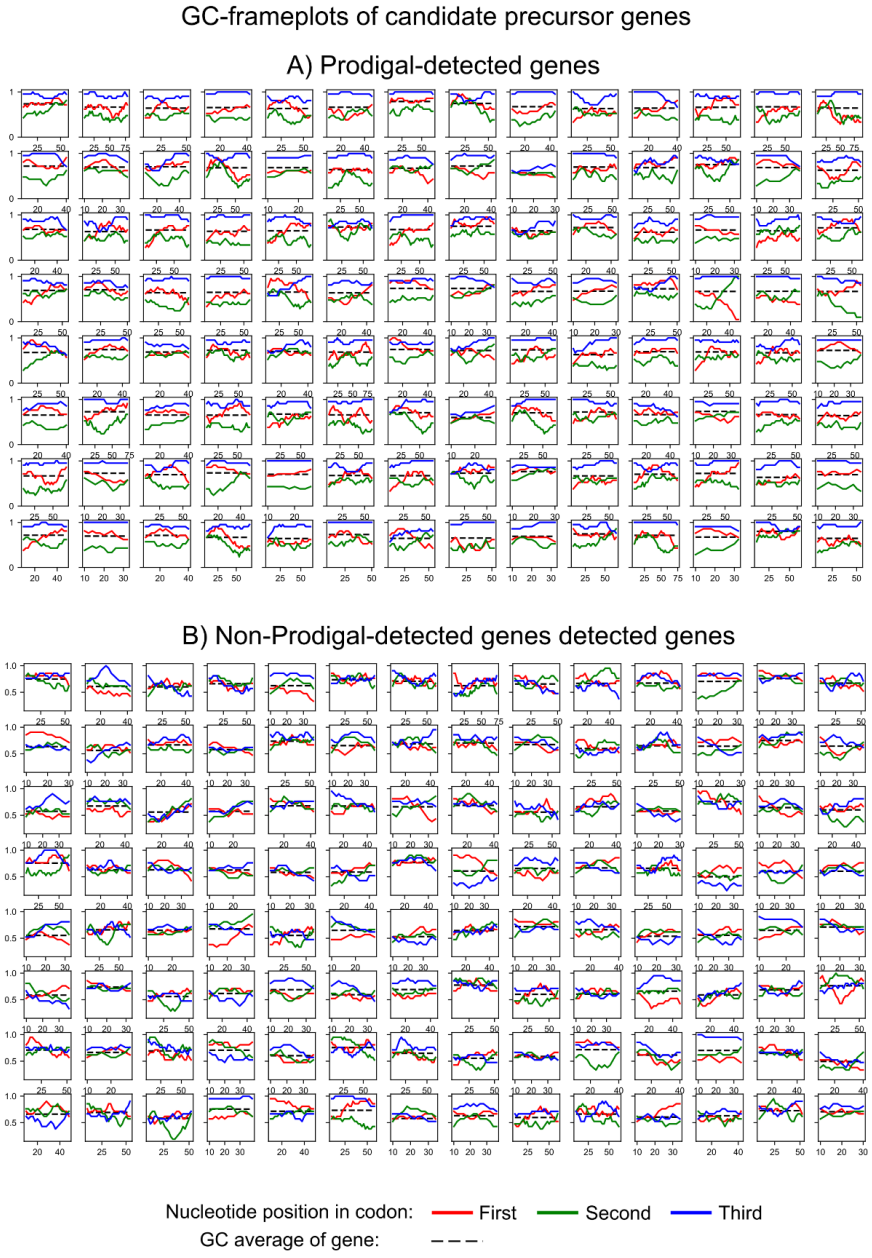


Figure S4. GC-plots of randomly sampled Prodigal-detected precursor hits (A) and intergenic precursor hits (B). GC values are shown as the moving average of the first, second and third positions, using a window-size of 5 and a step-size of 2. Only a small percentage of intergenic hits showed clear distinction between the three moving averages as in the Prodigal-detected hits, suggesting the majority of these are not encoding genes.

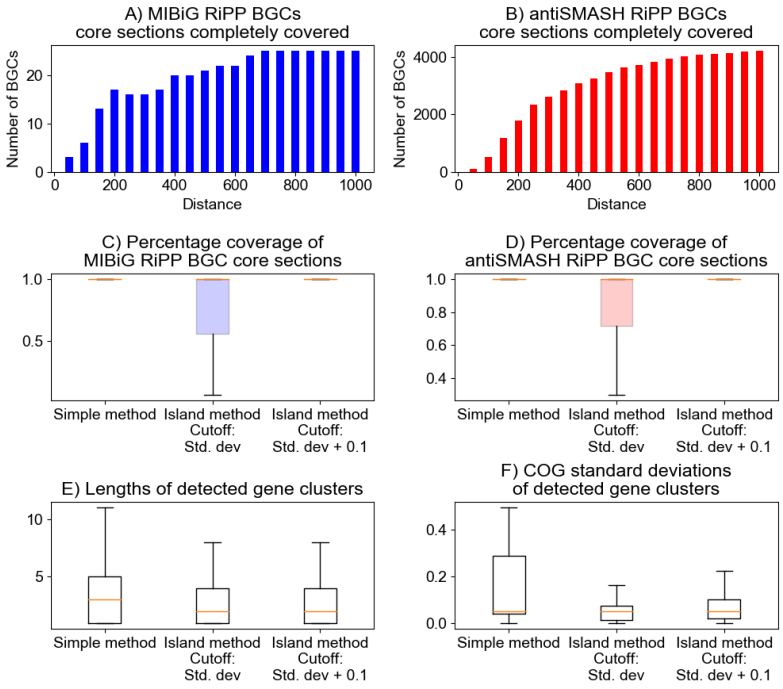


Figure S5. Gene cluster formation effectively covers antiSMASH and MIBiG BGC core gene sections. In the simple gene cluster formation method, genes are sequentially added as long as they are in the same strand orientation, within a certain distance. At a distance of 700 nucleotides, all MIBiG core gene sections are covered (A), as well as 91% (3947/4321) of antiSMASH core gene sections. (B). In the 'island method', genes are first fused into islands, which may be further fused if their average COG-scores are within a cutoff. Using just the standard deviation of the islands as a cutoff resulted in incomplete coverage of both the MIBiG and the antiSMASH core sections (C, D, middle boxes). Increasing the cutoff to the standard deviation plus 0.1 resulted in comparable coverage (C, D, right boxes) of these sections when compared to the simple method (C, D, left boxes). In addition, the overall gene cluster length (E) and variation of COG scores (F) within all formed gene clusters decreased.

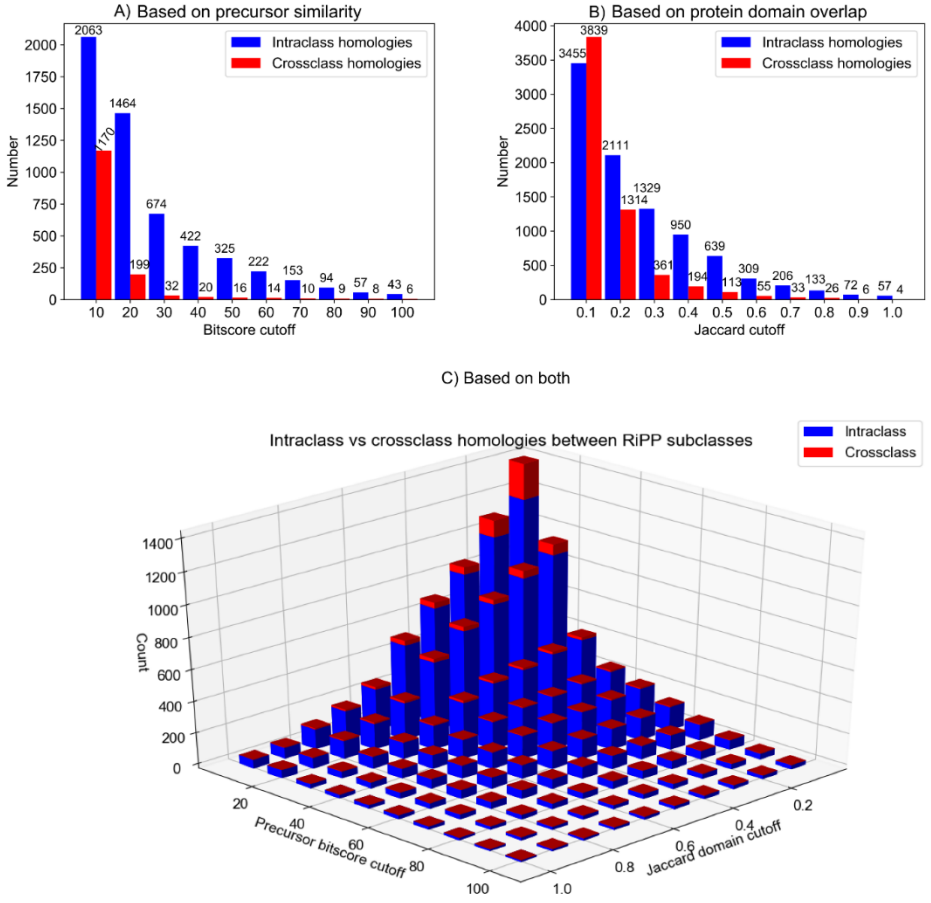


Figure S6. Combining precursor similarity with domain similarity is an effective strategy to group RiPP subclasses. Starting at precursor similarity bitscore cutoffs of 20 and Jaccard scores of overlapping protein domains found in MIBiG RiPP BGCs of 0.4, the number of intraclass homologies is larger than the number of crossclass homologies. Combining the two methods greatly decreases the number of cross-class homologies found, proving it as an effective method to group RiPP BGCs of different subtypes.

Table S1. RiPP classes in positive training data of decRiPPter.

RiPP class	Amount of precursors
Bottromycin	3
Cyanobactin	14
Glycocin	1
head-to-tail cyclized peptide	10
Lanthipeptide	79
LAP	4
Hybrid	4
lasso peptide	13
Linaridin	2
Microcin	7
Microviridin	4
Proteusin	1
Sactipeptide	4
Thiopeptide	12
Unclassified	17

Table S2. decRiPPter detects most RiPP precursors of known classes found by RODEO. RODEO results were extracted from previous studies [55, 72-74, 86].

RiPP Class	Number detected by RODEO	Scored ≥ 0.9 by decRiPPter
Lanthipeptide	453	329
Lasso peptide	5270	3738
Linaridin	2152	1127
Sactipeptide/ranthipeptide	1524	953
Thiopeptide	399	387
Total	9798	6534

Table S3. Comparison of detected BGCs with antiSMASH and RODEO. Note that not all genomes were analyzed by RODEO. Results from earlier RODEO genome mining [55, 72-74, 86] where only used if within the 1,295 *Streptomyces* genomes.

RiPP Class	RODEO BGCs	Overlap (no filter)	Overlap (mild filter)	Overlap (strict filter)	antiSMASH BGCs	Overlap (no filter)	Overlap (mild filter)	Overlap (strict filter)
Lanthipeptide	1530	1447	421	102	2768	2570	850	175
Lasso peptide	397	175	112	14	878	742	315	59
Linaridin	97	85	33	4	229	199	82	5
Thiopeptide	71	45	23	4	612	584	264	57
Sactipeptide/ ranthipeptide	1	1	1	0				
Bacteriocin					2735	1402	184	41
Bottromycin					2	2	0	0
Cyanobactin					31	27	3	1
Proteusin					2	2	2	0
RiPP hybrid					321	312	96	32

4

Characterization of a novel lanthipeptide class discovered with a machine-learning genome mining tool

Alexander M. Kloosterman*
Somayah S. Elsayed*
Chao Du
Marnix H. Medema
Gilles P. van Wezel

*These authors contributed equally to this work.

The work described in this chapter is part of the publication:

Kloosterman, et al., *Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics.* PLOS Biol, 2020. **18**(12): e3002016.

Abstract

Ribosomally synthesized and post-translationally modified peptides (RiPPs) represent a highly diverse and quickly expanding class of natural products that is divided into genetically and chemically distinct subclasses. The identification of novel subclasses is an excellent opportunity to identify chemical scaffolds and expand our knowledge of biosynthetic pathways, but unsuitable for similarity-based genome mining. Here, we report on the characterization of a novel RiPP subclass that has been identified using decRiPPter, a bioinformatic tool for the discovery of novel RiPP subclasses. This RiPP subclass is commonly found among streptomycetes, with one BGC present in every ten genomes. A representative gene cluster from *Streptomyces pristinaespiralis* was selected for characterization. Placing a nearby regulator behind a constitutive promoter resulted in the activation of the BGC, and several masses were detected in the crude chemical extracts of cultures with LCMS. MS/MS fragmentation analysis, chemical labeling and NMR were combined to elucidate the structure one of the most abundant masses, which could be linked to one of the predicted precursor peptides encoded in the gene cluster. Structural analysis showed that this gene cluster specifies lanthipeptides, called the pristinins, despite the fact that the gene cluster did not contain genes for canonical enzymes that create the required lanthionine bridges. The lanthipeptide therefore belongs to a novel class, which we call class V. Through comparisons with previously identified RiPPs, two genes are proposed to encode the enzymes that form the lanthionine bridge in this subclass. These genes are present in a wide variety of genetic contexts, both within *Streptomyces*, but also in other *Actinobacteria* and in *Firmicutes.* This work not only showcases the potential of decRiPPter, but also expands the list of RiPP subclasses and identifies promising enzyme queries that can be used in further genome mining studies.

Introduction

The continuing increase in available sequence data has fueled the identification of many new natural product biosynthetic gene clusters (BGCs) through genome mining [26, 185]. Homology-based genome mining methods expand classes of natural products through the identification of key genetic markers in contexts. Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a highly diverse collection of natural products, which are split into several subclasses that share biosynthetic pathways [42, 48]. Identification of novel RiPP subclasses using bioinformatics alone is a difficult challenge, as these classes by definition lack any known marker. Despite a large pool of available sequence data, this process still mostly depends on traditional high-throughput screening (HTS).

Although many compounds have been found with high-throughput screening, not all BGCs are activated under laboratory conditions and have therefore likely been missed [186]. This means that a large part of the chemical space of natural products is still to be discovered. Especially for streptomycetes, the number of BGCs far exceeds the natural products that have so far been identified. Growing a strain under many different conditions followed by extensive metabolic profiling, the so-called One Strain MAny Compounds (OSMAC) approach, is a good way to find natural products produced by a single strain [187, 188]. Eliciting strategies are complementary to OSMAC, aiming to mimick the ecological growth conditions of the producer strain and hence the activation of cryptic compounds [33, 189, 190]. In contrast to these general methods, BGC-specific methods study the activation of a single BGC of interest, or even force their expression through engineered promoters and heterologous expression [33, 34].

In this study, we characterize a RiPP family previously discovered by decRiPPter (Chapter 2). DecRiPPter identifies RiPP BGCs with a Support Vector Machine (SVM) classifier that identifies RiPP precursors. It is not limited by biosynthetic domains, and could therefore identify new RiPP subclasses. The RiPP family studied here is prevalent in streptomycetes, with one representative BGC present for every ten *Streptomyces* genomes analyzed. A BGC from *S.*

pristinaespiralis from this family is silent under the growth condition tested, but can be activated by placing a nearby regulator behind a strong promoter. The BGC specifies a novel lanthipeptide, called pristinin A3. Since the BGC lacks any homologs of the lanthionine-forming modifying enzymes, a new route must be required for their biosynthesis, meaning pristinin is a class V lanthipeptide. Based on similarities with enzymes encoded in other BGCs, two gene products, called SprPT and SprH3 are proposed as candidates for their biosynthesis. We further show that their encoding genes are found in a wide variety of different contexts, meaning that they could be used as a new handle for RiPP genome mining. Our work not only validates decRiPPter's capabilities to detect novel RiPP subclasses, but also provides new genome mining rules for the expansion of one the best-studied RiPP subclasses.

Results and Discussion

Discovery of a novel family of lanthipeptides

In Chapter 3, we described the applicability of decRiPPter for the mining of *Streptomyces* genomes for RiPP BGCs. To validate the capacity of decRiPPter to find novel RiPP subclasses, we set out to experimentally characterize one of the candidate families (Chapter 3, Figure 4, 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 (Text S1). 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 [43].

One of the gene clusters belonging to this candidate family was identified in *Streptomyces pristinaespiralis* ATCC 25468 (Figure 1A; Table 1). *S. pristinaespiralis* is known for the production of pristinamycin, and was selected for experimental work since the strain was readily available and genetically tractable [191, 192]. 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 (S1 Text). 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 (Fig 1A). Occurrence of two distinct genes for precursors within a single RiPP BGC is typical of two-component lanthipeptides [193].

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

Table 1. Annotation of the *Streptomyces pristinaespiralis* RiPP (*spr*) gene cluster.

Gene name	NCBI Genbank Accession	NCBI Annotation of gene product	Protein domains	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 PF05193	RiPP maturation protease
sprF1	ALC22065.1	Flavoprotein	PF01636	Cysteine decarboxylation
sprF2	ALC22066.1	Flavoprotein	PF02441	Cysteine decarboxylation
sprOR	ALC22067.1	5,10-methylene tetrahydromethanopterin reductase	PF00291	Reduction of dehydroalanine and dehydrobutyric acid
sprT1	ALC22068.1	ABC transporter ATP-binding protein	PF00005 PF00664	Transport
sprT2	ALC22069.1	ABC transporter	PF12698	Transport
sprT3	ALC22070.1	ABC transporter ATP-binding protein	PF00005 PF13732	Transport
sprMe	ALC22071.1	carminomycin 4-O-methyltransferase	PF00891	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	Dehydration/cyclization
sprPT	ALC22076.1	hypothetical protein	PF01636	Dehydration/cyclization
sprX	ALC22077.1	hypothetical protein		Unknown

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) [194]. 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. Genes encoding cysteine-decarboxylating enzymes are present in non-RiPP gene clusters (Table S3) and are also associated with other metabolic pathways [195]. In theory, though, this BGC could have been detected using a bait-based approach using these genes as queries.

A more detailed comparison with the gene clusters in MIBiG [196] 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 [197]. Thioviridamide-like compounds are primarily known for their thioamide residues, for which a TfuA-associated YcaO is thought to be responsible [52, 95]. 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 [198]. Both flanking regions were cloned into this vector, creating the vector pAK3. Subsequent homologous recombination resulted in a strain where the gene cluster was replaced by the *aac3(IV)* gene, called *spr*::apra (Materials and Methods). 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 [199]. 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 (Fig 1B). Aside from the regulator itself, six out of the sixteen other proteins were detected in

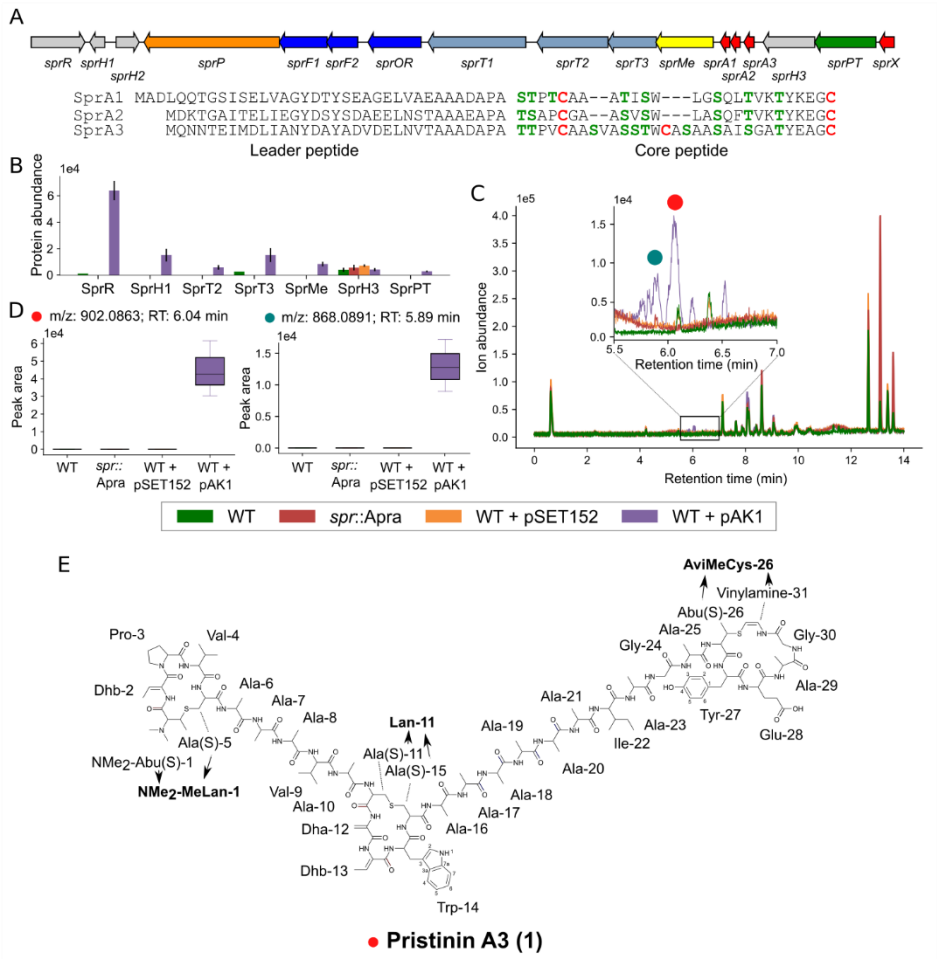


Figure 1. The pristinin BGC (*spr*) of *S. pristinaespiralis* produces a highly modified RiPP. A) The *spr* gene cluster encodes three putative RiPP precursors, three transporters, a peptidase and an assortment of modifying enzymes (see Table 1). Alignment of the predicted precursor peptides is given below. B) Protein abundance of the products of the *spr* gene cluster in *S. pristinaespiralis* ATCC 25468 and its derivatives. Strains were grown in NMMP and samples were taken after 7 days. Enhanced expression of the regulator (from construct pAK1) resulted in the partial activation of the gene cluster. Proteins that could not be detected are not illustrated. C) Overlay chromatogram of crude extracts from strains grown under the same conditions as under B), samples after 7 days. Several peaks were detected in the extract from the strain with expression construct pAK1 between 7 and 8 minutes. D) Boxplot of two peaks detected only in the strain with pAK1. The two masses could be related to two of the three precursors peptides. E) 2D structure of pristinin A3 (**1**), derived from the SprA3 precursor. The compound has a mass of 2703.235 Da.

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 silent BGC under the conditions tested.

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 adsorb 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 (Figure 1C), 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 S1). The isotope patterns of these peaks showed that six of the identified ions were triply charged. Careful analysis of adduct ions and looking for mass increases consistent with Na- or K-addition, led to the conclusion that these peaks corresponded to the $[M+3H]^+$ adduct, suggesting monoisotopic masses in the range of 2,604.273 and 2,754.242 Da. The highest signal came from the compound with a 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 S4). We therefore reasoned that this mass was the product of one of the precursor peptides, while others were incompletely processed peptides. Another mass of 2,601.2433 could not be directly linked to the mass of 2,703.245. This mass was nevertheless only detected in extracts of the strain harboring pAK1 (Figure 1D), suggesting it is the product of another precursor peptide, although whether or not it is the final product remains unclear.

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 [200], creating strain Δ*spr* (Materials and Methods). The expression construct pAK1 and an empty pSET152 vector were transformed to the *spr* null mutant. When these transformants were grown under the same conditions, the aforementioned peaks were not detected, further suggesting that they were products of this gene cluster (Figure 2A).

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 [201]. 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 2B). 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. Many more masses were detected, most of which could be related to one of these two masses, suggesting these are the final products, related to two distinct precursors (Figure 2CD, Table S4 and Table S5).

We then performed MS/MS analysis of the extracts of the pAK2 transformants to identify the metabolites. Building on the hypothesis that the abundant peaks corresponded to the modified SprA1-A3 peptides, we used their peptide sequences to map the fragments. The fragmentation pattern of the peak with a mass of 2,703.245 Da could indeed be assigned to the sprA3 precursor sequence, but only when several mass adjustments of -16 Da, -18 Da, +28 Da and -46 Da were applied (Figure S2A, Table S6). Similarly, fragments for the mass of 2,553.260 could be matched to the SprA2 precursor sequence considering the same mass adjustments (Figure S2B, Table S7). The compounds were named the pristinins, and the individual compounds were named after their precursors (pristinin A3 and pristinin A2, respectively).

All of the -18 and -16 Da adjustments were predicted on serine and threonine fragments. These mass differences are typical of dehydration (-18 Da) of the residues to dehydroalanine (Dha) and dehydrobutyrine (Dhb). Reduction of these dehydrated amino acids (+2 Da) would then give rise to alanine and

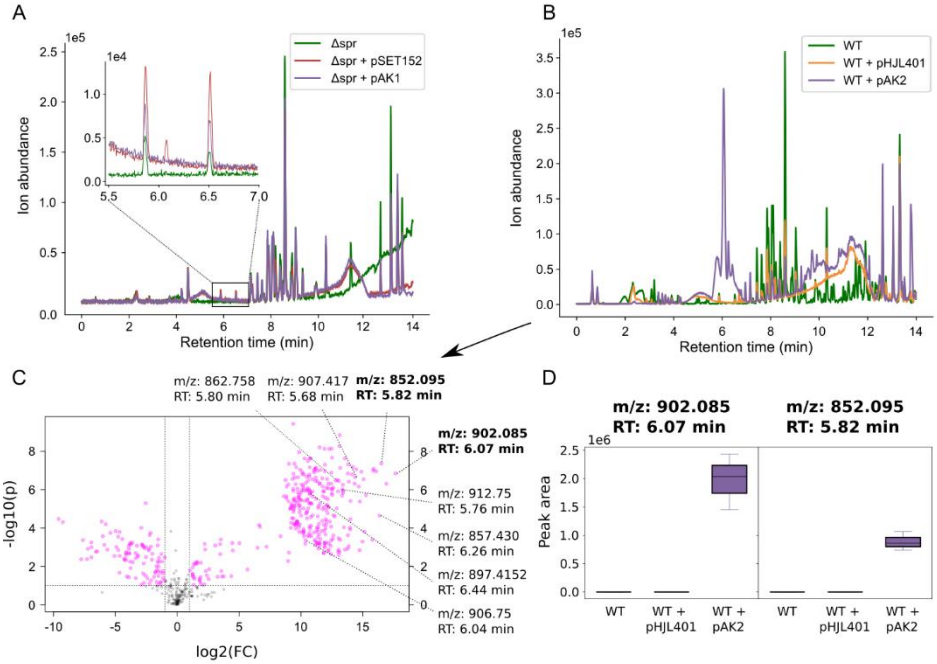


Figure 2. Chromatograms comparing the extracted compounds in knockout strains and highly producing strains. A) Strains lacking the *spr* gene cluster are unable to produce the extracted products, even when transformed with pAK1. B) Chromatogram of methanol extracts made from *S. pristinaespiralis* harboring no vector (WT) an empty pHJL401 vector (pHJL401), or pAK2 (pHJL401 with *sprR* behind p_{gap}). A large peak can be seen in the extracts of strains harboring pAK2, not seen in extracts of the other strains. C) Volcano plot comparing extracts of the strain containing pAK2 with the strain containing pHJL401. Peaks in pink had p-value ≥ 0.1 and a fold-change of ≥ 2. A large collection of peaks can be identified with log2(fold-change) ≥ 10. The two largest peaks (bold) corresponding to different monoisotopic masses could be related to the SprA2 and SprA3 precursors by MS/MS (S11 Fig). Many of the other masses eluted at comparable times, and had masses that were close to the two major peaks, suggesting they were derived from them. Clear mass differences could be identified for some of the identified masses (Table S5). Whether the largest peaks indeed correspond to the final product remains to be determined. D) Extracted ion chromatograms of the two major peaks identified from the volcano plot. The two masses were only detected in the strain harboring pAK2.

butyric acid residues, a modification that has been reported for lanthipeptides [202]. A modification of +28 Da suggests a dual methylation among the five N-terminal residues, which is consistent with the methyltransferase SprMe that is encoded by the *spr* gene cluster. The loss of -46 Da could be attributed to the C-terminal cysteine. This mass difference correlates to oxidative decarboxylation,

which is consistent with the cysteine decarboxylase SprF2 that is encoded by the cluster. The loss of -18 Da in a threonine residue close to the modified cysteine suggests the presence of an AviMeCys group at the C-terminal end of the peptide. The lack of fragments for the residues T^{-18}YEAGC^{-46} in the fragments pristinin A3 further supports the presence of an AviMeCys-containing C-terminal ring.

Surprisingly, no fragments were found of the residues $S^{-18}S^{-18}T^{-18}$WC in the center of pristinin A3, or for the N-terminal $[T^{-18}T^{-18}PVC]^{+28}$ region. Considering the other modifications typical of lanthipeptides, and the likely presence of a thioether crosslink in the AviMeCys group, we hypothesized the presence of thioether crosslinks between the Dhbs 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 [203]. In agreement with our hypothesis, treatment with iodoacetamide did not affect the observed masses, despite the presence of three cysteines in the peptide (Figure S7).

NMR confirms the presence of lanthionine bridges in the pristinins

To further ascertain the presence of the proposed modifications, we purified the peak corresponding to pristinin A3. Since the products were not detected when cultures were grown in 500 mL cultures, we grew 100 × 20mL cultures (2L total) of a transformant harboring the expression plasmid pAK2. The culture was then extracted and the extract was subjected to a series of chromatographic fractionations, which resulted in the purification of pristinin A3 (**1**) (Materials and Methods). The purified compound was dissolved in deuterated DMSO (DMSO-d_6) for NMR analysis. Extensive purification allowed us to purify 1.1 mg of the compound. While the amount of material meant that the NMR signal was low, we could derive many key features of the peptide in the ^{1}H NMR spectrum (Figure S3, Figure S4A). The NH signals in the ^{1}H NMR spectrum were very broad using DMSO-d_6 as solvent. We therefore changed to CD_3CN:H_2O 9:1 as the solvent, which showed very good NH signals for the recently identified similar peptide cacaoidin [204]. Indeed, sharper peaks and better HMBC correlations could be observed (Figure S5 and S6). Re-analysis of pristinin A3 (**1**) using LC-MS showed that the compound was partially oxidized, i.e. a mixture of compounds

was analyzed in the NMR run using $CD_3CN:H_2O$ as a solvent (Table S11). MS/MS fragmentation suggested that the oxidation occurred consistently in the center and N-terminal ring structures (Table S12).

Combined analysis of the 2D COSY, TOCSY, HSQC, HMBC and NOESY NMR spectra obtained in DMSO-d_6 (Figure S4, Table S9) supported the proposed structure of pristinin A3 (**1**) (Figure 1E). In the 2D spectra several spin systems were identified, which were consistent with the amino acid sequence of SprA3 and the MS/MS fragmentation data (Figure S3). These amino acid residues were 2 Val, 2 Gly, 1 Pro, 1 Trp, 1 Ile, 1 Tyr, 1 Glu and multiple mostly overlapping Ala. Additionally, we identified spin systems consistent with the proposed modified amino acid residues. These were 2 Dhbs, 2 β-thioalanines (Ala(S)), 1 Dha, 1 β-thioaminobutyric acid (Abu(S)), and 1 aminovinyl group. Due to weak signals, we could not use the HSQC-TOCSY spectra to further support the identified residues. There was no clear evidence in the NMR spectra of the presence of Thr or Ser amino acid residues, which corroborated the hypothesis that all the Thr and Ser residues identified in SprA3 had been modified.

We next sought evidence for the connectivity of the identified amino acids. The connectivity of the amino acid residues through NMR could be readily established through the Hα-NH (i, i+1), Hβ-NH (i, i+1), and NH-NH (i, i+1) NOESY correlations. Based on this, the AviMeCys-containing C-terminal ring and its extension up to Ala-21 could be unambiguously established to be in accordance with the proposed structure through the MS/MS data (Figure S3). Importantly, the same structural fragment could be clearly observed in the sample analyzed in $CD_3CN:H_2O$ 9:1, supporting the observation from the MS/MS data that the oxidation of pristinin A3 (**1**) was in the rings closer to the N-terminus. The NMR data in $CD_3CN:H_2O$ confirmed the sequence of Ala-25 up to Glu-28, because some of the Hα and NH signals for these residues, that were overlapping in DMSO-d_6, were well separated in $CD_3CN:H_2O$ (Figure S4, S5 and S6, Table S10). Additionally, HMBC correlations could be observed to the carboxyl group of Glu in $CD_3CN:H_2O$. The NOESY correlations in DMSO-d_6 further unambiguously confirmed the peptide sequence observed in MS/MS for Dhb-2 to Ala-10, Dha-12 to Dhb-13, and Trp14 to Ala-16 (Table S9). The sequence of Ala-17 to Ala-20 had overlapping Hα and NH signals. However, the correlation pattern observed

Table 2. Summary of the different methods used to identify the amino acid residues of pristinin A3. Symbols indicate whether residues and their connectivity were confirmed (+), partly confirmed (±), or not confirmed (-).

Amino acid residues	Gene seq.	HRMS/MS	NMR	Acid hydrolysis[a]	Amino acid residues	Gene seq.	HRMS/MS	NMR	Acid hydrolysis[a]
NMe_2-MeLan-1*	-	+	±[b]	+	Ala-17	-	+	+	+
Dhb-2	-	+	+	-	Ala-18	+	+	+	+
Pro-3	+	+	+	+	Ala-19	+	+	+	+
Val-4	+	+	+	+	Ala-20	-	+	+	+
Ala-6	+	+	+	+	Ala-21	+	+	+	+
Ala-7	+	+	+	+	Ile-22	+	+	+	+
Ala-8	-	+	+	+	Ala-23	-	+	+	+
Val-9	+	+	+	+	Gly-24	+	+	+	-
Ala-10	+	+	+	+	Ala-25	+	+	+	+
Lan-11*	-	+	±[c]	+	AviMeCys-26*	-	+	±	-
Dha-12	-	+	+	-	Tyr-27	+	+	+	+
Dhb-13	-	+	+	-	Glu-28	+	+	+	+
Trp-14	+	+	±[d]	-	Ala-29	+	+	+	+
Ala-16	+	+	+	+	Gly-30	+	+	+	-

[a] Acid hydrolysis only confirms the amino residues, but not their connectivity.

[b] Only Ala(S)-5 could be observed in NMR;

[c] Only Ala(S)-15 could be observed in NMR

[d] Trp-14 and its connectivity to Ala(S)-15 could be confirmed by NMR, but its connectivity to Dhb-13 could not be confirmed.

* NMe2-MeLan-1 = NMe_2-Abu(S)-1 + Ala(S)-5; Lan-11 = Ala(S)-11 + Ala(S)-15; AviMeCys-26 = Abu(S)-26 + Vinylamine-31

and the peak integration support a series of alanine residues to be the connection between Ala-16 and Ala-21, as was also indicated by the MS/MS data.

It was not possible to establish the connection between Dhb-13 and Trp-14 using NMR. At the same time, a Dha–Dhb sequence could be clearly established using NMR. The fact that Dha and Dhb are the products of modified Ser and Thr residues, respectively, and the fact that the only Ser–Thr sequence in the SprA3 precursor lies before Trp, inevitably means that the observed Dha–Dhb structural fragment is connected to Trp-14 and positioned as Dha12 and

Dhb-13. Finally, the thioether crosslinks of the proposed N-terminal and center ring structures could not be completely resolved based on NMR data alone. This is because the ^{1}H NMR resonance for a CH/CH_2 group attached to a sulfur atom should be around δ_H 3 ppm, which is close to the area where the water signal in DMSO-d_6 (δ_H 3.3 ppm) is suppressed in the NMR experiments. Water suppression greatly affected the smaller signals around this area. Nevertheless, we managed to establish and position Ala(S)-5 and Ala(S)-15, both of which have to be part of a thioether bond as proven through the IAA labelling experiment discussed earlier. This left only one residue in each of the two additional rings observed in MS/MS, which was not accounted for by NMR (Figure S3). Based on this, an NMe_2-Abu(S)-1 and Ala(S)-11 could be proposed to form thioether bridges with Ala(S)-5 and Ala(S)-15, respectively, resulting in the formation of *N*,*N*-dimethyl-β-methyllanthionine (NMe_2-MeLan) and lanthionine (Lan) residues, respectively. As a further evidence, 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 [205]. The resulting mixture of amino acids was analyzed using LC-HRMS and was indeed found to contain peaks with exact masses corresponding to NMe_2-MeLan and Lan (Table S8). Thus, the primary sequence of the peptide, the MS/MS fragmentation data, the NMR data, acid hydrolysis and labelling experiments (Table 2) allowed us to elucidate the 2D structure of pristinin A3 (**1**; Figure 1E).

The RiPPs characterized here contain a number of modifications that have previously been identified in other RiPPs. A recent study, which appeared around the time of submission for this paper, describes a RiPP found by activity-based screening, called cacaoidin, that has many of the same modifications [204], and is additionally glycosylated. The serines converted to alanines in cacaoidin were all D-alanines. It therefore seems probable that the converted serines in pristinin A3 (**1**) were also converted to D-alanines, which could be determined by further chemical analyses. BLAST analysis shows that the genes of the cacaoidin BGC show low similarity to those in the *spr* BGC, and the precursor genes do not seem directly related. However, the same Pfam domains are found in both BGCs, indicating that both BGCs belong to the same RiPP class. The authors describing cacaoidin remark that these modifications were found previously in linaridins and lanthipeptides, and therefore named this class the

lanthidins. While some enzymes encoded by the BGCs of this RiPP class indeed show low similarity to enzymes involved in the biosynthesis of characterized RiPPs, the combination of modifications makes it a novel RiPP subclass that was not previously detected by other RiPP genome mining tools. Overall, these findings further support the potential of decRiPPter to identify novel RiPP BGCs.

SprH3/SprPT are candidates for the enzymes that install lanthionines

Taken together, we have shown that pristinin A3 contained a number of posttranslational modifications that are typical of lanthipeptides. 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. No homologs of known lanthionine-forming enzymes 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 [94, 206]. Another RiPP subclass with an AviCys moiety is the linaridin subclass. Dehydration of the required serine is thought to be catalyzed by LinH or LinL, neither of which show similarity to the proteins encoded by the thioviridamide BGC or the *spr* BGC. Of note, the cacaoidin BGC also encoded two proteins with the same domains as SprH3 and SprPT (i.e. PF01636 and PF17914). Since the thioviridamide, cacaodin and *spr* gene clusters share a common modification for which the enzyme is unknown, we hypothesize that SprH3 and SprPT carry out the dehydration and cyclization reactions and are therefore likely involved in the maturation of many different RiPPs, with dehydrated residues, AviCys moieties, or thioether bridges. In the latter group, these enzymes candidate as core modifying enzymes of a new lanthipeptide subclass, which we designated lanthipeptide class V.

To find experimental support for the hypothesis that SprH3 and SprPT are the sought-after modifying enzymes, we replaced the gene pair *sprH3*/*PT* with an apramycin resistance cassette (aac3(IV); Materials and Methods). To this end, the flanking regions were amplified with PCR, and placed into the shuttle vector pWHM3. An apramycin resistance was placed between the flanks

through restriction and ligation. The resulting vector pAK8 was transformed to *S. pristinaespiralis* and the exchange of genes was confirmed by PCR. The strain was named *sprH3PT*::Ap. This strain was then transformed with the pAK2 vector in order to activate the *spr* BGC, or with an empty pHJL401 vector as a control. The three resulting strains were grown and extracted using the same conditions as described above. Despite this, no masses were detected with HPLC that relate 6to the products of the *spr* BGC (Figure S8A). The genetic modifications made to the BGC therefore disrupt the biosynthesis of the pristinin A2 and A3 precursors.

The extracts obtained from the *sprH3PT*::Ap strain and derivatives suggest that the two products of removed gene pair are indeed involved in the biosynthesis of pristinin A2 and A3. An alternative explanation is that the genetic modifications themselves disrupt the transcription and/or translation of the *spr* BGC, which appears to be in an operon-like structure. We aimed to rule out the latter explanation by providing an additional copy of the removed genes to the *sprH3PT*::Ap strain. The gene pair was amplified with PCR, either with or without the native promoter. When no native promoter was amplified, the amplified gene pair was placed behind the upstream region of XNR_3799, a strong promoter for streptomycetes amplified from *S. lividans* (Zhang, L., personal communication). The two different gene pair regions were then placed either in pHJL401, to create the control constructs pAK4 (native promoter) and pAK6 (XNR promoter), or the pAK2 vector, creating the constructs pAK5 (native promoter) and pAK7 (XNR promoter). A t0 terminator was placed between the DNA fragment harboring the *sprH3/PT* gene pair and the fragment harboring the *sprR* gene in the pAK5 and pAK7 vectors to prevent transcriptional read-through.

The resulting four vectors were transformed to the *sprH3PT*::Ap strain, cultured and extracted as described above. Under these conditions, none of the masses related to the *spr* BGC were detected in the extracts (Figure S8B). This meant that the complementation vectors either do not express the gene pair, or disruption of the *spr* BGC extends beyond the targeted gene pair to also affect the rest of the BGC. The removed fragment is evidently important for the production of the mature RiPP product. However, whether this is due the presence of promoter regions within that fragment that regulate the expression of the BGC, or due to the relevance of the encoded products could not be

determined from these experiments. Further experiments, such as *in vitro* enzymology experiments are still required to confirm the function of the SprH3/PT proteins.

The *sprH3/sprPT* gene pair is present in a wide variety of predicted RiPP BGCs

Lanthipeptide core modifying enzymes catalyze the most prominent reaction in lanthipeptide maturation, and as such, are present in many different genetic contexts [68, 73]. To find support for the proposed role of the gene products we studied the distribution of the *SprH3*/*PT* gene pair across *Streptomyces* genomes analyzed by decRiPPter. Using CORASON [185] with the s*prPT* gene as a query yielded 195 homologs in various gene clusters (Figure 3, Materials and Methods). 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. A total of 391 orthologs of the gene pair were found outside *Streptomyces*, particularly in Actinobacteria (219) and Firmicutes (161; Figure S9). Distantly similar homologs of the gene pair were also identified in Cyanobacteria, Planctomycetes 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 (Table 1), 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 [39]. 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. This domain is often associated with small prepeptides, such as the precursor peptides of methanobactin [207]. 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 3, right side). While some gene clusters were mostly homologous to the *spr* gene

cluster (Figure 3, group g-h), others shared only a few genes (groups a and d), and some only shared the gene pair itself (Figure 3, b, c and e; Table 3). 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. We would like to emphasize, however, that not all of these BGCs necessarily specify lanthipeptides. Assuming that the proposed role for the products of *sprH3/PT* in dehydration of serine and threonine residues is correct, these modifications could also lead to AviCys moieties, such as in thioviridamide-like products, or simply remain dehydrated residues. Further genetic and biochemical elucidation of the role of these enzymes is necessary to completely determine the scope of their reactions.

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 S13).

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

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

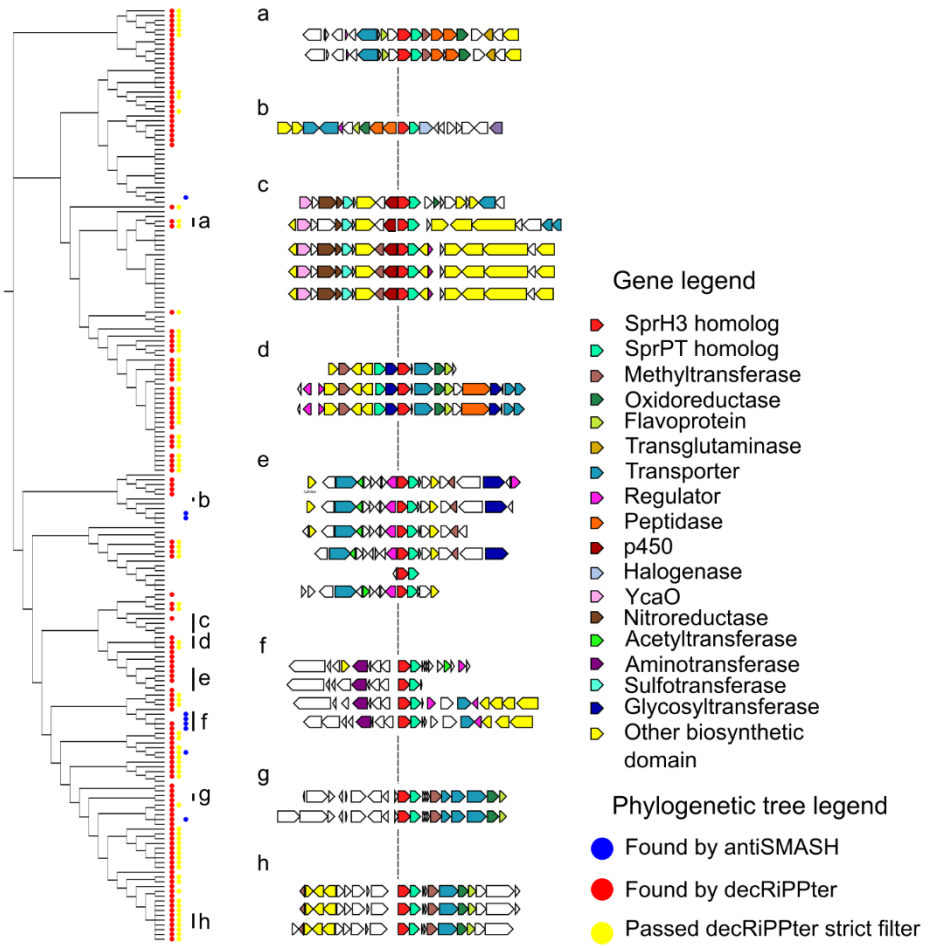


Figure 3. Orthologs of *sprPT* and *sprH3* cooccur in a wide variety of genetic contexts. (Left side) Phylogenetic tree of gene clusters containing homologs of *sprPT* and *sprH3*, visualized by CORASON. A red dot indicates that the genes were present in a gene cluster found by decRiPPter, a yellow dot that it passed the strict filter (Chapter 2). 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.

Another modification dependent on the presence of dehydrated serine and threonine residues is the conversion of these to alanine and butyric acid, respectively. This conversion is catalyzed either by a zinc-dependent dehydrogenase ($LanJ_A$, also known as as LtnJ) or an NAD(P)H-dependent FMN reductase family enzyme ($LanJ_B$, also known as CrnJ) in lanthipeptides [202]. Outside *Streptomyces*, the genomic surroundings of the *sprPT/sprH3* gene pair occasionally contained homologs of the *lanj*$_A$ gene (40/391; 10.1%). An example of such a BGC is that of pediocin A, a known antimicrobial compound of which the structure has yet to be resolved [208]. These gene associations further imply that the SprH3/SprPT gene products apply the canonical dehydration reactions.

A similar modification was observed for pristinin A2 and A3, despite that no homologs of the genes encoding $LanJ_A$ or $LanJ_B$ were identified within the *spr* gene cluster. However, *sprOR* encodes a putative oxidoreductase, and thus is a candidate 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 S13). 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 [209]. 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 S10). A recent paper describes a BGC with a gene also encoding a luciferase-like monooxygenase. The product of this BGC contains serine residues that are converted to alanine residues [73], further suggesting that this enzyme applies this modification.

Interestingly, many *sprOR*, *sprF1* and *sprF2* homologs were found not present in either a lanthipeptide BGC or close to the *sprPT/H3* gene pair. These three genes products all require a dehydrated serine or threonine residue to carry out their reaction. The presence of these homologs therefore provides a promising lead for core-dependent genome mining. Assuming the products of the homologs still carry out the same reaction, investigation of these homologs could lead to the discovery of even more lanthipeptide core modifying enzymes.

Conclusions and final perspectives

Most RiPP genome mining strategies expand previously characterized RiPP subclasses. These efforts can lead to novel natural products when new RiPP precursors are identified in conjunction with previously characterized modification machinery. However, the detection of completely novel RiPP subclasses remains a more challenging ordeal, and currently used genome mining tools can only identify these if there are similarities between the known and the novel RiPP subclass.

In this work, we have characterized a candidate novel RiPP subclass, whose BGC was identified with decRiPPter. The product of one of the gene clusters associated with this candidate class was characterized as the first member of a new class of lanthipeptides (termed 'class V'). BGCs of this class were not detected by any other RiPP genome mining tool. Variants of this gene cluster are widespread across *Streptomyces* species, further expanding one of the best-studied RiPP subclasses. The fact that no less than five different sets of lanthionine-forming enzymes have been reported highlights the importance of this crosslink. Furthermore, this subclass is one of the few RiPP subclasses that has been prioritized purely through the use of bioinformatics, showcasing the potential of these methods for natural product genome mining when properly applied. Since no fewer than 42 different candidate families were discovered in *Streptomyces* alone, the potential of decRiPPter to further expand the list of RiPP subclasses is an exciting prospect.

In addition, two core genes were proposed based on their similarity to genes associated with other RiPP subclasses, which share a common modification. These genes were used to expand the family by finding additional homologs in Actinobacteria and Firmicutes. These homologs could be present in many different genetic contexts, suggesting that a wide variety of new RiPPs and RiPP modifications could be identified among these BGCs. 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.

Materials and Methods

Experimental procedures

Bacterial strain and growth conditions

Streptomyces pristinaespiralis ATCC 25468 was purchased from DSMZ (DSM number 40338). Media components were purchased from Thermo Fisher Scientific, Sigm-Aldrich or Duchefa Biochemie. For strain cultivation on solid media, *Streptomyces* spores were spread on mannitol soya flour agar (SFM; 20 g/L Agar, 20 g/L mannitol, 20 g/L soya flour, supplemented with tap water) prepared as described previously [210], and incubated at 30°C. Spores were harvested after 4-7 days of growth when the strain started to produce a grey pigment, by adding water directly to the plate and releasing the spores with a cotton swab. Spores were centrifuged and stored in 20% glycerol.

For cultivation in liquid media, 20-50 µL of a dense spore stock was inoculated into 100 mL shake flasks with coiled coils containing 20 mL of the medium of interest. For extractions, NMMP was used (0.60 mg/L $MgSO_4$, 5 mg/L NH_4SO_4, 5 g/L Bacto casaminoacids, 1 mL trace elements (1 g/L $ZnSO_4.7H_2O$, 1 g/L $FeSO_4.7H_2O$, 1 g/L $MnCl_2.4H_2O$, 1 g/L $CaCl_2$, anhydrous)), while for genomic DNA isolation, a 1:1 mixture of TSBS: YEME with 0.5% glycine and 5 mM $MgCl_2$ was used (TSBS: 30 g/L Bacto Tryptic Soy Broth, 100 g/L sucrose; YEME: Bacto Yeast Extract: 3 g/L, Bacto Peptone 5 g/L, Bacto Malt Extract 3 g/L, glucose 10 g/L, sucrose 340 g/L).

E. coli strains JM109 and ET8 were used for general cloning purposes and demethylation, respectively. Strains were cultivated in liquid LB and on LB-agar plates at 37°C.

Molecular biology

All materials and primers were purchased from Sigma-Aldrich or Thermo Fisher Scientific unless stated otherwise. Restriction enzymes and T4 ligase were purchased from NEB. Restriction and ligation protocols were followed as per manufacturer's description. For amplification of DNA fragments with PCR, Pfu polymerase was used. Primers were designed with T_m of the annealing region roughly equal to 60°C. Standard PCR protocols consisted of 30 cycles (45 second DNA melting @ 95 °C, 45 second primer annealing @55°C-65°C, 60s-180s primer elongation @ 72°C), but PCR protocols were optimized where necessary.

Deletion mutants were created by replacing the gene cluster or targeted genes with an *aac(3)IV* apramycin resistance cassette via homologous recombination, as described [211]. For the deletion of the entire gene cluster, the -1507/-39 and +135/+1641 regions upstream and downstream of the cluster were amplified by PCR with the spr_LF_F/spr_LF_R and spr_RF_F/spr_RF_R primer pairs (table S1) respectively, and inserted into the pWHM3-oriT vector (Table S2) into the EcoRI/HindIII sites. The *aac(3)IV* apramycin resistance cassette was inserted into the XbaI site, creating the vector pAK3. pAK3 was transformed to *E.coli ET8* for DNA demethylation, purified, and transformed to *S. pristinaespiralis* by protoplast transformation. Transformation mixtures were plated out on R5, prepared as described earlier [210]. After 14-18 hours, the plates were overlaid with 1.2 mL H_2O containing 10 µg thiostrepton and 25 µg apramycin. Three colonies were picked after 4 days of growth and spread onto SFM plates without added antibiotic to allow for homologous recombination. Colonies containing the correct phenotype (apramycin-resistant, thiostrepton-sensitive) were picked and the homologous recombination was confirmed by PCR, using the spr_del_check_F/spr_del_check_R primer pair.

For the deletion of the gene pair *sprH3/sprPT*, the primers sprH3PT_LF_F/ sprH3PT_LF_R and sprH3PT_RF_F/ sprH3PT_RF_R were used to amplify the -1430/-54 and +2/+1483 flanking regions. These resulting fragments were used to create the pAK8 vector, and the mutants were created as above. Confirmation of the mutants was done with PCR using the primer pair sprH3PT_check_F/sprH3PT_check_R.

Removal of the apramycin cassette was done by transforming the pUWLCRE shuttle vector to the mutant strain as described above. Three colonies were picked and grown on SFM without antibiotics. The antibiotic resistance phenotype was monitored by growing the spores on plates with the relevant antibiotics. Strains that were apramycin-resistant, thiostrepton-sensitive were picked as candidate full deletion mutants, which was confirmed by PCR using the same checking primers as used for the apramycin resistance mutants.

Constructs for the overexpression of the *sprR* regulator were constructed as follows: the *sprR* gene was amplified from the genomic DNA of *S. pristinaespiralis* using the sprR_F/sprR_R primer pair, and placed into the EcoRI/XbaI site of the pSET152 vector. The -0/-457 upstream region of glyceraldehyde 3-phosphate dehydrogenase amplified from the genome of *S. coelicolor*, was obtained from previous studies [212, 213] and inserted into the EcoRI site and the engineered NdeI site, placing it directly upstream of the *sprR* gene. To create vector pAK2, the entire region between the EcoRI and XbaI sites was excised and inserted into the pHJL401 vector.

To make the complementation constructs for the *sprH3/PT* deletion strain, we aimed at placing a DNA fragment containing both genes (-0/+0), preceded by a promoter, in the XbaI/HindIII site behind the regulator in the pAK2 vector. An additional terminator sequence was placed between the the *sprR* gene and the amplified fragment, to prevent transcriptional read-through. To this end, a single fragment containing both genes with the preceding 519 bp was amplified with the sprH3PT_compl_F_t0_prom/sprH3PT_R primer pair, and placed in either pHJL401 (creating pAK4) or in the pAK2 vector (creating pAK5). The SprH3PT_compl_F_t0_prom primer contains a t0 terminator sequence. Overexpression constructs with the XNR_3799 promoter were created by amplifying the -695/+3 upstream region of XNR_3799 of *S. lividans* with a preceding t0 terminator sequence from an in-house plasmid on which these two sequences were adjacent, using the XNR_t0_F and XNR_t0_R primers. Using the XbaI/NdeI restriction site, this fragment was placed behind the *sprH3/PT* genes, amplified without their native promoter using the sprH3PT_compl_F/ sprH3PT_compl_R primer pair. The resulting fragment containing the t0 terminator, the XNR3170 promoter and the *sprH3*/*PT* gene pair was placed on pHJL401 (creating pAK6) and on pAK2 (creating pAK7).

Extractions

Strains were cultured in 100 mL shake flasks containing 20 mL NMMP, with coiled coils at 30°C for 7 days. 20 µg/mL thiostrepton was added to cultivate strains containing pHJL401. Mycelium was collected by centrifugation, washed twice with sterile MiliQ water and extracted with 5 mL methanol by shaking overnight at 4°C. The methanol was collected and centrifuged at 4°C to clear it of cellular debris and precipitates. The crude extracts were dried and weighed, and dissolved in methanol at a concentration of 1 mg/mL for further analysis.

Peptide purification

For large-scale extraction, 2L NMMP prepared as above was inoculated with 2.5 mL of a dense spore stock *S. pristinaespiralis* with pAK3, and split over one hundred 100 mL shake flasks. The

cultures were grown for 14 days, pooled together and extracted with an equivalent volume of butanol. The butanol extracted was then evaporated *in vacuo* to yield 1.7g of crude extract. The resulting crude extract was adsorbed on silica gel 60 (40–60 µm, Sigma Aldrich), and dry loaded on a VLC column (3 × 30 cm) packed with the same material. The column was eluted with 200 mL fractions of a gradient comprised of (v/v): hexane, hexane–EtAc (1:1), EtAc, EtAc-MeOH (3:1), EtAc-MeOH (1:1), EtAc–MeOH (1:3), and finally MeOH. The fractions containing the compound of interest were pooled, concentrated and further purified using Waters preparative HPLC system comprised of 1525 pump, 2707 autosampler, and 2998 PDA detector. The pooled fraction (112.9 mg) was injected into a SunFire C_{18} column (10 µm, 100 Å, 19 × 150 mm). The column was run at a flow rate of 12.0 mL/min, using solvent A (0.1% FA in H_2O) and solvent B (0.1% FA in ACN), and a gradient of 30–60% B over 20 min. HPLC purification was monitored at 254 nm, and eventually resulted in compound **1** (1.1 mg).

LCMS analysis

LC-MS/MS acquisition was performed using Shimadzu Nexera X2 UHPLC system, with attached PDA, coupled to Shimadzu 9030 QTOF mass spectrometer, equipped with a standard ESI source unit, in which a calibrant delivery system (CDS) is installed. The dry extracts were dissolved in MeOH to a final concentration of 1 mg/mL, and 2 µL were injected into a Waters Acquity Peptide BEH C_{18} column (1.7 µm, 300 Å, 2.1 × 100 mm). The column was maintained at 40 °C, and run at a flow rate of 0.5 mL/min, using 0.1% formic acid in H_2O as solvent A, and 0.1% formic acid in acetonitrile as solvent B. A gradient was employed for chromatographic separation starting at 5% B for 1 min, then 5 – 85% B for 9 min, 85 – 100% B for 1 min, and finally held at 100% B for 4 min. The column was re-equilibrated to 5% B for 3 min before the next run was started. The LC flow was switched to the waste the first 0.5 min, then to the MS for 13.5 min, then back to the waste to the end of the run. The PDA acquisition was performed in the range 200 – 400 nm, at 4.2 Hz, with 1.2 nm slit width. The flow cell was maintained at 40 °C.

The MS system was tuned using standard NaI solution (Shimadzu). The same solution was used to calibrate the system before starting. System suitability was checked by including a standard sample made of 5 µg/mL thiostrepton; which was analyzed regularly in between the batch of samples. All the samples were analyzed in positive polarity, using data dependent acquisition mode. In this regard, full scan MS spectra (*m/z* 400 – 4000, scan rate 20 Hz) were followed by three data dependent MS/MS spectra (*m/z* 400 – 4000, scan rate 20 Hz) for the three most intense ions per scan. The ions were selected when they reach an intensity threshold of 1000, isolated at the tuning file Q1 resolution, fragmented using collision induced dissociation (CID) with collision energy ramp (CE 10 – 40 eV), and excluded for 0.05 s (one MS scan) before being re-selected for fragmentation. The parameters used for the ESI source were: interface voltage 4 kV, interface temperature 300 °C, nebulizing gas flow 3 L/min, and drying gas flow 10 L/min.

LC-MS based comparative metabolomics

All raw data obtained from LC-MS analysis were converted to mzXML centroid files using Shimadzu LabSolutions Postrun Analysis. The converted files were imported and processed MZmine 2.5.3 [214]. Throughout the analysis, *m/z* tolerance was set to 0.002 *m/z* or 10.0 ppm, RT tolerance was set to 0.05 min, noise level was set to 2.0E2 and minimum absolute intensity was set to 5.0E2 unless specified otherwise. Features were detected (polarity: positive, mass detector: centroid) and their chromatograms were built using the ADAP chromatogram builder [215] (minimum group

size in number of scans: 10; group intensity threshold: 2.0E2). The detected peaks were smoothed (filter width: 9), and the chromatograms were deconvoluted (algorithm: local minimum search; Chromatographic threshold: 90%; search minimum in RT range: 0.05; minimum relative height: 1%; minimum ratio of peak top/edge: 2; peak duration 0.03 – 3.00 min). The detected peaks were deisotoped (maximum charge: 5; representative isotope: lowest *m/z*). Peak lists from different extracts were aligned (weight for RT = weight for *m/z*; compare isotopic pattern with a minimum score of 50%). Missing peaks detected in at least one of the sample were filled with the gap filling algorithm (RT tolerance: 0.1 min). Among the peaks, we identified fragments (maximum fragment peak height: 50%), adducts ($[M+Na]^+$, $[M+K]^+$, $[M+NH_4]$, maximum relative adduct peak height: 3000%) and complexes (Ionization method: $[M+H]^+$, maximum complex height: 50%). Duplicate peaks were filtered. Artifacts caused by detector ringing were removed (*m/z* tolerance: 1.0 *m/z* or 1000.0 ppm) and the results were filtered down to the retention time of interest. The aligned peaks were exported to a MetaboAnalyst file. From here, peaks were additionally filtered to keep only peaks present in all three replicates, using in-house scripts. The resulting peak list was uploaded to MetaboAnalyst [216], log transformed and normalized with Pareto scaling without prior filtering. Missing values were filled with half of the minimum positive value in the original data. Heatmaps and volcano plots were generated using default parameters.

Mass spectrometry-based quantitative proteomics

20 uL of dense spore stocks were inoculated in NMMP and grown for 7 days as described above. 1 mL samples were taken after 2 and 7 days. Mycelium was gathered by centrifugation and washed with disruption buffer (100 mM Tris-HCl, pH 7.6, 0.1 M dithiothreitol). The samples were sonicated for 5 minutes (in cycles off 5s on, 5s off) to disrupt the cell wall, and centrifuged at max speed for 10 minutes to collect the proteins. Proteins were then precipitated using chloroform-methanol [217]. The dried proteins were dissolved in 0.1% RapiGest SF surfactant (Waters) at 95°C. Protein digestion steps were done according to van Rooden et al [218]. After digestion, formic acid was added for complete degradation and removal of RapiGest SF. Peptide solution containing 8 µg peptide was then cleaned and desalted using the STAGETipping technique [219]. Final peptide concentration was adjusted to 40 ng/µl with 3% acetonitrile, 0.5% formic acid solution. 200 ng of digested peptide was injected and analysed by reverse-phase liquid chromatography on a nanoAcquity UPLC system (Waters) equipped with HSS-T3 C18 1.8 µm, 75 µm X 250 mm column (Waters). A gradient from 1% to 40% acetonitrile in 110 min was applied, [Glu^1]-fibrinopeptide B was used as lock mass compound and sampled every 30 s. Online MS/MS analysis was done using Synapt G2-Si HDMS mass spectrometer (Waters) with an $UDMS^E$ method set up as described [218].

Mass spectrum data were generated using ProteinLynx Global SERVER (PLGS, version 3.0.3), with MS^E processing parameters with charge 2 lock mass 785.8426 Da. Reference protein database was downloaded from GenBank with the accession number GCA_001278075.1. The resulting data were imported to ISOQuant [220] for label-free quantification. TOP3 quantification result from ISOQuant was used when further investigating the data.

Iodoacetamide treatment

Reaction mixtures were prepared based on earlier reported studies [203]. 20 µL reaction mixtures containing 0.25 mg/mL purified peptide, 13 mM TCEP, 25 mM IAA and 250 mM HEPES (pH = 8.0) in H_2O were left at room temperature for 1 hour in the dark. Reaction mixtures were cleaned using the STAGETipping technique [219].

Protein hydrolysis

0.2 mg of purified peptide was dissolved in 3 mL 6M HCl and sealed inside a glass ampule, based on earlier studies[221]. The mixture was heated to 110°C for 24 hours. The HCl was removed by repeated drying and dissolving of the peptide with H_2O. The peptide was afterwards dissolved in 50 µL H_2O and analyzed with LCMS as described above.

NMR

NMR data were recorded on Bruker Ascend 850 NMR spectrometer (Bruker BioSpin GmbH), equipped with a 5 mm cryoprobe. The sample was measured in a 3 mm NMR tube through the use of an adapter. All NMR experiments were performed with suppression of the water peak in the solvent.

Data analysis

Genomic context analysis

CORASON [185] was used with the number of flanking genes set to 15, on the *Streptomyces* genomes analyzed with the query of interest. Results were parsed using in-house scripts and compared to decRiPPter output. NCBI BLAST was used to find additional homologs of genes of interest within the clusters, with a cutoff of 30 percent ID similarity.

Supplementary information for Chapter 4

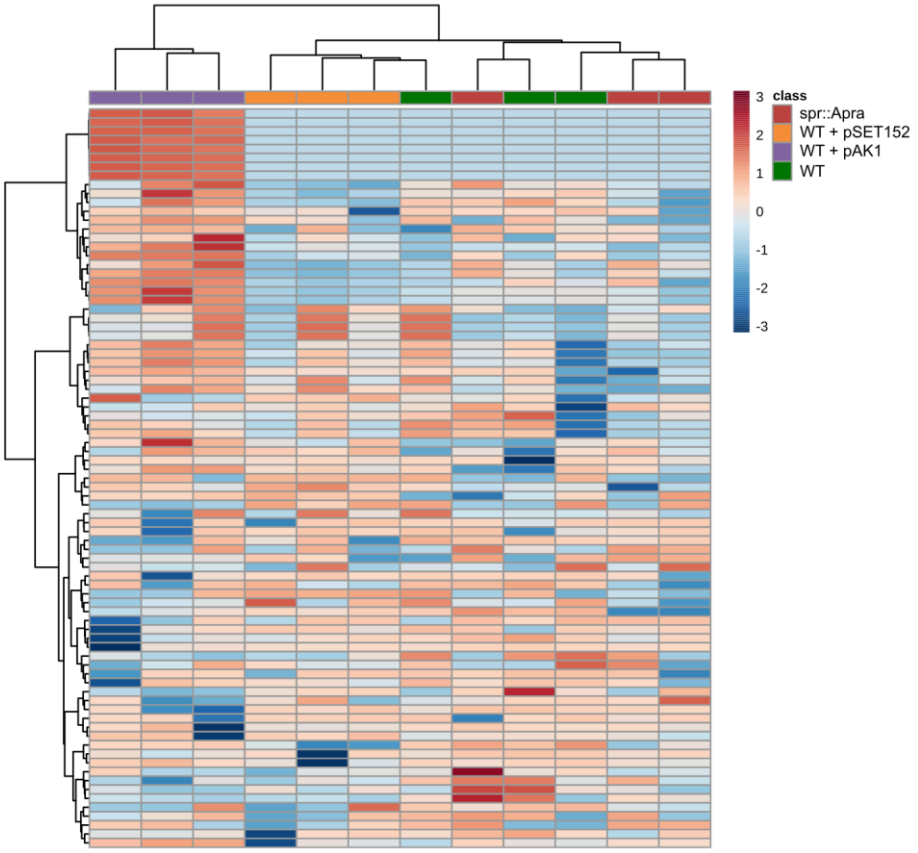


Figure S1. Heatmap of extracted peaks reveals seven peaks that are uniquely observed in strains containing the expression construct pAK1. Each row represents a single mass feature and each column represents a single extract, while the colour scale indicates the $\log_{10}$-scaled intensity of the mass features for each extract.

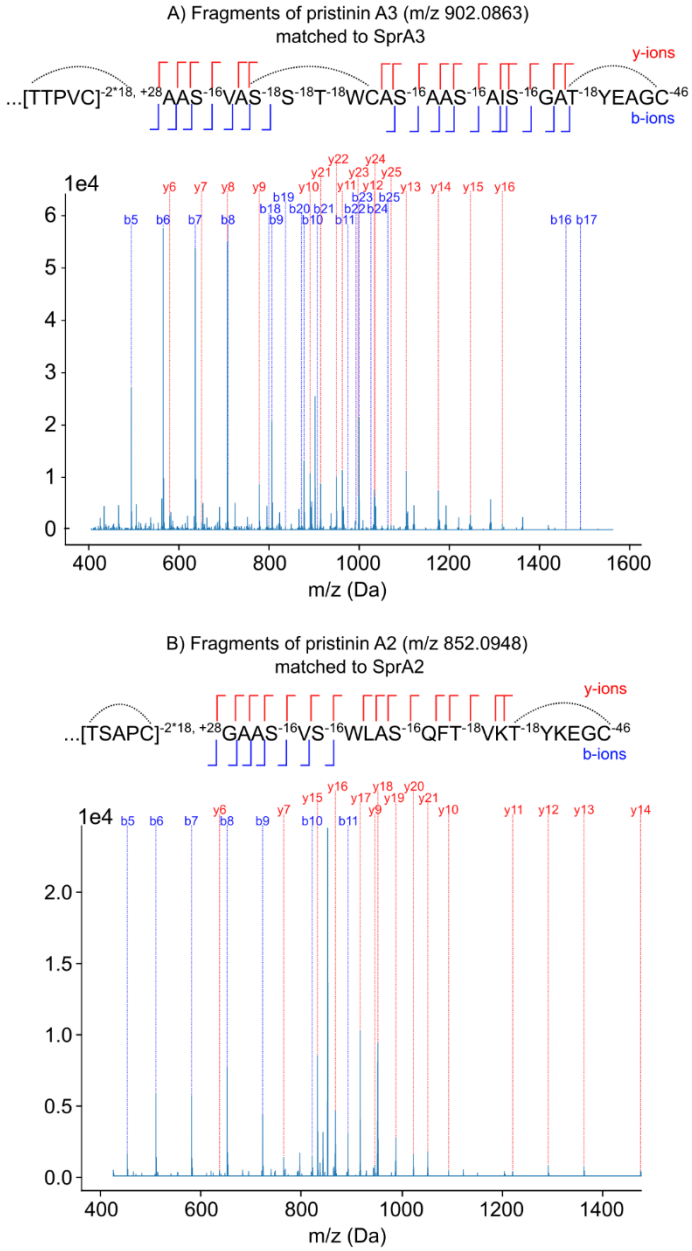


Figure S2. Fragmentation patterns of two highly extracted peaks can be matched to the SprA2 and SprA3 precursors. A full list of the modifications applied can be found in Table S6 and Table S7.

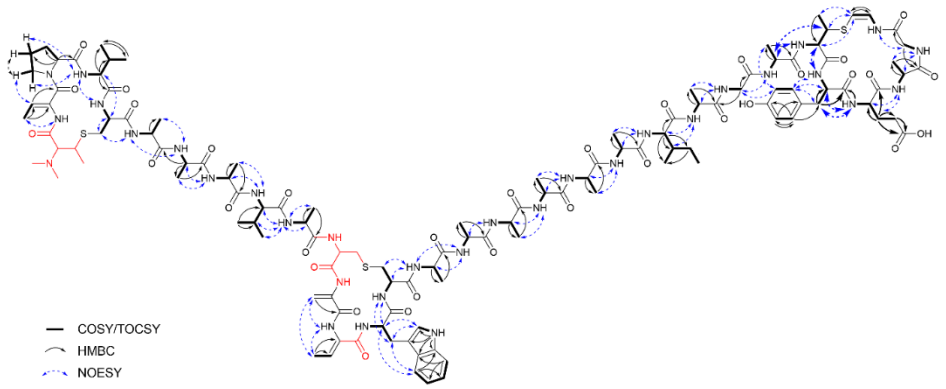


Figure S3. Key 2D NMR correlations observed for pristinin A3 (1). No clear correlations could be observed for the red parts of the structure, which were confirmed through other techniques. Bold arrows are for correlations which were better observed in CD_3CN:H_2O 9:1.

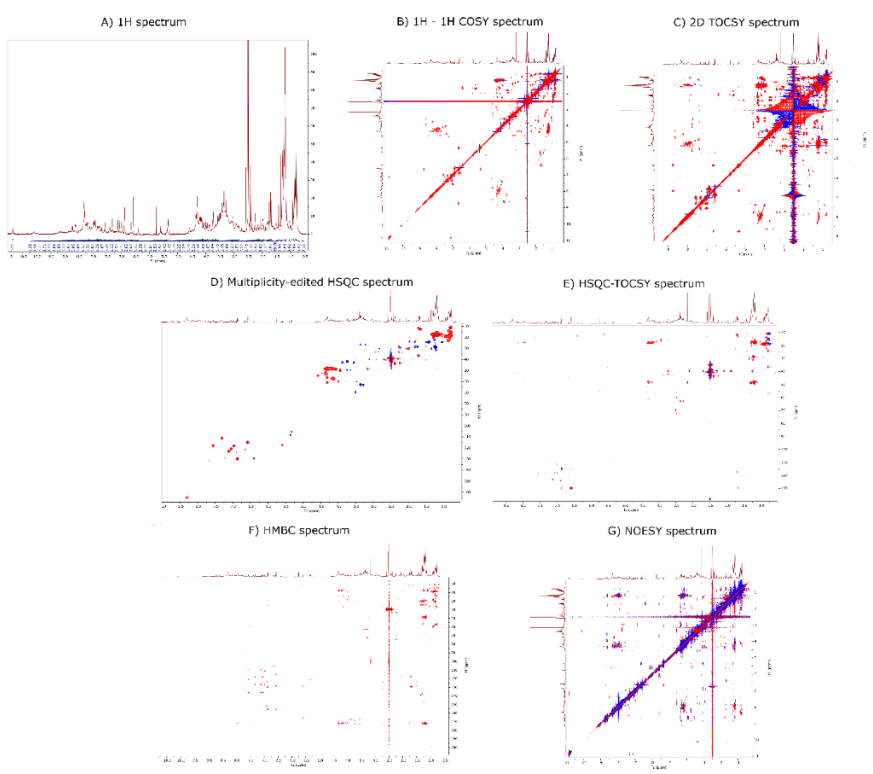


Figure S4. NMR spectra of pristinin A3 (850 MHz, in DMSO-*d6*, 298K). A) ^{1}H NMR spectrum with water suppression. The peak at 3.17 ppm is due to traces of methanol in the sample. B) ^{1}H–^{1}H COSY spectrum. C) 2D TOCSY spectrum. D) Multiplicity-edited HSQC spectrum. E) HSQC-TOCSY spectrum. F) HMBC spectrum. G) NOESY spectrum. Full data available on request.

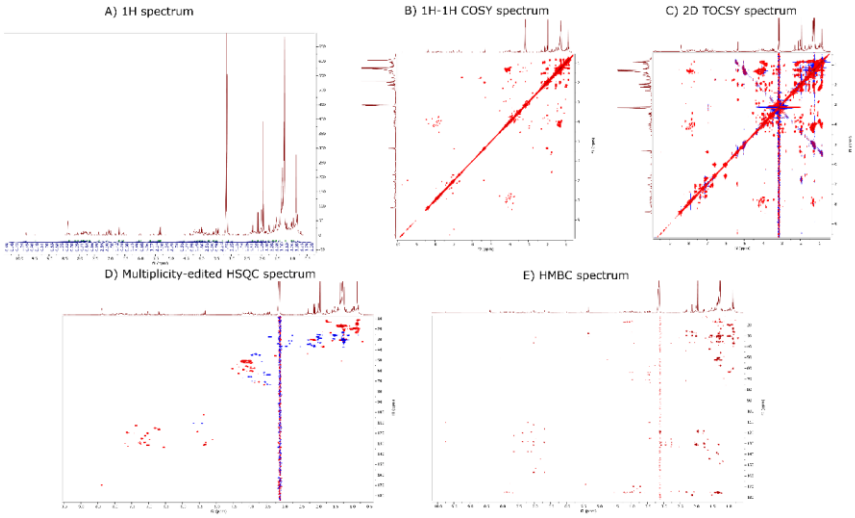


Figure S5. NMR spectra of pristinin A3 (850 MHz, in CD_3CN:H_2O 9:1, 297 K, first run). A) ^1H NMR spectrum with water suppression. B) ^1H–^1H COSY spectrum. C) 2D TOCSY spectrum. D) Multiplicity-edited HSQC spectrum. E) HMBC spectrum. Full data available on request.

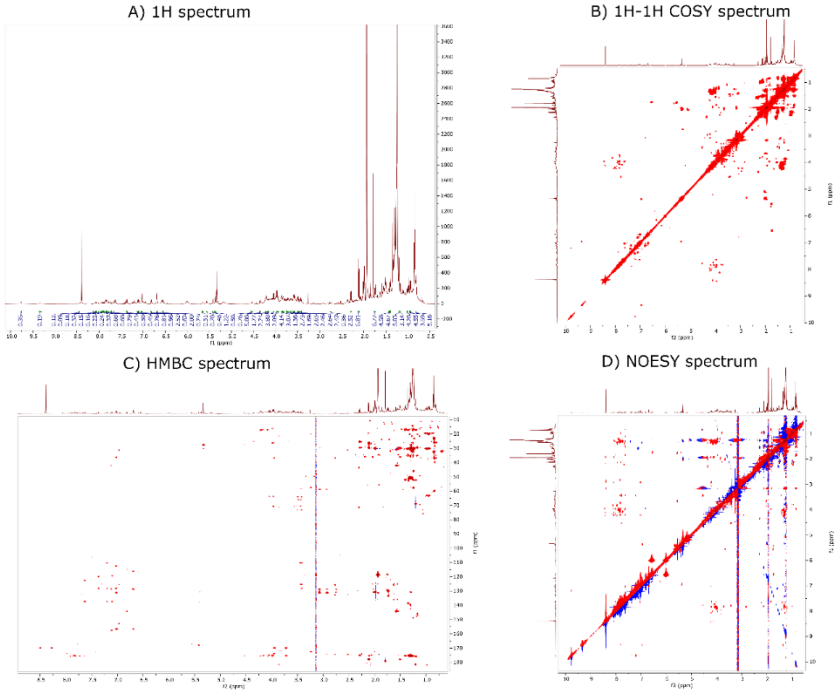


Figure S6. NMR spectra of pristinin A3 (850 MHz, in CD_3CN:H_2O 9:1, 297 K, second run). A) ^1H NMR spectrum. B) ^1H–^1H COSY spectrum. C) NOESY spectrum. D) HMBC spectrum. Full data available on request.

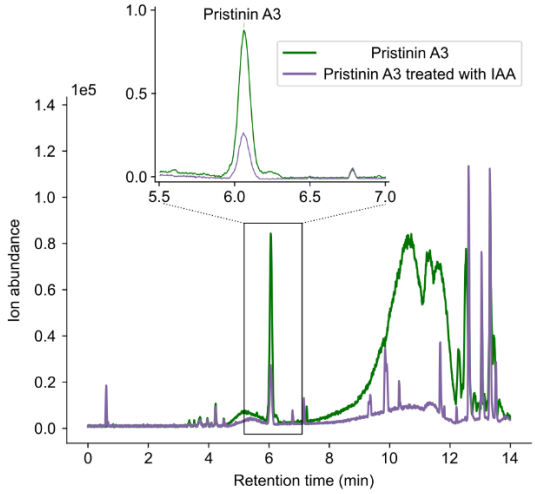


Figure S7. Labeling experiments with iodoacetamide (IAA) provide further support for the proposed structure of pristinin A3. (Purple) IAA covalently attaches to free sulfur groups of cysteines. However, the SprA3 peak was unaltered by IAA treatment, despite the presence of three cysteines in the peptide, strongly suggesting that these cysteines are not free.

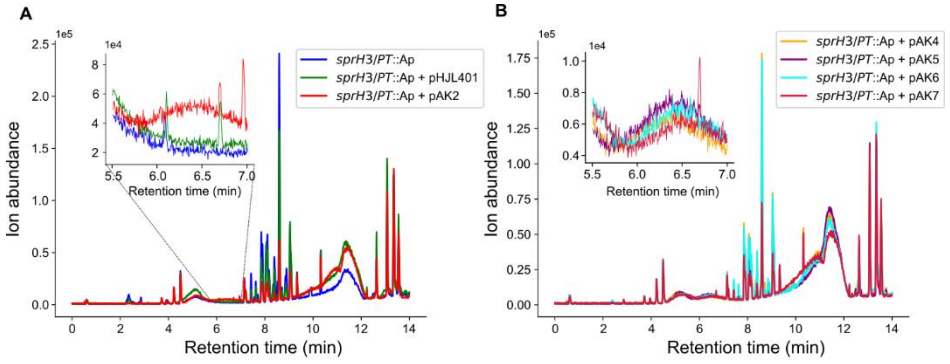


Figure S8. A mutant strain in which *sprPT* and *sprH3* are deleted no longer produces the *spr* RiPPs, but the production is not restored by complementation. A) LCMS analysis of crude extracts made of *sprH3PT*::Ap, with pHJL401, pAK2 or no vector. Lacking the *sprH3/PT* gene pair, the strain no longer produces the previously identified RiPPs. B) Complementation of the *sprH3PT* genes does not restore RiPP production. Whether its own native promoter was used (pAK4, pAK6) or the strong XNR_3799 promoter (pAK5, pAK7), even in combination with the *sprR* gene behind the gap promoter (pAK6, pAK7), the masses corresponding to the RiPPs were no longer detected.

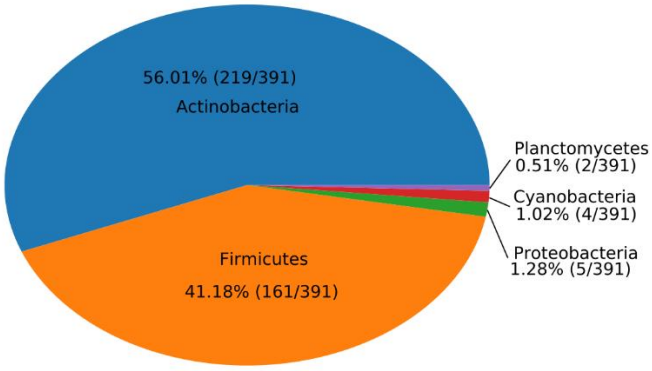


Figure S9. Homologs of the sprPT and sprH3 gene pair are present outside Streptomyces. Most homologs were found in Actinobacteria and Firmicutes, although a few additional candidates were found in Proteobacteria, Cyanobacteria and Planctomycetes.

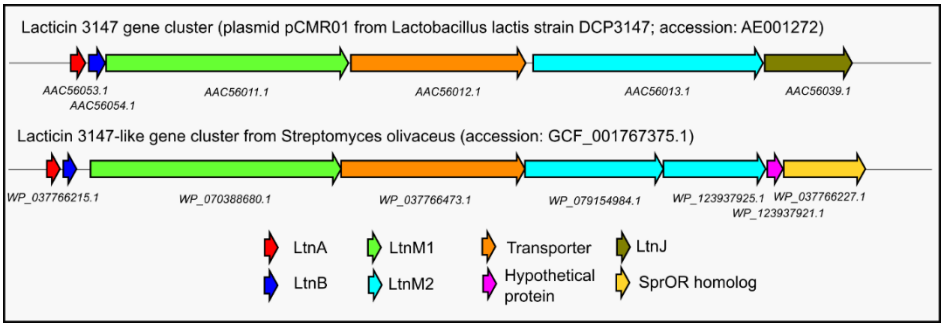
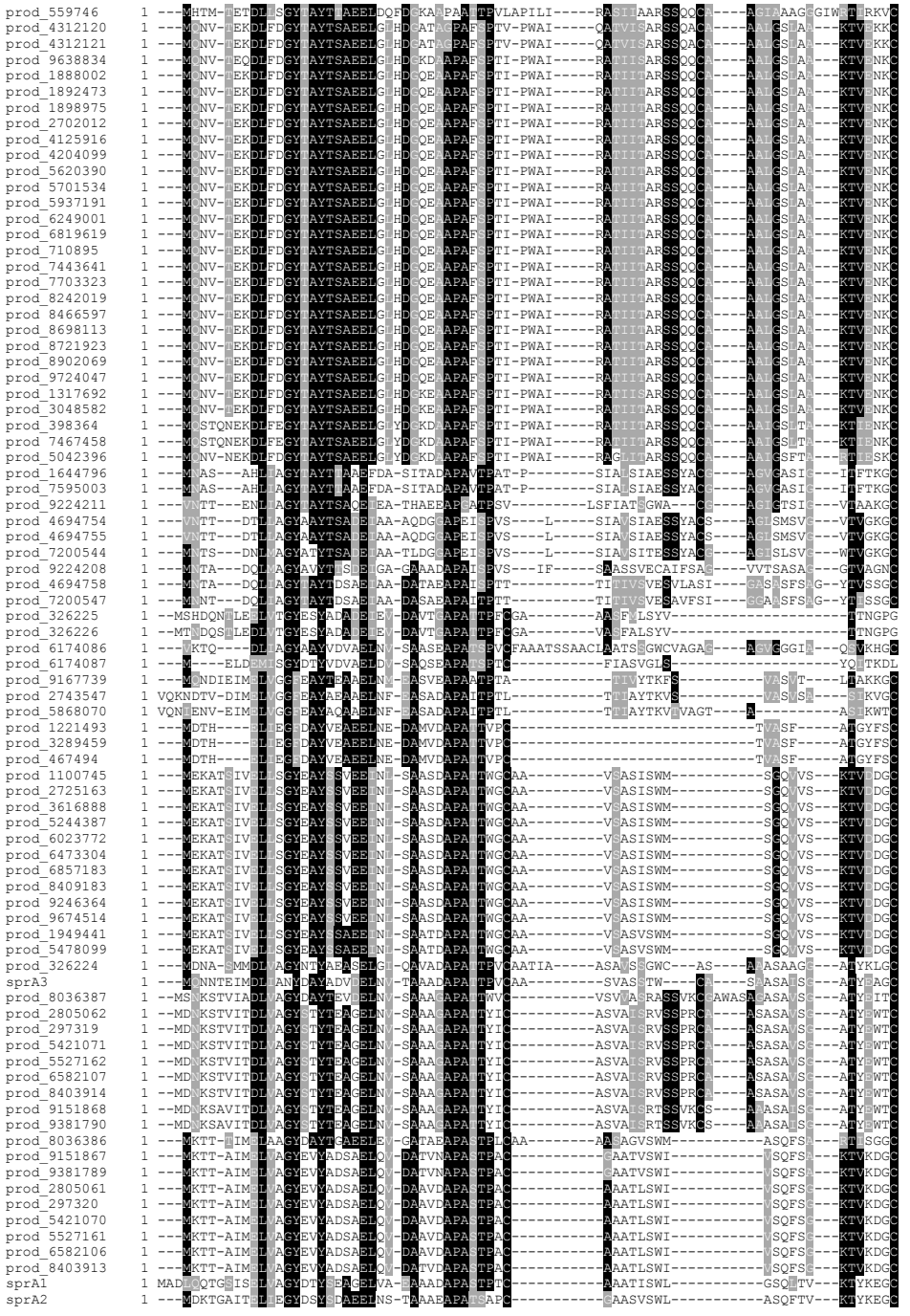


Figure S10. Comparison of the lacticin 3147-like gene cluster from Lactococcus lactis with a homologous cluster from Streptomyces olivaceus. Genes encoding both precursors, both LanM-like modifying enzymes and the transporter are well conserved between the clusters. The gene encoding LtnJ, however, responsible for the reduction in the conversion to alanine and butyric acid, was not conserved. Instead, a homolog to sprOR was found, suggesting it may carry out a similar function.



Text S1. Alignment of precursors belonging to the characterized family of type V lanthipeptides.
Precursors were aligned with MUSCLE [153] and visualized with BoxShade.

Table S1. Primers used in this study.

Primer name	Primer sequence
sprR_F	gatc GAATTC CAT ATGACCGTCAACGACCTGTCC
sprR_R	gatc TCTAGA CGCGGCCCACGGATCAGACC
spr_LF_F	gatc GAATTC CTCGCGGCCCTCGGCATTCTGG
spr_LF_R	gcta TCTAGA GTGGCGTGCGCGGCGTTGG
spr_RF_F	gcta TCTAGA CGCCCGGAAACAGGCATGAAGG
spr_RF_R	gcta AAGCTT ATGTCGCGGTGGACGACACCC
spr_del_check_F	GGGCTACATGCCTACTTTGC
spr_del_check_R	GTGCCCTCTGATTCCTTTCC
sprH3PT_LF_F	gcta GAATTC CCTCCTTGCGGAAGGCAGC
sprH3PT_LF_R	gcta TCTAGA CCTTACGACGGCTGAGGCGG
sprH3PT_RF_F	gcta TCTAGA CGTCCCGAAGCGCTCTGAC
sprH3PT_RF_R	gcta AAGCTT GCTTTCTTCCTCGTCATCGGCGG
sprH3PT_check_F	gcta TCTAGA TTCCTCGTTCGCGCTTTCTCCG
sprH3PT_check_R	gcta TCTAGA CCTTACGACGGCTGAGGCGG
sprH3PT_compl_F_t0_prom	gatc TCTAGA TTGTTCAGAACGCTCGGTCTTGCACACCGGGCG TTTTTTCTTTGTGAGTCCA GGGTGCCCTCTGATTCCTTTCCG
sprH3PT_compl_F	gatc TCTAGA CAT ATGACAGTGATGCTGGAGGCCACG
sprH3PT_compl_R	gatc AAGCTT TCAGCGGCGAGGCAGATTCC
XNR_t0_F	gcta TCTAGA TTGTTCAGAACGCTCGGTCTTGC
XNR_t0_R	gcta AAGCTT gatc CATATGCCGACCTCCCCCTTCG

Table S2. Plasmids used in this study.

Plasmid	Description	Reference
pSET152	Integrative *E. coli* / *Streptomyces* shuttle vector.	Bierman *et al.* [199]
pHJL401	*E. coli/Streptomyces* shuttle vector with intermediate copy number.	Larson *et al*. [201]
pWHM3	Unstable *E. coli/Streptomyces* shuttle vector with high copy number; used for homologous recombination.	Vara *et al*. [198]
pUWLCRE	Unstable *E. coli/Streptomyces* shuttle vector containing the Cre recombinase enzyme, behind a constitutive promoter.	Fedoryshyn *et al*. [200]
pAK1	pSET152 containing *sprR* behind GAPDH promoter from *S. coelicolor* (SCO1947).	This work.
pAK2	pHJL401 containing *sprR* behind GAPDH promoter from *S. coelicolor* (SCO1947).	This work.
pAK3	pWHM3 containing regions flanking the *spr* gene cluster.	This work.
pAK4	pHJL401 containing the *sprH3/PT* gene pair with native promoter.	This work.
pAK5	pHJL401 containing *sprR* behind GAP promoter from *S. coelicolor*, a t0 terminator, the *sprH3/PT* gene pair with their native promoter.	This work.
pAK6	pHJL401 containing the *sprH3/PT* gene pair behind the XNR_3170 promoter.	This work.
pAK7	pHJL401 containing *sprR* behind GAP promoter from *S. coelicolor*, a t0 terminator, the XNR_3170 promoter and the *sprH3/PT* gene pair.	This work.
pAK8	pWHM3 containing regions flanking the *sprH3/PT* gene pair.	This work.

Table S3. Proteins containing a flavoprotein domain (PF02441) are present in both RiPP and non-RiPP BGCs. While proteins with this domain are known in RiPP biosynthesis for the decarboxylation of C-terminal cysteines, their presence is not restricted to RiPP BGCs.

MIBiG BGC ID	BGC class	RiPP class (if applicable)	Protein accession
BGC0000157	Polyketide		ABI94381.1
BGC0000158	Polyketide		ABV91288.1
BGC0000171	Polyketide		CCC21124.1
BGC0000203	Polyketide		ADI71473.1
BGC0000203	Polyketide		ADI71437.1
BGC0000373	NRP		EFG10345.1
BGC0000807	Saccharide		ADD45285.1
BGC0000889	Other		BAM73626.1
BGC0000932	Other		AFO93363.1
BGC0001115	NRP/Polyketide		CBK62752.1
BGC0001193	NRP		AJI44175.1
BGC0001362	Other		AFO93363.1
BGC0001592	Other		AVI10267.1
BGC0000508	RiPP	Lanthipeptide	CAA44255.1
BGC0000514	RiPP	Lanthipeptide	ABC94905.1
BGC0000527	RiPP	Lanthipeptide	CAB60260.1
BGC0000529	RiPP	Lanthipeptide	ADK32557.1
BGC0000530	RiPP	Lanthipeptide	EMC15126.1
BGC0000531	RiPP	Lanthipeptide	AAG48568.1
BGC0000533	RiPP	Lanthipeptide	AAD56146.1
BGC0001618	RiPP	Lanthipeptide	ARD24448.1
BGC0001669	RiPP	Lanthipeptide	AVH76813.1
BGC0000582	RiPP	Linaridin	ADR72965.1
BGC0000583	RiPP	Linaridin	YP_001827875.1
BGC0000625	RiPP	Thioamide-containing peptide	BAN83921.1
BGC0001802	RiPP	Thioamide-containing peptide	ATJ00796.1
BGC0001803	RiPP	Thioamide-containing peptide	BAN83921.1
BGC0001696	RiPP	Thioamide-containing peptide	BBC15202.1

Table S4. Peaks unique to strains containing pAK1 appear to be mostly derived from a single mass. Charges were predicted from isotope patterns, and monoisotopic masses were calculated on assuming M+H ions.

Peak *m/z*	Predicted charge	Monoisotopic mass	Description
707.3534	1	706.3454	Fragment of 2703.2349
868.0891	3	2601.2433	
902.0863	3	2703.2349	
903.4186	3	2707.2318	2703.2349 + 4 Da (2*H_2)
907.4167	3	2719.2261	2703.2349 + 16 Da (O)
908.7487	3	2723.2221	2703.2349 + 20 Da (O + 2*H_2)
914.9003	NA	NA	
918.0807	3	2751.2181	2703.2349 + 48 Da (3*O)

Table S5. Many detected masses from strains containing the expression construct pAK2 appear to be derived from two masses. The two base masses were also the most abundant, making it likely these form final products, while the other masses may be incompletely processed products.

Description	Calculated *m/z* M+3H (Da)	Observed *m/z* pAK2	Δppm
Most abundant mass #1	902.088	902.085	3.3
+ oxygen	907.42	907.417	2.8
	907.42	907.417	2.4
+2 oxygen	912.751	912.748	3.6
	912.751	912.75	1.3
+3 oxygen	918.083	918.08	2.6
	918.083	918.081	1.6
	918.083	918.082	1.2
+4 oxygen	923.414	923.412	3.1
+ methyl	906.76	906.757	3.5
- methyl	897.416	897.415	0.8
- 2 methyl	892.744		
Most abundant mass #2	852.097	852.095	2.3
+ oxygen	852.376	857.426	0.9
+2 oxygen	862.758	862.758	0
+3 oxygen	868.09	868.09	0.6
	868.09	868.089	0.7
- 2 methyl	842.751	842.747	4.2

Table S6. Observed masses for fragments of a mass of 2703.235 Da can be matched to the SprA3 precursor. See also Figure S2A.

Ion	Calculated *m/z*	Calculated *m/z* (z = 2)	Observed *m/z*	Δppm	Ion	Calculated *m/z*	Calculated *m/z* (z = 2)	Observed *m/z*	Δppm
b1	112.0762				y1	76.0221			
b2	195.1134				y2	133.0436			
b3	292.1661				y3	204.0807			
b4	391.2345				y4	333.1233			
b5	494.2437		494.2437	0.05	y5	496.1866			
b6	565.2808		565.2808	0.06	y6	579.2237		579.224	0.49
b7	636.3179		636.3185	0.87	y7	650.2608		650.2596	1.88
b8	707.3551		707.3555	0.61	y8	707.2823		707.2825	0.30
b9	806.4235		806.4248	1.64	y9	778.3194		778.322	3.33
b10	877.4606		877.4608	0.25	y10	891.4035		891.4061	2.96
b11	946.4821		946.4810	1.11	y11	962.4406		962.4424	1.90
b12	1015.5035				y12	1033.4777		1033.4777	0.00
b13	1098.5406	549.7742			y13	1104.5148		1104.5161	1.17
b14	1284.6200	642.8139			y14	1175.5519		1175.5535	1.35
b15	1387.6291	694.3185			y15	1246.5890	623.7984	1246.5909	1.50
b16	1458.6663	729.8370	1458.6639	1.61	y16	1317.6261	659.3170	1317.6302	3.09
b17	1529.7034	765.3556	1529.7031	0.17	y17	1420.6353	710.8216		
b18	1600.7405	800.8742	800.8757	1.94	y18	1606.7146	803.8612		
b19	1671.7776	836.3927	836.3925	0.25	y19	1689.7518	845.3798		
b20	1742.8147	871.9113	871.9155	4.86	y20	1758.7732	879.8905		
b21	1813.8518	907.4298	907.4316	1.96	y21	1827.7947	914.4013	914.4006	0.72
b22	1926.9359	963.9719	963.9677	4.31	y22	1898.8318	949.9198	949.9199	0.09
b23	1997.9730	999.4904	999.4840	6.41	y23	1997.9002	999.4540	999.4547	0.68
b24	2054.9945	1028.0011	1028.0005	0.62	y24	2068.9373	1034.9726	1034.9719	0.66
b25	2126.0316	1063.5197	1063.5239	3.95	y25	2139.9744	1070.4911	1070.4915	0.34
b26	2209.0687	1105.0383			y26	2211.0116	1106.0097		
b27	2372.1320	1186.5699			y27	2314.0207	1157.5143		
b28	2501.1746	1251.0912			y28	2413.0892	1207.0485		
b29+1	2572.2117	1286.6098			y29	2510.1419	1255.5749		
b30+1	2629.2332	1315.1205			y30	2593.1790	1297.0934		
b31+1	2686.2369	1343.6224			y31	2704.2475	1352.6276		

Table S7. Observed masses for fragments of a peak corresponding to a monoisotopic mass of 2553.260 Da can be matched to the SprA2 precursor. See also Figure S2B.

Ion	Calculated *m/z* (*z*=1)	Observed *m/z*	Δppm	Ion	Calculated *m/z* (*z*=1)	Calculated *m/z* (*z*=2)	Observed *m/z*	Δppm
b1	112.0771			y1	76.02193			
b2	181.0986			y2	133.0434			
b3	252.1357			y3	262.086			
b4	349.1885			y4	390.1809			
b5	452.1977	452.1976	0.1	y5	553.2443			
b6	509.2191	509.219	0.2	y6	636.2814		636.2841	4.3
b7	580.2562	580.2552	1.8	y7	764.3764		764.3755	1.1
b8	651.2933	651.2933	0.0	y8	863.4448			
b9	722.3304	722.3316	1.6	y9	946.4819		946.4829	1.1
b10	821.3989	821.3968	2.5	y10	1093.55		1093.554	3.5
b11	892.436	892.4338	2.4	y11	1221.609		1221.606	2.1
b12	1078.515			y12	1292.646		1292.651	3.9
b13	1191.599			y13	1363.683		1363.672	8.1
b14	1262.636			y14	1476.767		1476.769	1.0
b15	1333.674			y15	1662.846	831.9272	831.9276	0.4
b16	1461.732			y16	1733.884	867.4458	867.4469	1.3
b17	1608.801			y17	1832.952	916.98	916.9802	0.2
b18	1691.838			y18	1903.989	952.4986	952.4948	3.9
b19	1790.906			y19	1975.026	988.0171	988.0172	0.1
b20	1919.001			y20	2046.063	1023.536	1023.54	4.0
b21	2002.038			y21	2103.085	1052.046	1052.043	3.6
b22	2165.101			y22	2206.094			
b23	2293.196			y23	2303.147			
b24	2422.239			y24	2374.184			
b25	2479.26			y25	2443.205			

Table S8. Cysteines linked to serine and threonine residues are detected after acidic hydrolysis of pristinin A3. Most of the predicted masses of the amino acids can be detected by HPLC-MS, including the cysteines linked to dehydrated serine and threonine residues.

Amino acid	Calculated *m/z* ($M+H^+$)	Observed *m/z*	Δppm
Glycine	76.04	NA	NA
Serine $^{-18}$	88.04	NA	NA
Alanine/Serine^{-16}	90.056	90.055	13
Threonine $^{-18}$	102.056	NA	NA
Proline	116.072	116.071	8.9
Valine	118.087	118.086	9.3
Isoleucine	132.103	132.102	7.9
Glutamate	148.062	148.06	7.1
Decarboxylated cysteine – threonine	177.07	NA	NA
Tyrosine	182.082	182.081	5.5
Tryptophan	205.098	NA	NA
Cysteine – Serine	209.06	209.059	2
Cysteine – Threonine (twice methylated)	251.108	251.106	5.5

Table S9. ^{1}H and ^{13}C NMR data for pristinin A3 (DMSO-d_6, 850 MHz, 298 K).

Residue	Position	δ_C, type	δ_H, mult.[a] (*J* in Hz)
Dhb-2	α	129.1, C	
	β	117.4, CH	5.59, q (7.2)
	γ	11.5, CH_3	1.71, d (7.2)
	CO	165.5, C	
	NH		9.20
Pro-3	α	60.7, CH	4.31
	β	29.4, CH_2	2.23
	γ	23.8, CH_2	a: 1.97
			b: 1.82
	δ	49.5, CH_2	a: 3.75
			b: 3.65
	CO	172.2, C	
Val-4	α	59.7, CH	4.00, t (8.2)
	β	28.2, CH	2.26
	γ	18.9, CH_3	0.94, d (6.9)
	γ′	19.1, CH_3	0.90, d (6.9)
	CO	171.0, C	
	NH		7.40
Ala(S)-5	α	53.8, CH	4.36
	β	33.1[b], CH_2	a: 3.08
			b: 2.84
	CO	ND	
	NH		7.89
Ala-6	α	48.1, CH	4.4.29
	β	17.5	1.19
	CO	171.6	
	NH		8.13
Ala-7	α	48.3, CH	4.17
	β	17.6, CH_3	1.22
	CO	171.8	
	NH		7.78
Ala-8	α	48.1, CH	4.27
	β	17.6, CH_3	1.21
	CO	172.1	
	NH		7.84
Val-9	α	57.6, CH	4.12
	β	30.3, CH	1.99
	γ	17.8, CH_3	0.83, d (6.8)
	γ′	19.0, CH_3	0.83, d (6.8)
	CO	ND	
	NH		7.90
Ala-10	α	48.1, CH	4.32
	β	17.9, CH_3	1.28
	CO	172.1	
	NH		8.23
Dha-12	α	ND	
	β	108.2, CH_2	5.36, d (16.1)
	CO	166.0, C	
	NH		ND
Dhb-13	α	129.9, C	
	β	129.4, CH	6.39, q (6.9)
	γ	12.5, CH_3	1.68, d (6.9)
	CO	ND	
	NH		10.1
Trp-14	α	54.4, CH	4.58
	β	26.4, CH_2	a: 3.37
			b: 3.22
	1 (indole)		10.84, br s
	2 (indole)	123.1, CH	7.12, br s
	3 (indole)	109.9, C	
	3a (indole)	126.9, C	
	4 (indole)	118.0, CH	7.56, d (8.0)
	5 (indole)	118.2, CH	6.98, dd (8.0, 7.6)
	6 (indole)	120.7, CH	7.05, dd (8.3, 7.6)
	7 (indole)	111.2, CH	7.32, d (8.3)
	7a (indole)	136.0, C	
	CO	ND	
	NH		7.76
Ala(S)-15	α	48.1, CH	4.30
	β	33.1[b], CH_2	2.94
	CO	ND	
	NH		7.70
Ala-16	α	48.8, CH	4.07
	β	17.4, CH_3	1.14, d (7.2)
	CO	171.8, C	
	NH		7.60
Ala-17	α	48.3, CH	4.15
	β	17.4, CH_3	1.20
	CO	172.9, C	
	NH		7.92
Ala-18	α	48.2, CH	4.22
	β	18.1, CH_3	1.19
	CO	172.9, C	
	NH		7.93
Ala-19	α	48.2, CH	4.18
	β	17.4, CH_3	1.19
	CO	171.8, C	
	NH		7.97

Table S9 (continued).

Ala-20	α	48.2, CH	4.22	**Tyr-27**	α	48.1, CH	4.32
	β	18.1, CH_3	1.19		β	34.5[b], CH_2	a: 3.04
	CO	172.9, C					b: 2.92
	NH		7.93		1 (phenol)	128.1, C	
Ala-21	α	48.1, CH	4.29		2/6 (phenol)	129.8, CH	6.87, d (8.0)
	β	17.4, CH_3	1.17		3/5 (phenol)	114.8, CH	6.57, d (8.0)
	CO	171.8, C			4 (phenol)	155.5, C	
	NH		7.98		CO	ND	
Ile-22	α	56.9, CH	4.15		NH		8.31
	β	36.4, CH	1.70	**Glu-28**	α	51.4, CH	4.35
	γ	24.1, CH_2	a: 1.41		β	28.4,CH_2	a: 1.81
			b: 1.04				b: 1.74
	δ	10.8, CH_3	0.78		γ	33.5, CH_2	2.16
	β-CH_3	15.0, CH_3	0.78		δ	ND	
	CO	170.4, C			CO	ND	
	NH		7.82		NH		8.29
Ala-23	α	48.1, CH	4.32	**Ala-29**	α	49.8, CH	3.94
	β	17.6, CH_3	1.18		β	15.7, CH_3	1.19
	CO	172.2, C			CO	173.0, C	
	NH		8.14		NH		8.61
Gly-24	α	41.7, CH_2	3.73	**Gly-30**	α	42.7, CH_2	a: 3.88, dd (17.0, 6.7)
	CO	168.4, C					b: 3.55, dd (17.0, 4.7)
	NH		8.22		CO	167.7, C	
Ala-25	α	48.1, CH	4.32		NH		8.63
	β	17.4, CH_3	1.29	**Vinyl-amine-31**	α	121.3, CH	6.64, dd (9.9, 8.5)
	CO	172.1, C			β	105.3, CH	5.33, d (8.5)
	NH		8.23		NH		8.73
Abu(S)-26	α	56.2, CH	4.31	ND: Not clearly detected			
	β	39.7[b], CH	3.04	a Multiplicities not given to overlapping or broad signals			
	γ	17.9, CH_3	1.29	b Very weak 13C NMR signal in the HSQC			
	CO	ND					
	NH		8.81				

Table S10. ^{1}H and ^{13}C NMR data for the G24, A25 and the C-terminal ring of pristinin A3 (CD_3CN:H_2O 9:1, 850 MHz, 297 K)

Residue	Position	δ_C, type	δ_H, mult.[a] (*J* in Hz)
Gly-24	α	44.1, CH_2	3.83, dd (11.9, 5.6)
	CO	171.9	
	NH		8.15 [8.12]
Ala-25	α	50.5, CH	4.23
	β	17.2, CH_3	1.27
	CO	175.2	
	NH		8.06 [7.96]
Abu(S)-26	α	57.6, CH	3.87 [3.95]
	β	45.6, CH	3.26
	γ	20.2, CH_3	1.24
	CO	172.5, C	
	NH		8.46 [8.44]
Tyr-27	α	58.8, CH	4.20 [4.22]
	β	36.4, CH_2	a: 3.07 [3.06] b: 2.94 [2.92]
	1 (phenol)	128.7, C	
	2/6 (phenol)	131.3, CH	7.02
	3/5 (phenol)	116.3, CH	6.69, d (8.0)
	4 (phenol)	156.8, C	
	CO	174.0	
	NH		8.12 [8.08]
Glu-28	α	55.7, CH	4.09 [4.11]
	β	28.7, CH_2	1.94 [1.92]
	γ	34.3, CH_2	2.24 [2.18]
	δ	180.7, C	
	CO	174.4, C	
	NH		8.35 [8.41]
Ala-29	α	50.6, CH	4.22
	β	17.3, CH_3	1.28
	CO	175.3	
	NH		8.06
Gly-30	α	44.9, CH_2	a: 3.95 b: 3.73
	CO	169.6, C	
	NH		7.81 [7.82]
Vinylamine-31	α	125.8, CH	6.98
	β	102.2, CH	5.37, d (7.5)
	NH		9.46 [9.35]

* Following the long time acquisition of the COSY, TOCSY, HSQC and HMBC spectra; the solvent evaporated from the NMR tube. The sample was then re-prepared for additional NOESY and longer HMBC experiments. It was noted from the ^{1}H NMR spectrum of the second run that some ^{1}H signals, especially those of the NH, have slightly shifted. Accordingly, a COSY experiment was repeated in the second run, to relate the ^{1}H NMR resonances of the two sets of data. The ^{1}H NMR resonance in the second run is given in square brackets, if it is different from the first one.

[a] Multiplicities not given to overlapping or broad signals

Table S11. Ratio of oxidized product in samples analyzed by NMR. The relative ares indicate the integrated areas divided over the integrated areas of the unmodified base peak.

Extract	NMR solvent	Base peak +16 Da relative area	Base peak +48 Da relative area
Crude extract	NA	0.152	0.066
Pristinin A3	DMSO-d_6	0.103	ND
Pristinin A3	CD3CN:H2O 9:1	0.287	3

NA: Not applicable; ND: Not determined

Table S12. Fragmentation data of oxidized products. X's indicate that a mass was observed within 10 ppm. A mixture of oxidized and non-oxidized fragments can be observed when the fragments do not contain the center ring structure. When the fragments do contain the center ring structure, they are always oxidized, suggesting the center ring contains the oxidation.

Ion	Obs. *m/z*	Obs. *m/z* (+16 Da)	Obs. *m/z* (+32 Da)	Obs. *m/z* (+48 Da)	Ion	Obs. *m/z*	Obs. *m/z* (+16 Da)	Obs. *m/z* (+32 Da)	Obs. *m/z* (+48 Da)
b1					y1				
b2					y2				
b3					y3				
b4					y4				
b5	x				y5				
b6	x	x			y6	x			
b7	x	x			y7	x			
b8	x	x			y8	x	x		
b9	x	x			y9	x	x		
b10	x	x			y10	x	x		
b11	x				y11	x	x		
b12					y12	x			
b13					y13	x			
b14					y14	x			
b15					y15	x			
b16					y16				
b17				x	y17				
b18			x	x	y18				
b19				x	y19				
b20				x	y20				
b21				x	y21				x
b22				x	y22				x
b23					y23				x
b24					y24				x
b25				x	y25				x
b26					y26				
b27					y27				
b28					y28				
b29					y29				
b30					y30				
b31					y31				

Table S13. Homologs of the genes *lanJ$_A$*, *sprF1, sprF2 and sprOR* are found associated with both known lanthipeptide BGCs and close to the *sprPT/sprH3* gene pair. Homology was determined at a cutoff of 30% amino acid identity of the gene products. Within *Streptomyces* genomes, all homologs were found within the analyzed 1,295 genomes. It was then checked whether these homologs overlapped with an antiSMASH-detected lanthipeptide BGC, or were within 15 genes of the *sprPT/sprH3* gene pair. *sprOR* homologs were found within canonical lanthipeptide BGCs as well as associated with the *spriPT/sprH3* gene pair, suggesting its association with lanthipeptide BGCs. For non-*Streptomyces* genomes, the *sprPT/sprH3* gene pair was first detected, and homologs of the given queries were found within the 15 surrounding genes. Homologs of *lanJ$_A$* and *sprF1* are often found associated with *sprPT/sprH3* gene pair, suggesting they are involved in lanthipeptide biosynthesis.

Streptomyces genomes

Query	Overlap with lanthipeptide BGC	*sprPT/sprH3* gene pair	Overlap with both	Overlap with neither
lanjA	0	0	0	5
sprOR	124	137	2	199
sprF1	0	124	2	16
sprF2	13	135	2	348

Non-*Streptomyces* gene clusters

Query	Overlap with lanthipeptide BGC	*sprPT/sprH3* gene pair	Overlap with both	Overlap with neither
lanjA	0	40	0	0
sprOR	0	108	0	0
sprF1	0	111	0	0
sprF2	0	146	0	0

5

Characterization of a novel RiPP BGC identified in *Streptomyces* sp. MBT27

Alexander M. Kloosterman
Somayah S. Elsayed
Jasper van der Peet
Chao Du
Marnix H. Medema
Gilles P. van Wezel

Abstract

Actinobacteria are the most prolific producers of bioactive molecules. Their biosynthetic arsenal includes some two thirds of the clinical antibiotics and many other compounds of clinical and agricultural importance. The increased numbers of available sequenced genomes have revealed an enormous reservoir of biosynthetic gene clusters (BGCs). Mining of genomes for truly novel families of BGCs, however, requires a different approach. Here, we report the discovery of a candidate RiPP BGC, called *trc*, in *Streptomyces* sp. MBT27. The *trc* gene cluster was identified using our machine learning-based pipeline decRiPPter and encodes two candidate precursors containing a repeated TTGWQ-motif, as well as a radical SAM enzyme and a PGM1-like ATP-grasp ligase, which have been previously associated with RiPP biosynthesis. Constitutive expression of a *luxR*-like regulatory gene located within the BGC resulted in strongly increased expression of the *trc* gene cluster. Comparative LC-MS analysis of culture extracts revealed 113 mass features that were produced by strains expressing the *trc* gene cluster but were not detected in extracts of a *trc* null mutant. Grouping these mass features with GNPS networking revealed two major networks containing 73 of these mass features, suggesting they are derived from similar compounds. Taken together, our data support that the *trc* gene cluster specifies a range of small RiPPs, likely derived from a TTGWQ-motif present in all predicted precursor peptides. Further research is required to unveil how these compounds were modified, their biological role and possible application.

Introduction

Natural products are compounds synthesized by bacteria and fungi, which have widespread clinical applications [222]. Actinobacteria are filamentous bacteria that live in both soil and aquatic environments, and are the most prolific producers of bioactive molecules with clinical and biotechnological application. These include antibiotics and compounds with anticancer, antifungal, immunosuppressant or herbicidal activity [32, 158]. The majority of these natural products are produced by members of the genus *Streptomyces*. The enzymes specifying these natural products are encoded by clusters of genes, organized in one or more operons, which are referred to as biosynthetic gene clusters (BGCs). Due to chemical redundancy, the return of investment of high-throughput screening is decreasing rapidly [25, 223]. Still, it is expected that we have only scratched the surface of the chemical space of natural products [224]. Genome sequencing has uncovered that even the best-studied model actinomycetes possess many yet underexplored resources for natural products [27, 225, 226]. Most natural products discovered belong to previously characterized classes rather than new classes [227]. The key question that scientists need to answer now is, can we find truly novel classes of natural products that have hitherto been overlooked? These molecules are likely products of so-called cryptic BGCs that are poorly expressed under routine laboratory conditions, but require specific molecular signals or intensive genetic manipulation [30, 33].

The challenge of identifying novel BGCs poses an interesting conundrum: how can novel BGCs be detected without prior knowledge of specific genetic elements, while retaining a high detection accuracy? Many genome mining efforts aimed at BGC detection target one or more core enzymes required for the biosynthesis of these classes. These have proven highly effective in the detection of BGCs for natural products that have been well-characterized and contain highly conserved genes, such as those encoding nonribosomal peptide synthetases (NRPS) and polyketide synthases (PKS) [36, 228]. When such conserved genetic markers are missing, a more innovative approach is required. Ribosomally synthesized and post-translationally modified peptides (RiPPs) form a class of natural products where this is often the case.

While RiPPs all share the same generic biosynthetic procedure, in which a precursor peptide is modified and cleaved to form a natural product, the modifications and responsible enzymes vary enormously. This makes it impossible to design a single genome mining strategy for the detection of all RiPP BGCs. Nevertheless, the large diversity covered by this class of natural products makes it an excellent candidate for the discovery of novel biosynthetic pathways.

Recently, we reported on a novel pipeline for the detection of novel RiPP BGCs, called decRiPPter (Chapter 3). This pipeline combines Support Vector Machine (SVM) models to detect candidate genes encoding RiPP precursors, with a pan-genomic analysis to prioritize novel candidate RiPP BGCs. A thorough analysis of 1,295 *Streptomyces* genomes resulted in the identification of 42 novel candidate RiPP families. While these candidate BGCs were not detected by conventional RiPP genome mining methods, some of them contained genes found among many different RiPP subclasses, such as genes encoding radical S-adenosyl methionine (SAM) utilizing enzymes, YcaO enzymes or ATP-grasp enzymes. Some of these genes have been used previously as 'bait' queries to identify novel RiPPs, efforts which have led to the discovery of the spliceotides [103], the WGK RiPPs [104], and the thiovarsiolins [52]. The presence of such RiPP-associated genes inside a gene cluster is no guarantee for the discovery of novel RiPPs, however. For example, a search for homologs of *pgm1*, a gene encoding an ATP-grasp ligase involved in the biosynthesis of pheganomycin, led to the discovery of the ketomemicins, which are non-RiPP natural products [163]. This example nevertheless illustrates that characterizing BGCs that show some relation to known BGCs may be a fruitful approach leading to the discovery of novel types of natural products.

Here, we describe the characterization of a gene cluster from *Streptomyces* sp. MBT27, which was detected by decRiPPter as a candidate RiPP BGC. The gene cluster contains several interesting RiPP markers, including a *pgm1* homolog, as well as a gene encoding a radical SAM enzyme, of which we study their relation to homologs from known BGCs. In addition, two closely related predicted precursors are encoded, which contain highly conserved TTGWQ-repeats. Five other gene clusters with similar features are discovered,

which form the candidate RiPP family. We studied the expression of the gene cluster by quantitative proteomics, and showed that it is naturally expressed under laboratory conditions. Comparative LC-MS analysis reveals that when the gene cluster is expressed at a higher level, several masses within the mass range 400 – 600 Da are detected at significantly higher levels, providing interesting leads for further research and characterization of any peptides that might be produced.

Results and Discussion

Bioinformatic characterization of the *trc* gene cluster

Given decRiPPter's capabilities of detecting BGCs of new RiPP subclasses, we aimed to find additional RiPP BGCs encoded by the analysed Streptomyces species. The large number of candidate RiPP BGCs detected by decRiPPter allows for many possibilities to filter them for candidates of interest, using numbers of encoded enzymes, transporters and regulators. Many candidate RiPP BGCs were identified by decRiPPter using a "strict filter". This filter requires the presence of two enzyme-coding genes, one gene encoding a transporter, one encoding a peptidase, and one encoding a regulator. In addition, the "Clusters of Orthologous Genes" (COG) score (which signifies the fraction of genomes within a taxon that contain a homolog of a given gene) of all genes in the gene cluster should be no higher than 0.1 on average. In other words, no more than 10 percent of the genomes analysed should contain an orthologue of any of the genes in the (core operon of the) gene cluster (Chapter 3). While using this strict filter increased the saturation of known RiPP BGCs among the results, 93% of known RiPP BGCs identified by antiSMASH [39] were filtered out in the process, which suggests that many more unknown candidate RiPP BGCs were filtered out as well. To investigate this, we examined the BGCs mined from 1,295 *Streptomyces* genomes, also considering BGCs that passed the mild filter (two encoded enzymes, one encoded transporter, average COG score ≤ 0.25). From these, we selected a promising candidate RiPP BGC that encodes a unique combination of enzymes, and was discovered in *Streptomyces* sp. MBT27 (from now on referred to as 'MBT27') [229]. The gene cluster consists of two putative operons in an opposing strand orientation, each starting with a predicted precursor gene. The shared sequence of the putative precursors is completely identical, with one precursor being 14 amino acids longer than the other. Interestingly, the precursors contained 3 and 4 repeats of a TTGWQ sequence, respectively, lending the cluster its preliminary name *trc* (TTGWQ-Repeat Containing RiPP candidate).

Enzymes encoded by the *trc* cluster include an ATP-grasp ligase (TrcC), a radical SAM protein (TrcD), oxidoreductases (TrcF, TrcG, TrcH3) and two aminotransferases (TrcJ, TrcK) (Figure 1). Directly adjacent to the operons lies a

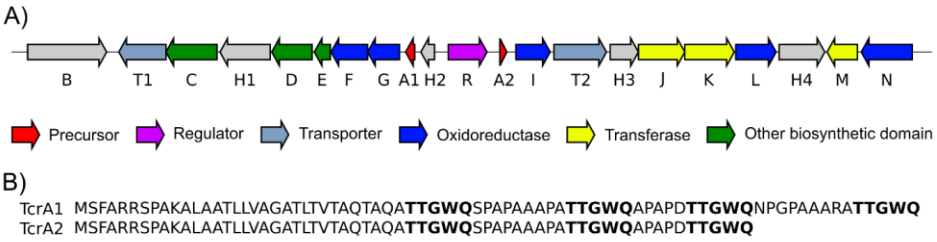


Figure 1. The *trc* gene cluster from *Streptomyces* sp. MBT27. A) The *trc* cluster consists of clusters of genes that likely form separate operons, each preceded by a putative RiPP precursor. B) Both predicted precursor proteins are highly similar to one another, and contain multiple repeats of a TTGWQ-motif. Further details on the annotation can be found in Table 1.

gene encoding a tryptophan halogenase (TrcN), although it is unclear whether this gene is part of the *trc* cluster. In addition, genes encoding a transporter (TrcT) and a regulator (TrcR) were found, which could have a cluster-specific role. We also searched for RiPP Recognition Elements (RREs), which facilitate precursor peptide binding in a wide variety of RiPPs [109], using RREFinder (Chapter 4). No RREs were found using either the conservative "precision mode" or the less restricted "exploratory mode". If this gene cluster indeed encodes proteins that produce a RiPP, precursor peptide recognition must be independent of an RRE.

To study the distribution of the *trc* cluster across bacteria, we employed MultiGeneBLAST [230] using all the genes in Figure 1 as a query. Five orthologous clusters of genes were found among the 1,295 streptomycetes studied (Figure S1). All of these contained both putative operons, showing little variation with respect to gene conservation and synteny. *trcN*, the gene encoding a tryptophan halogenase, was also well conserved, suggesting it was indeed part of the gene cluster. Outside of *Streptomyces*, only distantly related gene clusters were found in Actinobacteria, which share up to four different genes in a different genetic context (Figure S2). No gene clusters with significant similarity were identified outside of Actinobacteria, or among characterized BGCs from the MIBiG database [29].

Table 1. Annotation of the *trc* gene cluster.

Gene name	NCBI Accession	NCBI product annotation	Best Pfam hit	Pfam description
TrcB	WP_167162477.1	Threonine—tRNA ligase	PF00587	tRNA synthetase class II core domain (G, H, P, S and T)
TrcT1	WP_167162479.1	MFS transporter	PF07690	Major Facilitator Superfamily
TrcC	WP_167162481.1	Hypothetical protein	PF18604	Pre ATP-grasp domain
TrcH1	WP_167162483.1	Hypothetical protein		
TrcD	WP_167162485.1	Radical SAM protein	PF04055	Radical SAM superfamily
TrcE	WP_167162487.1	GFA family protein	PF04828	Glutathione-dependent formaldehyde-activating enzyme
TrcF	WP_167162489.1	Omega-3 fatty acid desaturase	PF00487	Fatty acid desaturase
TrcG	WP_167162880.1	2OG-Fe -dioxygenase family protein	PF10014	2OG-Fe dioxygenase
TrcA1	WP_167162491.1	Hypothetical protein		
TrcH2	WP_167162492.1	Hypothetical protein		
TrcR	WP_167162494.1	Helix-turn-helix transcriptional regulator		
TrcA2	WP_167162496.1	Hypothetical protein		
TrcI	WP_167162498.1	Phytanoyl-CoA dioxygenase family protein	PF05721	Phytanoyl-CoA dioxygenase (PhyH)
TrcT2	WP_167162500.1	Cation:proton antiporter	PF00999	Sodium/hydrogen exchanger family
TrcH3	WP_167162502.1	Hypothetical protein	PF00970	Oxidoreductase FAD-binding domain
TrcJ	WP_167162504.1	Aminotransferase class I/II-fold pyridoxal	PF00155	Aminotransferase class I and II
TrcK	WP_167162506.1	Aminotransferase class I/II-fold pyridoxal	PF01053	Cys/Met metabolism PLP-dependent enzyme
TrcL	WP_167162508.1	Hypothetical protein	PF02566	OsmC-like protein
TrcH4	WP_167162509.1	Hypothetical protein		
TrcM	WP_167162511.1	Class I SAM-dependent methyltransferase	PF13649	Methyltransferase domain
TrcN	WP_167162513.1	Tryptophan 7-halogenase	PF04820	Tryptophan halogenase

Predicted precursors encoded by the orthologous gene clusters are well conserved, particularly the N-terminal 31 aa (Figure S3). The C-terminal part of the peptides showed more variation, although all of them contain between two and four repeats of the TTGWQ sequence, a motif with unknown function. Possibly, the TTGWQ sequences form the core peptides, which are then processed to from the final product. This efficient usage of a precursor peptide, in which only a single leader peptide needs to be synthesized for multiple copies of the final product, has been reported for several other RiPPs, including cyanobactins [231, 232], orbitides [233], cyclotides [234], microviridins [235] and other omega-ester containing peptides (OEPs) [84], dikaritins [236-238], type II borosins [239], lyciumins [51] and pheganomycin [162].

Since decRiPPter also identifies putative precursors within non-RiPP BGCs, we looked for further evidence whether or not we could associate the *trc* cluster to the RiPP family of natural products. The gene *trcD* was predicted to encode a radical S-adenosyl methionine (radical SAM) enzyme. Radical SAMs typically share a conserved CxxxCxxC motif, containing a redox-active [4Fe-4S]-cluster binding an S-adenosyl methionine (SAM). These enzymes are highly divergent: a recent review grouped known examples of radical SAMs into 20 different families, which were further divided in almost 100 different subgroups [100]. Radical SAMs are encoded by the BGCs of many different RiPP subclasses [96]. Phylogenetic comparison of all radical SAM enzymes of characterized BGCs from MIBiG revealed several clades corresponding to enzymes involved in the biosynthesis of (multiple subclasses of) RiPPs (figure 2A). The protein sequences from these clades were mapped to the 20 different families of radical SAMs, by comparing their sequences with those of representatives of each family with BLAST. Nine of the 20 families were identified among all radical SAM enzymes, but only those containing a SPASM/Twitch-domain or a B12-binding domain were found among RiPP-related rSAMs. In this tree, the radical SAM enzymes with a B12-binding domain can be seen to clade together. These typically catalyse methylation, such as in bottromycins [97].

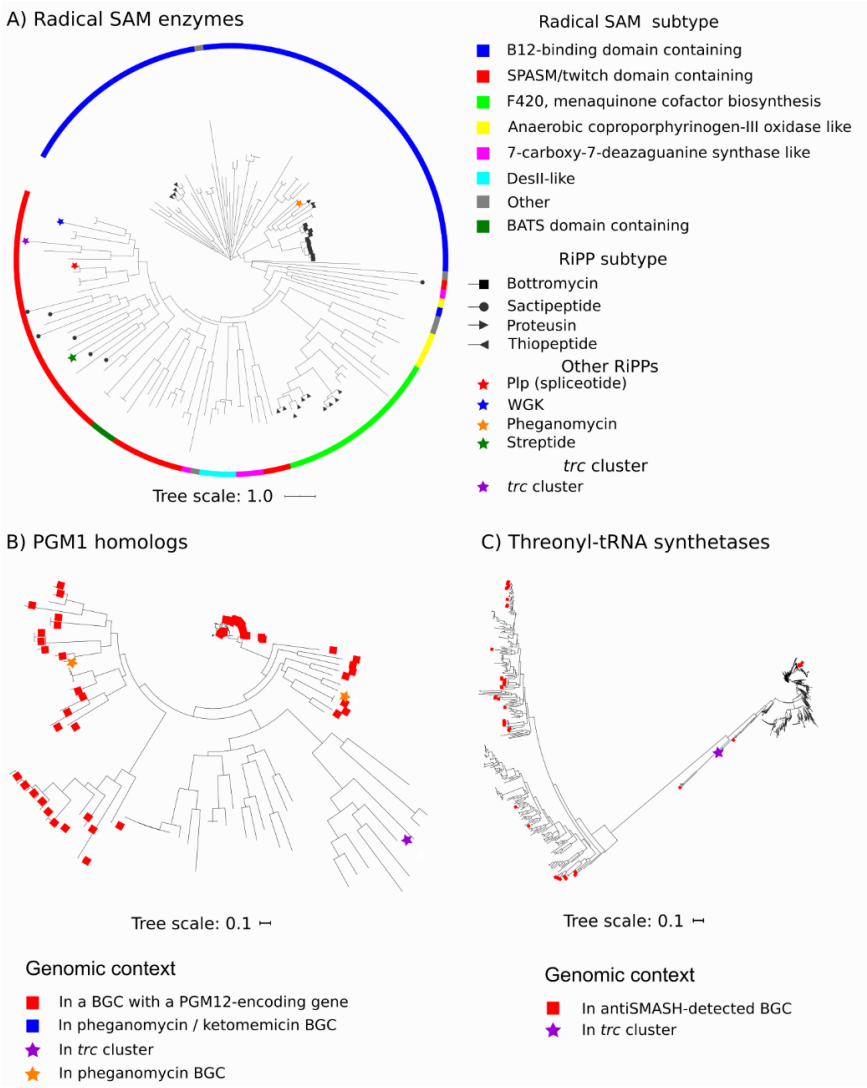


Figure 2. Homologs of TrcB, TrcC and TrcD overlap different types of BGCs predicted by antiSMASH. The relevant protein from the *trc* gene cluster is marked by a purple star. A) phylogenetic tree of rSAM enzymes detected in the MIBiG database. TrcD did not clearly overlap with any RiPP-associated clades. B) Phylogenetic tree of PGM1 homologs annotated in the antiSMASH database. Homologs are closely related when a PGM12-encoding gene was found in the same BGC (red). Other PGM1 homologs, including TrcC, are identified in a wider variety of BGCs and form a separate clade. C) Phylogenetic tree of 1,594 homologs of ThrRS found among 1,295 *Streptomyces* genomes. Among these homologs a closely related group was found, containing the majority of the homologs (1,255; righthand clade). The other homologs, including TrcB, showed a larger diversity, and were frequently found encoded in antiSMASH-detected BGCs (red).

The radical SAM of pheganomycin also belongs to this clade, marking it as distinct from TrcD, which fell within the clade of SPASM/Twitch-domain containing proteins. These radical SAM enzymes often form crosslinks, such as in sactipeptides and ranthipeptides [55, 99], or perform structural rearrangements, such as in spliceotides [103]. WGK and several recently identified RiPPs are all modified by radical SAM enzymes that belong to this family [104, 105, 107, 108]. These RiPPs are small (~500 Da) and contain a single crosslink applied by the radical SAM enzyme. The WGK radical SAM enzyme is closely related to TrcD, suggesting a similar modification is applied here. Still, as the majority of radical SAM enzymes (118 out of 142) were not encoded by a RiPP BGCs, and several of these enzymes were also closely related to TrcD, the presence of *trcD* alone did not provide conclusive evidence on whether the *trc* cluster encoded a RiPP.

trcC encodes a putative ATP-grasp ligase. ATP-grasp ligases catalyse nucleophilic additions onto carboxylic acids, which are activated as acylphosphate intermediates with ATP [181]. Members of this family have been shown to be involved in the biosynthesis of two RiPPs, pheganomycin [162] and omega-ester containing peptides (OEPs or graspetides) [85]. The ATP-grasp ligase predicted here does not show any significant similarity to any ATP-grasp ligase related to the OEP-family RiPPs found in a recent genome mining effort [85], but it is similar to PGM1, the ATP-grasp ligase encoded by the pheganomycin BGC. During biosynthesis of pheganomycin, the ATP-grasp ligase PGM1 catalyses the addition of an unnatural amino acid to a precursor peptide, via ATP-dependent formation of aminoacyl-phosphate derivative of this amino acid, followed by nucleophilic attack of the N-terminus of the peptide precursor. Several other proteins encoded by the pheganomycin BGC are thought to be involved in the biosynthesis of the unnatural amino acid, including PGM12, which shows high similarity to amidinotransferases. AntiSMASH detects BGCs related to the pheganomycin BGC using pHMMs built for the detection of PGM1 and PGM12. TrcC was detected by the pHMM built for PGM1, but a protein matching the pHMM for PGM12 was not encoded by the *trc* cluster. A phylogenetic tree containing all PGM1 homologs from the antiSMASH database [137] provided further insights in the relationship between the TrcC and PGM1 (figure 2B). Mapped onto this tree were those BGCs that contained genes for

PGM1 homologues. All PGM1 homologues in close proximity to a PGM12 homolog grouped together in two large clades, while the remaining PGM1 hits formed a third clade. The PGM1 homologue encoded by the *trc* cluster claded with the latter group of PGM1 homologues. In other words, the relationship of the *trc* cluster to pheganomycin was no stronger than that of other BGCs containing genes encoding PGM1 homologs, but no PGM12 homologs. Altogether, while *trcC* is not specific to BGCs that specify RiPPs, its presence, in combination with the presence of the radical SAM-encoding gene *trcD* and small ORFs encoding putative precursors with RiPP precursor-like conservation patterns, is at least strongly suggestive.

Lastly, a threonine-tRNA synthase (ThrRS) is likely encoded by *trcB*. Aminoacyl–tRNA synthases (AARS) are an essential part of primary metabolism, as they provide amino-acyl tRNA precursors used in ribosomal biosynthesis. Aminoacyl-tRNAs are used as a precursor in the biosynthesis of various secondary metabolites and antibiotics, such as type I lanthipeptides [47], 3-thiaglutamate [113, 147], cyclodipeptides [240] and valanimycin [241]. The BGCs for these antibiotics sometimes contain copies of the AARS to synthesize additional amino-acyl tRNA precursors, as is suspected for valanimycin [242, 243]. Another function of the gene could be to provide a variant that is resistant to the product of the BGC. This is the case for borrelidin, which targets AARSs, but has a resistant variant encoded in its BGC [244]. Bioinformatic analysis showed that genomes containing secondary copies of genes for ThrRS are not uncommon; among the 1,295 *Streptomyces* genomes, 1,594 ThrRS homologs were detected, averaging 1.23 per genome (Figure 2C; cutoff: 300 bitscore). 1,258 of these formed a closely related and highly conserved group. Only 11 of these overlapped with an antiSMASH-detected BGC, suggesting these were the ThrRS homologs involved in primary metabolism. In contrast, 57 of the remaining 337 homologs overlapped with an antiSMASH-detected BGC, showing that the presence of secondary household genes in BGCs is not an uncommon occurrence. TrcB was well removed from the highly conserved clade, further supporting the idea that the *trc* cluster from *Streptomyces* sp. MBT27 specifies a yet uncharacterized natural product.

Enhanced expression of the *trc* gene cluster and identification of the biosynthetic proteins

To experimentally characterize the *trc* cluster and its products, we aimed to enhance its expression *in vivo* by constitutive and strong expression of the transcriptional regulator. BGCs typically encompass a gene for a pathway-specific activator, which determines largely the timing and level of gene expression of the cluster [186, 245]. This property can be harnessed to efficiently over-express a BGC, and thus allow identification of the natural products that are overrepresented in the culture fluid of the recombinant strains [246]. The *trc* cluster contains a regulatory gene, *trcR* that encodes a putative LuxR-family regulator. These regulators often function as activators of BGCs in Actinobacteria [247]. We therefore placed a copy of *trcR* behind the strong and constitutive *gapdh* promoter from *Streptomyces coelicolor* [212, 213]. For this, we amplified the entire gene plus 30 nucleotides downstream from the *Streptomyces* sp. MBT27 genome and inserted it as an NdeI/XbaI fragment behind the *gapdh* promoter in the integrative vector pSET152. This construct was then introduced into MBT27, whereby the empty vector was used as the control. In this way, we created recombinant strains MBT27-*gaptrcR* and MBT27-pSET152. In parallel, we replaced the core region spanning the genes encoding both predicted precursors, the regulator and *trcH3* with the apramycin resistance cassette *aac(3)IV*. As this region contained both the regulator and both predicted precursors, we expected any secondary metabolite production of the gene cluster to be abolished in this strain. To this end, we used a method based on the unstable multi-copy vector pWHM3 [248]. We cloned the flanks upstream (-1507/-39) and downstream (+136/1641) of this region in pWHM3-oriT, inserted the *aac(3)IV* apramycin resistance cassette in-between and introduced this knock-out construct into MBT27 via conjugation. After several rounds of growth on non-selective media, followed by selection for the appropriate phenotype (apramycin^{R}, thiostrepton^{S}), we confirmed the replacement of the genes with PCR. The mutant strain was named MBT27-CR (Centre Replaced).

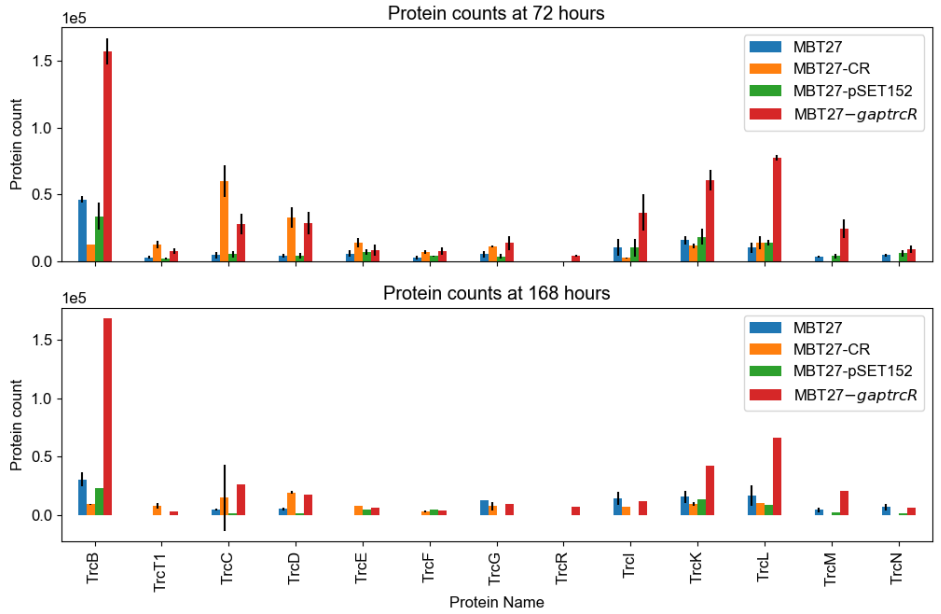


Figure 3. Expression of the *trc* cluster is affected by deletion of the core region and additional expression of *trcR*. All of the detected proteins were expressed at a higher level in a strain over-expressing TrcR. Surprisingly, in the proteome of the mutant MBT227-CR, which lacked the genes *trcA1-trcH3-trcR-trcA2*, several proteins encoded by the *trc* cluster were still detected, sometimes at a higher level than in the parental strain.

To establish the expression level of the *trc* gene cluster, and to see how gene expression would depend on the expression of *trcR*, we performed quantitative proteomics. As published previously, the expression level of BGCs corresponds very well to that of the metabolite produced from it, and hence the expression of the Trc proteins is a good measure of the expression of its cognate metabolite [249, 250]. For this, all strains were cultured in liquid minimal medium containing 0.5% (w/v) mannitol and 1% (w/v) glycerol as the carbon sources. Experiments were performed in triplicate. Mycelia were harvested after 72h and 168h, from which all proteins were isolated and analysed (Materials and Methods). Protein fragments were detected for 13 out of 19 proteins encoded by the *trc* cluster (Figure 3). Ten of these proteins were detected in all strains, showing that the gene cluster is expressed without genetic modifications. Strain MBT27-*gaptrcR* showed the highest overall expression of the *trc* cluster, in agreement with the function of TrcR as a transcriptional activator for the pathway.

Interestingly, in the knockout strain MBT27-CR the products of several genes in the *trcB-trcG* operon were detected at higher intensities than in the parent MBT27, despite the fact that the regulatory gene *trcR* had also been removed. Based on this, it appears that besides TrcR, other activating mechanisms exist. No other regulators were encoded on the contig that contains the *trc* cluster. Only one predicted regulator was detected at significantly higher levels in the proteome of both MBT27-CR and MBT27-*gaptrcR* was WP_167161651.1 (Figure S4). The gene encoding this protein was found in a terpene BGC, which suggests that it would function as a cluster-specific regulator for that BGC. Unfortunately, no other gene products of this BGC were detected by proteomics, so the involvement of this regulatory protein in the regulation of either the *trc* cluster or the terpene BGC could not be determined.

The proteins corresponding to the genes *trcI-trcH4* were still expressed in the mutant at levels comparable to those found in the parental strain MBT27. The only protein that was detected at significantly lower levels in the mutant was the threonine ligase TrcB, which also showed the strongest increase in MBT27-*gaptrcR* (~17-fold at 72 h). Taken together, these data show that TrcR functions as an activator of the *trc* cluster. However, the fact that the *trc* cluster was readily expressed in the wild-type strain indicates that TrcR it is not required *per se* for its expression. Interestingly, when *trcR* was removed alongside *trcA1*, *trcA2* and *trcH3*, most other gene products were more highly expressed. We cannot explain this based on the available data. Removal of these genes also changed the upstream regions of both putative operons, which could affect their transcriptional regulation. In addition, the product of the *trc* cluster itself may play a role in the regulation, as has been previously reported for other RiPPs [251]. Further experiments are required to unravel the exact regulatory mechanism of the *trc* cluster.

Analysis of the secreted metabolome by LC-MS

We then set out to identify the product synthesized by the *trc* cluster. We first removed the apramycin resistance cassette from the mutant strain via pUWLCRE-mediated Cre-Lox recombination [200], creating the in-frame deletion mutant MBT27-IFD. This strain together with MBT27 WT, MBT27-

pSET152 and MBT27-*gaptrcR* were cultured in liquid minimal medium and extracted with *n*-butanol, which has been previously used in the wide-scale extraction of different types of RiPPs simultaneously [252]. A total of six replicates were taken of each modified strain, while four replicates were taken for the wild-type strain. The crude extracts were analysed using LC-MS (Figure 4A and B), and mass features were extracted from the chromatograms with MzMine for comparative metabolomics using MetaboAnalyst (Materials and Methods).

Analysis of the heatmap based on the detected mass features revealed large differences among all the extracts (Figure S5). ANOVA analysis showed that 315 masses were differentially expressed among all the groups. For each of these masses, each possible pair of strains was compared with Fisher's LSD ($p \leq 0.05$) to find which masses were differentially expressed between which strains. Surprisingly, 25 different mass features were detected in higher intensities in the extracts of MBT27-IFD, as compared to the other groups. Apparently, removing part of the *trc* cluster altered the secondary metabolite profile of MBT27 in many ways. These masses may correspond to shunt products of the biosynthetic pathway of the *trc* cluster, as it was still partially expressed. Alternatively, abolishing production of the *trc* cluster may have freed up more resources for other BGCs, boosting their production, as previously reported [253]. In contrast, only two peaks were detected at higher intensities in extracts from MBT27-pSET152 compared to all other extracts. In total, 85 masses were detected in significantly higher levels in the extracts of MBT27-*gaptrcR*, compared to the extracts of all other strains (Fisher's LSD, $p \leq 0.05$).

Considering that the *trc* cluster was partially deleted, we selected candidate mass features which were not detected in the extracts of MBT27-IFD (peak area <= 3000), building on the assumption that secondary metabolite production of this gene cluster had been abolished. These criteria applied to 113 mass features, of which more than half (66) were detected at higher levels in the extracts of strain MBT27-*gaptrcR*. To further organize the results, we used molecular networking, through the Global Natural Products Social Molecular Networking (GNPS) platform to find relations between the identified mass features (Figure 4C) [254]. Using GNPS, mass features are represented as nodes,

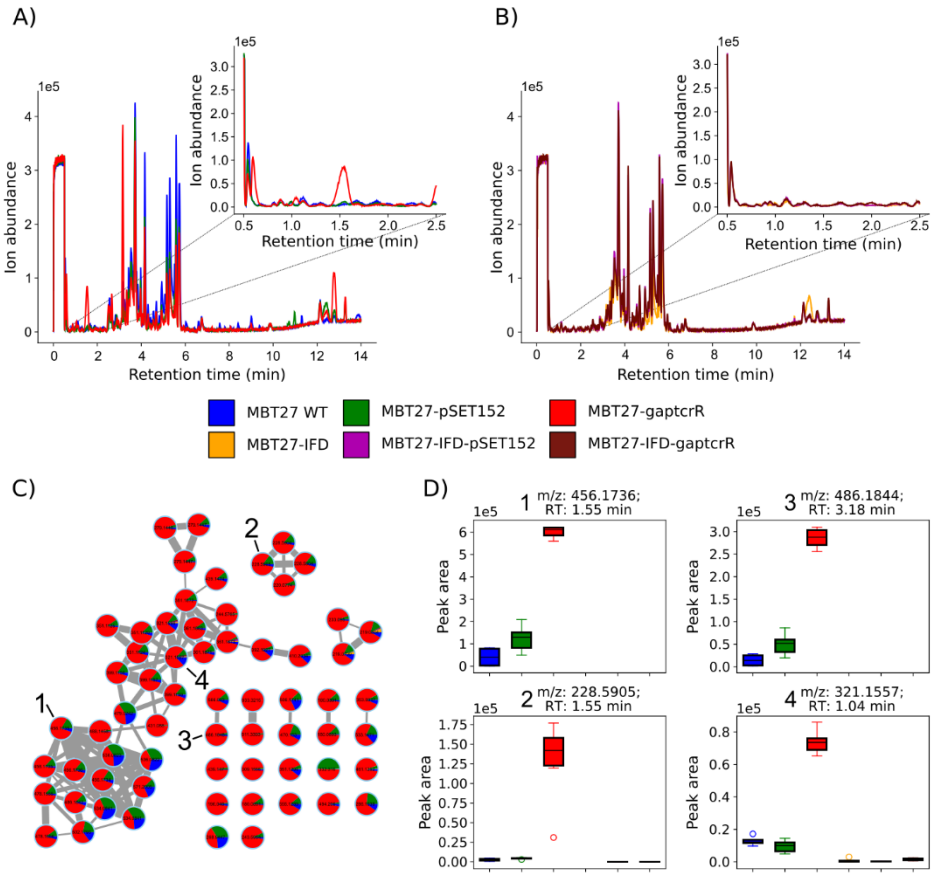


Figure 4. Comparative LC-MS analysis reveals large differences between the extracted strains. A) Overlay LC-MS chromatograms of MBT27 WT, MBT27-pSET152 and MBT27-*gaptrcR*. B) Overlay LC-MS chromatograms of MBT27-IFD, MBT27-IFD-pSET152 and MBT27-IFD-*gaptrcR*. C) Molecular families of the ions which were significantly enhanced in the extracts from a strain over-expressing TrcR as compared to the extracts of all other strains. At the same time, their production was very low or abolished in the strains lacking the core region of the *trc* cluster, regardless of whether trcR was over-expressed or not. Data of the complete network can be found in Data S1, and of this network in Data S2. D) Boxplot of selected peaks in the highlighted molecular families in C, compared among all six extracted strains.

which are connected to each other by edges due to similarities in their MS/MS spectra, or fragmentation patterns. When molecules share fragmentation patterns, it indicates that they have structural similarities. For 73 mass features

that were not detected in the extracts of MBT27-IFD, MS/MS spectra were obtained and the masses were compared. Most masses appeared as nodes connected into a single molecular family, suggesting that they belong to a related group of metabolites. The most highly expressed of these was a node with *m/z* 456.1736 (Figure 4D - 1). Several other highly abundant masses that were not linked with GNPS molecular families were still identified as related to this node, like it's doubly charged ion (*m/z* 228.5905, Figure 4D - 2), or nodes where the mass difference could be related to a number of atoms (*m/z* 486.1844, 456.1736 + mass of CH_2O group, figure 4D – 3). To look for further evidence that these peaks did not originate from effects of the enhanced expression of *trcR* that were unrelated to the *trc* cluster, we introduced the empty vector pSET152 and construct pAK10 into MBT27-IFD using conjugation, creating the strains MBT27-IFD-pSET152 and MBT27-IFD-*gaptrcR*, respectively. The mass features of interest were not detected in extracts of either of these strains, which makes it more plausible that they indeed originate from the *trc* cluster, and did not arise due to secondary effects from expressing the regulator. Further evidence for this could be gained by heterologously expressing the BGC in a different strain.

Of the masses described in the network, most had an *m/z* below 500 Da. While these masses are small for RiPPs, several RiPPs have been reported with a core peptide of only a few amino acids. These include the RiPPs modified by radical SAM enzymes that are closely related to TrcD (see above) [104, 105]. Similarly, the masses detected here may have been derived from the TTGWQ core sequence, which has a monoisotopic mass of 591.27 Da. However, we were unable to match the predicted mass to any of the identified masses in the LC-MS run, or identify amino acid residues from their respective MS/MS spectra, even when considering dehydration and deoxygenation of the precursor peptide. Additionally, it was not possible to identify a known structural class for these masses when the data were analysed using the MS2LDA tool [255]. Given the large number of enzymes encoded by the *trc* cluster, many different modifications to the precursor peptide are likely. Head-to-tail cyclization, for example, would make structure elucidation from MS/MS spectra difficult. Further structure resolution is required to completely resolve the structure of these metabolites and to determine whether or not they are RiPPs.

Conclusions and final perspectives

We have found a novel candidate RiPP BGC using decRiPPter, called the *trc* cluster, which was partially characterized via mutational analysis, expression, proteomics and mass spectrometry. The gene cluster contains several genes that relate to RiPP BGCs, such as a gene encoding an ATP-grasp ligase closely related to PGM1 and a gene encoding a radical SAM. The exact combination of enzymes has not been identified before, suggesting a novel natural product is specified by the *trc* cluster. We have not yet elucidated the natural product produced from the BGC. However, our bioinformatic analysis suggests that the gene cluster specifies a RiPP, whereby in particular the multiple TTGWQ repeats in the putative precursor peptides are suggestive of a RiPP, as short repeats are found in the precursors peptides of various RiPP sublasses. Enhanced expression of the regulatory gene *trcR* resulted in increased expression of the *trc* cluster, which could be correlated with the increased production of several secondary metabolites within the range of 400-600 Da. A few of these compounds were no longer produced when the cluster was partially removed, suggesting that these masses were products of the *trc* gene cluster. The mutant also lacked the genes *trcA1* and *trcA2*, suggesting that they may be involved in the biosynthesis of these products. We have so far been unable to confirm whether or not these products originated directly from the *trcA1* or *trcA2* precursors. Future studies will have to unveil the exact nature of the candidate masses, their relatedness, and the final product of the *trc* cluster.

Materials and Methods

Bioinformatics

Phylogenetic trees

For the generation of the phylogenetic trees, proteins were aligned with MUSCLE [153], and trees were generated with FastTree V2.1 [154] and visualized using iTOL [256].

Radical SAM

To create the radical SAM dataset, all proteins from the MIBiG database V2.0 [29] were scanned with hmmsearch [65, 134] against the Pfam [75] model of the radical SAM enzyme (PF04055), using the trusted cutoffs. The resulting proteins were mapped to radical SAM families by looking for the best hit among representatives of these families previously outlined [100] (Table S3). Phylogenetic trees were created as described above.

PGM1 homologs

The antiSMASH database [257] was queried for all BGCs containing a PGM1 homolog using the built-in query system. All BGCs were downloaded and parsed with BioPython [258] to detect PGM12 and PGM1 homologs. Phylogenetic trees were created as described above.

Threonyl-tRNA synthetases

Threonyl-tRNA synthetases were detected in 1,295 *Streptomyces* genomes analyzed previously with decRiPPter (Chapter 3). Protein homologs were detected with NCBI BLAST v.2.6 [56, 127] using a bit score cutoff of 300. Phylogenetic trees were created as described above.

Experimental procedures

Bacterial strain and growth conditions

Streptomyces sp. MBT27 was obtained from the Leiden University strain collection, which had been previously isolated from the Qingling Mountains [229]. Media components were purchased from Thermo Fisher Scientific, Sigma-Aldrich or Duchefa Biochemie. For strain cultivation on solid media, *Streptomyces* spores were spread on mannitol soya flour agar (SFM; 20 g/L Agar, 20 g/L mannitol, 20 g/L soya flour, supplemented with tap water) prepared as described previously [210], and incubated at 30°C. Spores were harvested after 4-7 days of growth when the strain started to produce a grey pigment, by adding water directly to the plate and releasing the spores with a cotton swab. Spores were centrifuged and stored in 20% glycerol.

For cultivation in liquid media, 20-50 µL of a dense spore stock was inoculated into 100 mL shake flasks with coiled coils containing 20 mL of the medium of interest. For extractions, NMMP was used (0.60 mg/L $MgSO_4$, 5 mg/L NH_4SO_4, 5 g/L Bacto casaminoacids, 1 mL trace elements (1 g/L $ZnSO_4.7H_2O$, 1 g/L $FeSO_4.7H_2O$, 1 g/L $MnCl_2.4H_2O$, 1 g/L $CaCl_2$, anhydrous)), while for genomic DNA isolation, a 1:1 mixture of TSBS: YEME with 0.5% glycine and 5 mM $MgCl_2$ was used (TSBS: 30 g/L Bacto Tryptic Soy Broth, 100 g/L sucrose; YEME: Bacto Yeast Extract: 3 g/L, Bacto Peptone 5 g/L, Bacto Malt Extract 3 g/L, glucose 10 g/L, sucrose 340 g/L).

E. coli strains JM109 and ET8 were used for general cloning purposes and demethylation, respectively. Strains were cultivated in liquid LB and on LB-agar plates at 37°C.

Molecular biology

All materials and primers were purchased from Sigma-Aldrich or Thermo Fisher Scientific unless stated otherwise. Restriction enzymes and T4 ligase were purchased from NEB. Restriction and

ligation protocols were followed as per manufacturer's description. For amplification of DNA fragments with PCR, Pfu polymerase was used. Primers were designed with T_m of the annealing region roughly equal to 60°C. Standard PCR protocols consisted of 30 cycles (45s DNA melting @ 95 °C, 45s primer annealing @55°C-65°C, 60s-180s primer elongation @ 72°C), but PCR protocols were optimized where necessary.

Following construction of the vectors (see below for specifics), constructs were transferred to MBT27 by conjugation [210]. Briefly, 50 µL of a dense MBT27 spore stock was added to 500 µL 2xYT, and a heat shock was applied at 50 °C for 10 minutes to trigger spore germination. In parallel, *E. coli* ET8 containing the construct of interest was grown until the OD_{600} measured 0.6 – 0.8 in 10 mL LB containing 50 µg/mL kanamycin, 50 µg/mL chloramphenicol and, as required, 20 µg/mL thiostrepton and 50 µg/mL apramycin. *E. coli* cultures were centrifuged, washed twice with LB to remove any remaining antibiotics, mixed with the germinated MBT27 spores and plated out on SFM plates containing 10 mM $MgCl_2$ and 10 mM $CaCl_2$. The plates were incubated at 30°C for 14-18 hours, and overlaid with 1.2 mL H_2O containing 417 µg/mL chloramphenicol, and, as required, 417 µg/mL thiostrepton and 1.04 mg/mL apramycin.

MBT27-CR knockout mutants were created by replacing the gene cluster with an *aac(3)IV* apramycin resistance cassette via homologous recombination. The -1553/-209 and +18/+1561 regions upstream and downstream of the *trc* cluster were amplified by PCR with the trc_LF_F/trc_LF_R and trc_RF_F/trc_RF_R primer pairs (table S1), respectively, and inserted into the pWHM3-oriT vector (Table S2) into the EcoRI/HindIII sites. The *aac(3)IV* apramycin resistance cassette was inserted into the created XbaI site, creating pAK9. pAK9 was transformed to *E.coli ET8* for DNA demethylation, which was used as a donor for transfer to MBT27 by conjugation [210]. Three colonies were picked after 4 days of growth and spread onto SFM plates without added antibiotic to allow for homologous recombination. Colonies containing the correct phenotype (apramycin-resistant, thiostrepton-sensitive) were picked and the homologous recombination was confirmed by PCR, using the trc_del_check_F/trc_del_check_R primer pair.

The strain MBT27-IFD was created by removal of the apramycin cassette from the strain MBT27-CR using the vector pUWLCRE [200]. This vector was conjugated to the strain MBT27-CR, and three separate colonies were picked and grown separately on SFM without antibiotics. After one round of growth, fresh spores were collected and plated at diluted concentrations to allow the spores to grow as individual colonies. From these, colonies were selected with the correct antibiotic resistance phenotype (apramycin-sensitive, thiostrepton-sensitive). Deletion was confirmed by PCR using the trc_del_check_F/trc_del_check_R primer pair.

Constructs for the overexpression of the *trcR* regulator were constructed as follows: the -0/+30 region of the *trcR* gene was amplified from the genomic DNA of MBT27 using the trcR_F/trcR_R primer pair, and placed into the EcoRI/XbaI site of the pSET152 vector. The -0/-457 upstream region of glyceraldehyde 3-phosphate dehydrogenase amplified from the genome of *S. coelicolor*, was obtained from previous studies [212, 213]. The promoter region was inserted into the EcoRI site and the engineered NdeI site, placing it directly upstream of the *trcR* gene. The resulting vector was named pAK10.

Extractions

Strains were cultured in 100 mL shake flasks containing 20 mL NMMP, with coiled coils at 30°C for 7 days. The entire culture was extracted by adding an equivalent volume of n-butanol and shaking overnight at 4°C. The mixture was collected and centrifuged at 4°C, after which the top butanol

layer was collected. The crude extracts were dried and weighed, and dissolved in methanol at a concentration of 1 mg/mL for LC-MS analysis.

LC-MS analysis

LC-MS/MS acquisition was performed using Shimadzu Nexera X2 UHPLC system, with attached PDA, coupled to Shimadzu 9030 QTOF mass spectrometer, equipped with a standard ESI source unit, in which a calibrant delivery system (CDS) is installed. The dry extracts were dissolved in MeOH to a final concentration of 1 mg/mL, and 2 µL were injected into a Waters Acquity Peptide BEH C_{18} column (1.8 µm, 100 Å, 2.1 × 100 mm). The column was maintained at 30 °C, and run at a flow rate of 0.5 mL/min, using 0.1% formic acid in H_2O as solvent A, and 0.1% formic acid in acetonitrile as solvent B. A gradient was employed for chromatographic separation starting at 5% B for 1 min, then 5 – 85% B for 9 min, 85 – 100% B for 1 min, and finally held at 100% B for 4 min. The column was re-equilibrated to 5% B for 3 min before the next run was started. The LC flow was switched to the waste the first 0.5 min, then to the MS for 13.5 min, then back to the waste to the end of the run. The PDA acquisition was performed in the range 200–400 nm, at 4.2 Hz, with 1.2 nm slit width. The flow cell was maintained at 40 °C.

The MS system was tuned using standard NaI solution (Shimadzu). The same solution was used to calibrate the system before starting. Additionally, a calibrant solution made from Agilent API-TOF reference mass solution kit was introduced through the CDS system, the first 0.5 min of each run, and the masses detected were used for post-run mass correction of the file, ensuring stable accurate mass measurements. System suitability was checked by including a standard sample made of 5 µg/mL paracetamol, reserpine, and sodium dodecyl sulfate, which was analyzed regularly in between the batch of samples.

All the samples were analyzed in positive polarity, using data dependent acquisition mode. In this regard, full scan MS spectra (*m/z* 100 – 1700, scan rate 10 Hz, ID enabled) were followed by two data dependent MS/MS spectra (*m/z* 100 – 1700, scan rate 10 Hz, ID disabled) for the two most intense ions per scan. The ions were selected when they reach an intensity threshold of 1500, isolated at the tuning file Q1 resolution, fragmented using collision induced dissociation (CID) with fixed collision energy (CE 20 eV), and excluded for 1 s before being re-selected for fragmentation. The parameters used for the ESI source were: interface voltage 4 kV, interface temperature 300 °C, nebulizing gas flow 3 L/min, and drying gas flow 10 L/min. The parameters used for the CDS probe were: interface voltage 4.5 kV, and nebulizing gas flow 1 L/min.

LC-MS based comparative metabolomics

All raw data obtained from LC-MS analysis were converted to mzXML centroid files using Shimadzu LabSolutions Postrun Analysis. The converted files were imported and processed MZmine 2.5.3 [214]. Throughout the analysis, *m/z* tolerance was set to 0.002 *m/z* or 10.0 ppm, retention time (RT) tolerance was set to 0.05 min, noise level was set to 2.0E2 and minimum absolute intensity was set to 5.0E2 unless specified otherwise. Features were detected (polarity: positive, mass detector: centroid) and their chromatograms were built using the ADAP chromatogram builder [215] (minimum group size in number of scans: 10; group intensity threshold: 2.0E2). The detected peaks were smoothed (filter width: 9), and the chromatograms were deconvoluted (algorithm: local minimum search; Chromatographic threshold: 90%; search minimum in RT range: 0.05; minimum relative height: 1%; minimum ratio of peak top/edge: 2; peak duration 0.03 – 3.00 min). The detected peaks were deisotoped (maximum charge: 5; representative isotope: lowest *m/z*). Peak lists from different extracts were aligned (weight for RT = weight for *m/z*; compare isotopic

pattern with a minimum score of 50%). Missing peaks detected in at least one of the sample were filled with the gap filling algorithm (RT tolerance: 0.1 min). Among the peaks, we identified fragments (maximum fragment peak height: 50%), adducts ($[M+Na]^+$, $[M+K]^+$, $[M+NH_4]$, maximum relative adduct peak height: 3000%) and complexes (Ionization method: $[M+H]^+$, maximum complex height: 50%). Duplicate peaks were filtered. Artifacts caused by detector ringing were removed (*m/z* tolerance: 1.0 *m/z* or 1000.0 ppm) and the results were filtered down to the retention time of interest. The aligned peaks were exported to a MetaboAnalyst file. From here, peaks were additionally filtered to keep only peaks present in all three replicates, using in-house scripts. The resulting peak list was uploaded to MetaboAnalyst [216], log transformed and normalized with Pareto scaling without prior filtering. Missing values were filled with half of the minimum positive value in the original data. Heatmaps and volcano plots were generated using default parameters.

Molecular networking

Raw LC-MS data were processed first in MZmine 2as described above, with added steps for MS2 mass detection (polarity: positive, mass detector: centroid, noise level: 0), and MS2 pairing (m/z range 0.05 Da, RT range 0.2 min).. The processed data were then exported to GNPS-FBMN as a .mgf spectra file and a .csv quantification table, with the following parameters: merge MS/MS enabled - spectra to merge across sample, m/z merge - mode most intense, intensity merge mode - maximum intensity, mass deviation - 0.005 or 20 ppm, cosine threshold - 60%, peak count threshold - 0%, Filter rows - only with MS2. The exported files, together with a metadata file describing the samples, were submitted to the Global Natural Products Social Molecular Networking (GNPS) tool for molecular networking [254]. The Feature-Based Molecular Networking (FBMN) workflow [259] was used adopting the default parameters apart from the maximum connected component size changed to 200, and disabling of filtration of peaks around precursor ion mass and peaks in 50Da window. The molecular networking job can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1699fc78a3b84cefb36271cf673d3b83.
Additionally, the data were analyzed using the MS2LDA tool for the identification of likely sub-structures in the extracts based on the obtained fragmentation pattern of the molecules [260]. The default parameters were used for TOF data apart from LDA free motifs being set to 300, and databases for urine and plant motifs being excluded in the analysis. The MS2LDA job can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2886d1b2edc44c31ba693f5eab7cec73. The results obtained from the MS2LDA workflow were mapped to the molecular network previously generated using the MolNetEnhancer workflow [255], and the obtained network was visualized using Cytoscape [133].

Mass spectrometry-based quantitative proteomics

20 µL of dense spore stocks were inoculated in NMMP and grown for 7 days as described above. 1 mL samples were taken after 2 and 7 days. Mycelium was gathered by centrifugation and washed with disruption buffer (100 mM Tris-HCl, pH 7.6, 0.1 M dithiothreitol). The samples were sonicated for 5 minutes (in cycles off 5s on, 5s off) to disrupt the cell wall, and centrifuged at max speed for 10 minutes to collect the proteins. Proteins were then precipitated using chloroform-methanol [217]. The dried proteins were dissolved in 0.1% RapiGest SF surfactant (Waters) at 95°C. Protein digestion steps were done according to van Rooden et al [218]. After digestion, formic acid was added for complete degradation and removal of RapiGest SF. Peptide solution containing 8 µg peptide was then cleaned and desalted using the STAGETipping technique [219].

Final peptide concentration was adjusted to 40 ng/µL with 3% acetonitrile, 0.5% formic acid solution. 200 ng of digested peptide was injected and analysed by reverse-phase liquid chromatography on a nanoAcquity UPLC system (Waters) equipped with HSS-T3 C18 1.8 µm, 75 µm X 250 mm column (Waters). A gradient from 1% to 40% acetonitrile in 110 min was applied, [Glu^{1}]-fibrinopeptide B was used as lock mass compound and sampled every 30 s. Online MS/MS analysis was done using Synapt G2-Si HDMS mass spectrometer (Waters) with an UDMS^{E} method set up as described [218].

Mass spectrum data were generated using ProteinLynx Global SERVER (PLGS, version 3.0.3), with MS^{E} processing parameters with charge 2 lock mass 785.8426 Da. Reference protein database was downloaded from GenBank with the accession number GCA_001278075.1. The resulting data were imported to ISOQuant [220] for label-free quantification. The TOP3 quantification result from ISOQuant was used when further investigating the data.

Supplementary information for Chapter 5

Data S1. Molecular network of all the ions detected in the extracts of MBT27 and derived strains. The relative intensities of the ions are mapped to the nodes as pie charts, and the nodes are labelled by the monoisotopic mass of their precursor ions. The edge thickness represents the cosine score, which indicates the degree of the relatedness of the MS/MS spectra. Available upon request.

Data S2. Molecular network of the ions detected in the extracts of MBT27 and derived strains, as represented in Figure 4C. The relative intensities of the ions are mapped to the nodes as pie charts, and the nodes are labelled by the monoisotopic mass of their precursor ions. The edge thickness represents the cosine score, which indicates the degree of the relatedness of the MS/MS spectra. Available upon request.

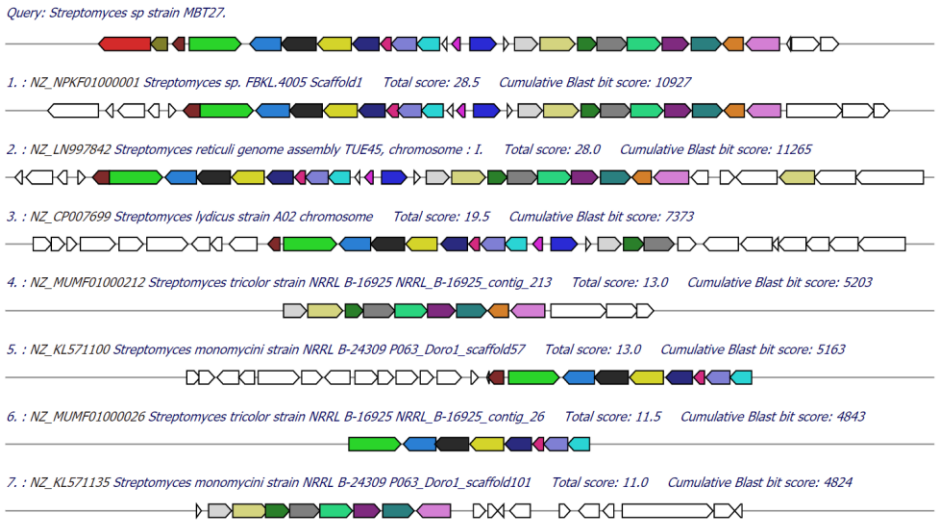


Figure S1. MultiGeneBlast analysis reveals five gene clusters closely related to the *trc* gene cluster. Homologous gene clusters all contained the same set of the genes, except for *Streptomyces lydicus* sp. A02, which a transporter (*trcT2*) and five other genes (*trcKLH4MN*). Homologs of *trcA1* and *trcA2* were also not always detected in genomes in which the cluster was detected on two different contigs.

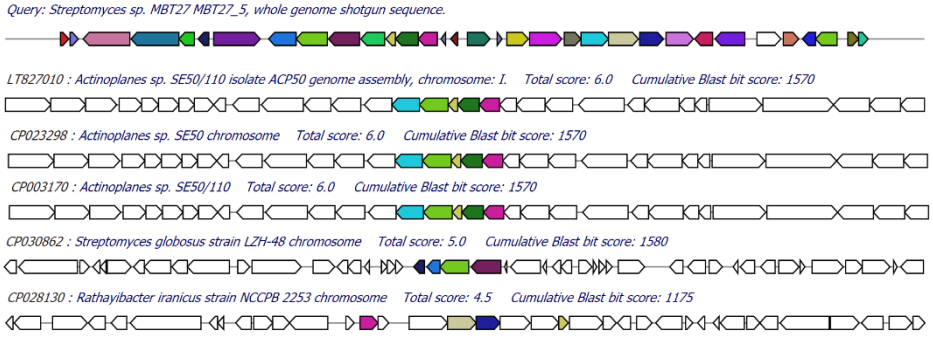


Figure S2. MultiGeneBlast shows several gene clusters distantly related to the *trc* cluster.

```
WP_059250216.1      MSFARRRTAKAVATLFLTSCAALAVTATTAQ--DTHHAAGRSTTGTH-----ALGATTGWQSPA----------PATVQDTTGWQ
WP_046926578.1      MPFARRRPAKALATIVLASCATLTATTTAQA-------AGHTTHRTATRAP-LAQHTTGWQTPA----------PIA-RSTTGWQ
TrcA1               MSFARRSPAKALAATLLVAGATLTVTAQTAQ----------ATTGWQSPAPAAAPATTGWQAPAPDTTGWQNPGPAAARATTGWQ
TrcA2               MSFARRSPAKALAATLLVAGATLTVTAQTAQ----------ATTGWQSPAPAAAPATTGWQAPA----------P----DTTGWQ
WP_030024344.1      MSFARRRPVKALAAISLAACAALTVTTTTAQTADTTTVAAPTTTGWQAPAP-TIQHTTGWQAPA----------P-----TTGWQ
                    *.**** ..**:*:  *.: *:*:.*: :           :*            *****:**          *     *****
```

Figure S3. Alignment of all detected precursors shows conserved TTGWQ motif. All predicted precursors varied between 61 and 75 amino acids in length. Alignment of the sequences showed a conserved N-terminal sequence, combined with a more varied C-terminal sequence. However, despite the variation, between two and four repeats of the TTGWQ motif were consistently detected.

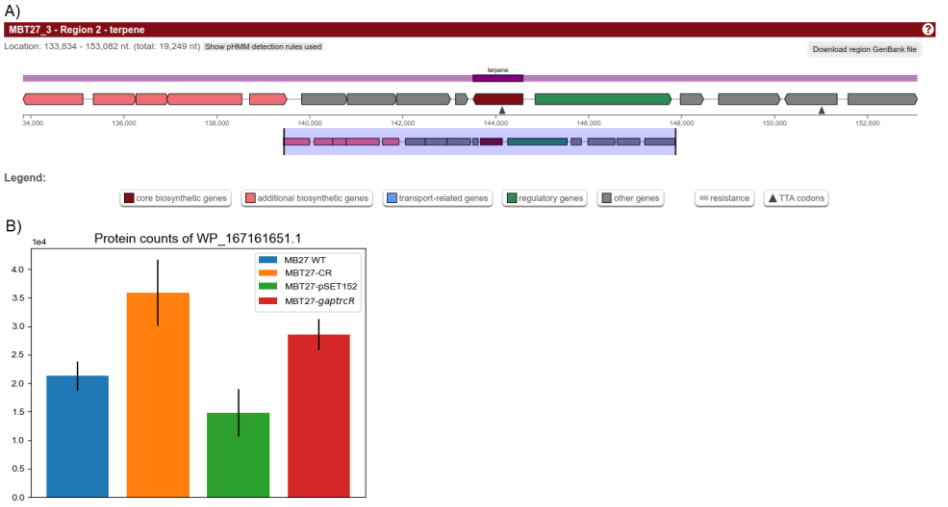


Figure S4. Protein levels of a differently expressed regulator. A regulatory protein encoded by a terpene BGC (A) is detected at significantly higher levels in the proteome of MBT27-CR and of MBT27-*gaptrcR* compared to the proteome of the control strains (B).

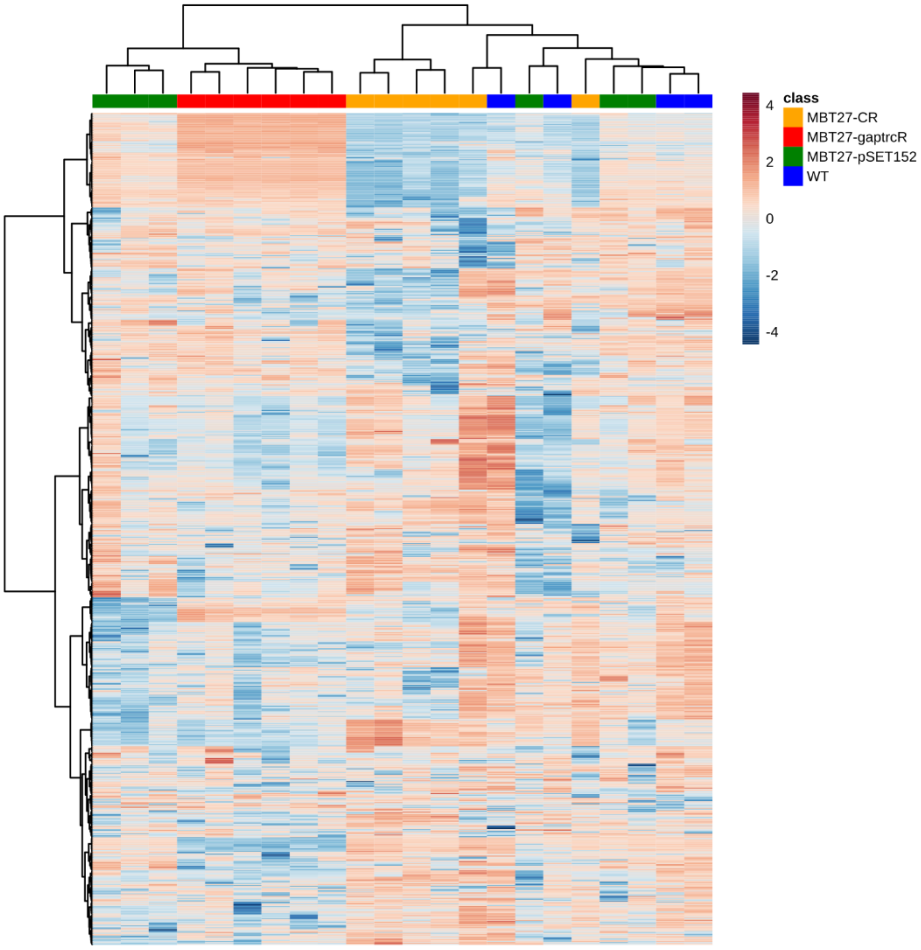


Figure S5. Heatmap of mass features detected by LC-MS and processed with MZmine shows large differences among the extracted strains. Each row represents a single mass feature and each column represents a single extract, while the colour scale indicates the $\log_{10}$-scaled intensity of the mass features for each extract. Large differences are even seen among replicates, arising from the different biological replicates of the modified strains which were extracted.

Table S1. Primers used in this study.

Primer name	Primer sequence
trcR_F	gatc GGTACC CAT ATGCCGGAGAACCTGGGC
trcR_R	gatc TCTAGA AAGCTT CCGTTGCACTACATGGTCGAAGCC
trc_LF_F	cgta GAATTC GCGACTGACAGCACCACTGG
trc_LF_R	cgta TCTAGA CACCGATCCACACCAGTGG
trc_RF_F	cgta TCTAGA GGCCGTAGGGACAATCAATCACC
trc_RF_R	cgta AAGCTT CACTGTCAGCCACGCAATGATGG
trc_del_check_F	GGTGTCGAAATCGGACATGG
trc_del_check_R	CCTTTGGCAGGTCGTCGCTGACC

Table S2. Plasmids used in this study.

Plasmid	Description	Reference
pSET152	Integrative *E. coli* / *Streptomyces* shuttle vector.	Bierman *et al.* [199]
pWHM3	Unstable *E. coli* / *Streptomyces* shuttle vector with high copy number; used for homologous recombination	Vara *et al*. [198]
pUWLCRE	Unstable *E. coli/Streptomyces* shuttle vector containing the Cre recombinase enzyme. behind a constitutive promoter.	Fedorshyn *et al*. [200]
pAK9	pSET152 containing *trcR* behind GAP promoter from *S. coelicolor* (SCO1947).	This work.
pAK10	pWHM3 containing regions flanking the *trc* cluster.	This work.

Table S3. Strains used in this study.

Strain	Description	Reference
MBT27 wildtype	*Streptomyces* sp. MBT27, previously isolated from the Qingling mountains.	Zhu *et al.* [229]
MBT27-pSET152	*Streptomyces* sp. MBT27containing an empty pSET152 (empty vector).	This work.
MBT27-*gaptrcR*	*Streptomyces* sp. MBT27containing pAK4.	This work.
MBT27-CR	*Streptomyces* sp. MBT27in which the **c**entre region from *trcA1* to *trcA2* was **r**eplaced by the apramycin resistance cassette aac3(IV).	This work.
MBT27-IFD	MBT27-CR from which the apramycin cassette was removed with the Cre-Lox system.	This work.
MBT27-IFD-pSET152	MBT27-IFD containing an empty pSET152 (empty vector).	This work.
MBT27-IFD -*gaptrcR*	MBT27-IFD containing pAK9.	This work.

Table S4. Overview of UniProt accessions of representatives of radical SAM subfamilies used to classify radical SAM proteins from MIBiG. Adapted from Holliday *et al* [100].

Class name	Uniprot Accession IDs
tRNA wybutosine-synthesizing	Q57705
spectinomycin biosynthesis (SpcY-like)	Q9S1L5
antiviral proteins	O70600
AviX12-like	Q93KV6
lipoyl synthase like	P60716
FeMo cofactor biosynthesis protein	P11067
DesII-like	Q9ZGH1
anaerobic coproporphyrinogen-III oxidase like	A0A060PWX2, V0VQG0, P9WP73, P52062, P32131, Q796V8, C6FX53, Q9FB10, I3NN68
BATS domain containing	P12996, Q58195, Q9X0Z6, Q46E78, P30140, C6FX51, Q6PSL4
7-carboxy-7-deazaguanine synthase like	O31677, A0A0H3KB22, O54060
PLP-dependent	Q841K7, A4J6G2, Q9XBQ8, P39280
methylthiotransferase	Q9WZC1, Q96SZ6, P0AEI4, P54462
F420, menaquinone cofactor biosynthesis	Q58826, Q57888, Q9XAP2, Q5SK48
organic radical-activating enzymes	Q84F14, P0A9N8, O87941, Q8GEZ7, P0A9N4, P39409
methyltransferase	P36979, Q9FBG4, Q58036
spore photoproduct lyase like	A4IQU1, Q97L63
elongater protein-like	Q02908
B12-binding domain containing	Q3ME29, Q2MFI7, A0A095DNL6, Q1Q0N1, P26168, A8R0J7, A8R0J8, D2KTX8, F8JND9, F8JNE0, Q58275, Q8GHB6, O24770, Q70KE5, B3QHD1, B9ZUJ4, D2KTX6, Q60AV6, Q5IW50, A8R0J3, Q56184, Q50258, Q6QVU0, Q8KCU0, C0JRZ9, Q8KBK9
SPASM/Twitch domain containing	A0A1I5E523, P69848, A1B2Q7, D0QZJ5, Q841K9, B8J367, P27507, Q8RAM6, Q8G907, A0A0E2Q059, Q46CH7, Q0TTH1, O31423, Q51741, Q53U14, P9WJ79, Q9X758, P71011, A0A095EC78, E5KJ95, C2TQ82, Q6E3K8

5

6

General discussion and conclusion

Alexander M. Kloosterman
Marnix H. Medema
Gilles P. van Wezel

Introduction

Ribosomally synthesized and post-translationally modified peptides (RiPPs), like many other natural products, comprise a dazzling array of chemical diversity [42, 48]. The simple biosynthetic logic – a precursor gene is translated, and the product is modified and cleaved – results in many different possible structures, depending on the precursor peptide and modifications applied to it. They are divided in more than 40 different subclasses, and the list of subclasses keeps steadily growing. Their functions are equally diverse, and include quorum sensing, acting as enzyme co-factors, roles in cellular development, mediating host-microbe interactions, but also the much sought-after antibacterial and antifungal properties that would make them interesting for clinical applications [261].

Next-generation sequencing efforts surprisingly revealed that the capacity of bacteria to produce natural products had been grossly underestimated. This has led to a revolution in drug discovery based on the efficient mining of the rapidly growing genome sequence data [26, 262]. Numerous tools and databases have been developed to explore, compare and catalogue biosynthetic gene clusters (BGCs) and their chemical products [29, 185, 228, 263, 264]. For most of the biosynthetic gene clusters (BGCs) discovered, the chemical compounds they produce are not yet known. These so-called cryptic BGCs represent a vast potential for new natural products with potentially interesting bioactivities. Even so, the BGCs that are currently easily detected are in some ways similar to characterized ones, since their detection is based on the identification of conserved protein domains [26, 39, 40, 228].

RiPPs present an interesting case when it comes to genome mining, as there is no single genetic marker that identifies them all, other than the prerequisite of an ORF that encodes a small peptide. Although some RiPP subclasses overlap on a genetic level, most require bioinformatic rules aimed at subclass-specific genetic markers. There is still plenty of room for innovative genome mining strategies aimed at identifying novel RiPP subclasses, each of which further charts undiscovered genetic space [52, 85, 88, 89, 114, 265, 266]. In this thesis, we have explored such strategies which prioritize novelty at the

cost of fidelity, with the primary aim being the identification of novel RiPP families. The main pipeline that was developed, decRiPPter, can function as a platform for explorative RiPP genome mining. In contrast to most tools developed for high-confidence RiPP genome mining, decRiPPter relies more heavily on user settings, and present several options for trade-offs between confidence and novelty. Using this tool, the pristinin BGC was discovered, which encodes a novel class V lanthipeptide.

Machine learning paves the way for class-independent precursor identification

Machine-learning-based and neural-network-based classifiers have risen in popularity over the last decades as tools to process and classify massive datasets with large numbers of features. The large databases of genome sequences now available provide many opportunities for these classifiers to exploit their high precision for the benefit of genome mining. Specifically for RiPPs, the identification of the precursor gene presents an interesting challenge for machine-learning classifiers. Precursor genes are not easily recognized by similarity-based methods, and are frequently missed by automatic gene annotation algorithms due to their small size. Several classifiers have been developed for RODEO that supplement classical RiPP genome mining by identifying precursors of known classes [45, 55, 72-74, 86], and several more tools have been reported for standalone precursor identification [88, 89, 114].

Detection of precursor peptides forms the core of decRiPPter (Chapter 3), and determines which genomic regions will be further investigated. As such, decRiPPter is the first reported genome mining tool that uses the detection of precursors, rather than of enzymatic domains as the basis for the identification of novel RiPP subclasses. Analysis of 1,295 *Streptomyces* genomes resulted in the discovery of 42 candidate RiPP families after manual curation. All of these families are specified by BGCs that are characterized by a promising combination of precursor, transporter, biosynthetic, regulator and peptidase genes, typically organized in a single operon-like genomic structure. While some of the BGCs contain genes previously reported in known RiPP BGCs, most of the biosynthetic genes encode enzymes not previously associated with RiPP biosynthesis,

suggesting that many more RiPP modifying enzymes exist than currently known. Characterization of these enzymes could then be translated to new RiPP genome mining rules for tools like BAGEL [62] and antiSMASH [39], standardizing their detection. Experimental investigations into one of the families showed that it did indeed encode a novel RiPP, namely a lanthipeptide, pristinin A3, that is modified by a newly discovered set of modifying enzymes. How many more of the 42 families actually specify RiPPs needs to be validated experimentally. However, if even half of these candidates encode actual RiPPs, it would represent a sizable contribution to expanding the RiPP chemical space.

Remarkably, the wide variety of precursor sequences of many different classes can be adequately covered by the SVM-based classifier of decRiPPter. Apparently, there are certain combinations of features that are typical of RiPP precursors regardless of class. These include the enrichment of certain amino acids, like cysteine, serine and threonine, which are often modified in known RiPPs, but also frequently found in the candidate RiPP BGCs. In addition, arginine residues are particularly rare across RiPP precursors. An evolutionary link between different RiPP classes could explain these conserved features, but is made unlikely by the large disparity in precursors and modifying enzymes. Cysteine, serine and threonine residues do have oxygen- and thiol-groups, respectively, making them easier to modify. This chemical property could drive the evolutionary process towards precursor peptides containing certain residues, even if they have evolved independently. If the latter is the case, it would explain why feature-dependent classifiers that focus on amino acid frequencies are so effective at detecting precursor peptides of many different classes, and it would suggest that many more RiPP classes can be detected by them.

A difficult challenge when applying these classifiers to a large genomic space is the number of false positives. The sheer number of candidates (71 million) as opposed to the number of expected precursor genes (~6500 if each genome encodes five RiPP precursors) makes it so that even a false discovery rate of 1% would result in many more false positives than true positives. Comparisons with other machine-learning-based classifiers revealed similar numbers of hits for those, meaning that this would be a general issue. As the

number of characterized precursors increases, and therefore the training set improves, the accuracy of newer classifiers may improve as well. Alternatively, restricting the set of precursors to those for which at least two different classifiers reach a consensus would reduce the number of hits substantially, as the overlap between the three studied methods is relatively low. However, it is questionable whether precursor identification itself can become reliable enough for precursor-based RiPP genome mining without considering their genetic context. The false discovery rate would have to drop substantially while still covering the wide variety of precursor sequences. Until then, using the genetic context as shown in Chapter 3 is a viable alternative to filter the identified precursors down to a more manageable set.

Another way to filter the predicted precursors without considering genetic context is to prioritize precursors with multiple core regions. Having multiple copies of the same core region allows for the efficient production of several RiPP variants, while only needing a single leader peptide. A similar pattern was also identified in the RiPP candidate discussed in Chapter 5. These repeats are found often in eukaryotic RiPPs [51, 239, 267], and could provide a handhold for their identification without prior knowledge of their primary sequence. If these patterns occur as exact copies, their identification would be algorithmically straightforward, by taking subsets of the sequence and finding exact matches of that sequence elsewhere in the peptide. If more variation of the pattern occurs, such as in thiovarsolins, identification of these patterns would have to be based on a local alignment algorithm, such as BLAST, or a motif discovery tool such as MEME [116, 268]. Alternatively, the presence of a repeated pattern can be used as a feature in a future iteration of the classifier, so that it is taken into account during precursor prediction itself. Flagging precursors in which these patterns can be found can be used to remove many false positives, albeit at the cost of removing RiPP families which do not contain these patterns. Their presence could therefore be used as an imperfect bioinformatic handle to fine-tune precursor-based RiPP genome mining.

Prioritizing novel RiPP BGCs from the genetic context

decRiPPter uses the genetic context of predicted precursor genes to prioritize candidate BGCs. The filtering process exemplifies the trade-off between confidence and novelty, and can be set up according to user preferences. At loose conditions (e.g., mild filtering), most known RiPP BGCs are left unfiltered, but the number of false positives is estimated upwards of 84.4%, making the dataset too large to manually process. It is likely that there are still many RiPP BGCs among this dataset, which is also highlighted by the promising candidate discussed in Chapter 5, but without additional filters, selecting a suitable candidate can become difficult. In order to simplify this, the HTML-based output allows a user to browse the results. In addition, the entire set can be filtered with additional criteria of interest, such as specific biosynthetic domains, or a specific number of transporters, proteases or regulators in or nearby the precursor gene. The resulting set can then be manually investigated and a BGC of interest can be selected. Expanding the output filtering options with additional parameters, such as specific motifs within precursors, would help users browsing this large dataset and find the exact sort of BGC they are looking for.

The strict filter applied is a middle ground between confidence and novelty. On the one hand, it is permissive in the sense that many different domains are considered as possible RiPP associated enzymes and proteins. On the other hand, it is restrictive in the sense that genes for a peptidase, regulator and transporter are all required. In theory, these encode peptidases for precursor cleavage, a dedicated transporter module, and a cluster-specific regulator. Many known RiPP BGCs do not contain all of these genes, and instead their encoded pathway and products are regulated, transported and cleaved by proteins encoded elsewhere in the genome. As a result, the remaining candidate BGCs are promising, and the false positive rate was lower than with the mild filter (estimated between 46.7 and 73.0%), although many known RiPP BGCs are filtered.

Several other methods for prioritizing gene clusters of interest can be envisioned, which would each represent a different trade-off in confidence and

novelty. Integrating these into decRiPPter would further expand the possibilities for more fine-tuned search strategies in which several criteria can be combined. The tool for one of these, RRE-Finder, was discussed in Chapter 2. RiPP Recognition Elements (RREs) are involved in the precursor recognition of many different RiPP classes, and could function as a class-independent bioinformatics handle for RiPP discovery. With RRE-Finder, RREs can be detected at a faster rate than with HHPred, allowing for the analysis of large amounts of queries. Exploratory mode of RRE-Finder, which is based on HHPred, detected several novel RRE-enzyme fusions in the UniProt database, which could lead to the discovery of novel RiPP modifying enzymes. Unfortunately, the false discovery rate of exploratory mode is higher than for precision mode, which makes it questionable which of the newly discovered RRE-enzyme fusions would be worth investigating. This disadvantage can be mitigated by imposing other mild criteria of decRiPPter, i.e. a predicted precursor gene nearby, one or two biosynthetic domains in an operon-like gene organization, and not being part of the core genome. Integration of RRE-Finder therefore would be a valuable addition to the decRiPPter pipeline, and help increase the confidence for both tools.

RRE-Finder itself could be further improved by using a machine-learning classifier for the detection of RREs. Like RiPP precursor peptides, RREs are generally no longer than 120 amino acids long. A candidate sequence of this length can be used completely as an input vector in a neural network, as is done in NeuRiPP, without having to select specific features. This approach would allow for detection of discrete RREs by using part of the sequence, e.g. the N- or C-terminal regions, as raw input for the network. These classifiers might be able to better distinguish between regulators and RREs, as they can recognize more complicated patterns than only secondary structure. A possible discriminatory feature are the sequence residues that are known to interact with the precursor peptide. Several of these residues have been shown to co-evolve with the precursor peptide, and likely stand out from a sequence-based point of view when compared to similar domains found in regulators. Further research is required to determine if machine-learning classifiers are indeed suitable for the detection of RREs.

Insights into RiPP evolution guide discovery of novel RiPPs

Understanding how different RiPPs have evolved can provide useful insights for the prioritization of RiPP BGCs, especially if these principles are class-independent. For secondary metabolism in general, it has been hypothesized that their enzymes have evolved from primary metabolism enzymes. An example of this can be seen for polyketide synthetases (PKSs), which descend from fatty acid synthetases, but have diverged to take in different substrates, and apply extra tailoring [269]. This property has been used earlier to mine for BGCs in EvoMining [160, 161]. By searching for enzymes that have evolved from primary metabolism enzymes, many BGCs of known classes like NRPS and PKS, but also of novel classes, can be identified.

Interestingly, the RiPP candidates prioritized by decRiPPter included several BGCs that encode proteins previously identified in a different context. HypD, HypE, MauD and MauE are thought to be involved in protein maturation, by creating crosslinks or modifying specific amino acids [182, 183]. These proteins could have easily evolved towards modifying small peptides rather than proteins, and could thus have become RiPP-modifying enzymes. A similar example was recognized earlier: QhpD, an enzyme that catalyzes the synthesis of a thioether bond in a protein, and radical SAM enzymes involved in thioether crosslink formation in sactipeptides and ranthipeptides, show moderate similarity [55, 270]. Protein modification is a widely occurring phenomenon in all branches of life, and it is possible that more RiPP modifying enzymes evolved from them. An approach similar to EvoMining, using protein-modifying enzymes as a query, could aid in the identification of more of these RiPP subclasses.

Another sizable contributor to RiPP BGC biodiversity is the occurrence of gene swaps. The genes for YcaOs [95], rSAMs [96], lanDs [194], for example, are encoded by BGCs of several RiPP subclasses. The newly reported lanthipeptide class V further contributes to this list, as its BGC contains elements from both linaridins [271] and thioamitides [52, 94, 272], further suggesting that gene swaps contribute significantly for RiPP diversity. An automated procedure might be able to prioritize genes present in many RiPP-like clusters, even if they were not previously functionally associated with RiPPs before. If, from a

candidate RiPP BGC, a gene or set of genes can be detected in BGCs of other candidate RiPP families as detected by decRiPPter, this would make it more likely that the gene product is involved in RiPP maturation. This can be seen to some extent with the *mauE* and *mauD* genes, which are present in three different RiPP families, and also with the core enzymes of the novel lanthipeptide subclass, described in Chapter 4. In a simple form, this procedure can be automated by searching for biosynthetic domains that are seen among several different RiPP families. A more sophisticated pipeline could involve the usage of CORASON to identify gene islands widespread across many different RiPP-like contexts. Successful identification of these islands would help prioritize RiPP modifying enzymes, and by extension, RiPP families.

Examples of novel RiPPs and their classification

To validate decRiPPter's capabilities to detect novel RiPP classes, we selected two BGCs of different candidate families to experimentally characterize. One of these encodes a novel lanthipeptide, pristinin A3, containing the classical thioether bridge, a C-terminal aminovinylcysteine and serine-to-alanine conversions (Chapter 4). Importantly, two candidate genes appear likely candidates for the formation of the thioether bridge. Their presence in many genetic contexts shows that this class is widespread across several taxonomic clades, and that these genes are excellent candidates to add to the rulesets of high-confidence RiPP genome mining tools. Furthermore, lanthipeptides frequently possess antimicrobial activity [273, 274], so the discovery of a novel class of these could in time lead to the discovery of novel antibiotics.

Another promising BGC (Chapter 5) has many features that suggest it specifies a RiPP. This BGC contains many genes that encode enzymes previously associated with RiPP biosynthesis, like an rSAM and an ATP-grasp ligase. Despite this, the BGC was not directly recognized by other RiPP genome mining tools, and encodes several more predicted modifying enzymes that were not recognized. The repeated, conserved patterns observed in the precursor peptides are likely multiple core regions. Several masses were detected exclusively when the gene cluster was activated, which were no longer present when the gene cluster was inactivated. These masses were within 200 Da of the

mass of the predicted core peptide. Unfortunately, none of the masses could be matched to the core peptide, and it remains unclear whether any of the masses are directly derived from it. It seems likely that the many predicted enzymes extensively modify the core peptides, meaning more sophisticated analytical chemistry is required to relate the structure to the peptide. Furthermore, heterologous expression of the BGC could help prioritize which masses are exclusively derived from the BGC, and not produced due to any secondary effects, like the activation of another BGC.

The two BGCs described in this work both contain genes that have homologs encoded by BGCs of other RiPP subclasses. Despite this, they both would still likely specify members of a novel RiPP class, due to a unique combination of modifications or novel enzymatic machinery that installs it. In general, however, the discovery of RiPP classes that are produced mostly by a combination of modifying enzymes already known makes their classification more complicated. The consensus for classification of RiPPs is based on designating modifications as core or accessory, and determining which core modifications are required for one RiPP family [42]. This methodology is becoming more and more difficult to uphold. Given that modifications can be swapped between different RiPP families, which one is considered a core modification and which one is considered an accessory one is context-dependent. If the lanthionine bridge of pristinin A3 is considered the core modification, as for other lanthipeptides, then all other modifications would be considered secondary. These include the formation of dehydrated serine residues, which are considered a core modification in linaridins.

As a result, what makes up a novel RiPP class becomes somewhat arbitrary. Lipolanthines, for example, are considered a standalone RiPP class, but they are clearly very related to other lanthipeptides [80]. By contrast, glycosylated lanthipeptides are not considered their own class. Since the definition of a RiPP class determines the rules for genome mining of that class, we should take care not to restrict ourselves too much with these definitions. Many more interesting RiPP variants can be found by alleviating the strictest of rules. Rather than focus on the identification of novel RiPP classes, which could be considered arbitrary, perhaps the priority should be the identification of

RiPP-associated reactions and their corresponding modifying enzymes. The RiPP classes can be considered examples in which specific modifications have been found combined. But any RiPP-associated enzyme could arguably lead to the discovery of new RiPP classes and variants, whether core or accessory.

Conclusion

Natural products and their BGCs come in many shapes and sizes, resulting in a rich diversity to explore. In this thesis, we have explored methods aimed at finding novel types of natural products, specifically novel RiPP subclasses. The biosynthetic logic of a RiPP can be made up of many different precursors and modifying enzymes. There are several features, however, which can be exploited for their detection. RiPP BGCs should always encode a precursor peptide, providing a handhold for identification with machine-learning classifiers. Encoded modifying enzymes in the BGC should be capable of recognizing the precursor peptide, which can be exploited through the detection of RREs or through their association with other RiPP classes. We have combined these methods to prioritize many different gene clusters, and illustrated that one of these gene clusters indeed specified a novel type of lanthipeptide (pristinin). The pipeline can be expanded further in many ways, including the integration of RRE-Finder, new precursor classifiers, or detection methods using evolutionary principles, which will help expanding the large chemical diversity of this class of natural products.

Nederlandse samenvatting

Secundaire metabolieten als bron van antibiotica

Planten, dieren en micro-organismen produceren een grote diversiteit aan metabolieten en andere natuurstoffen [1, 222]. Sommige van deze moleculen zijn essentieel voor elk organisme. Suikers, vetten en nucleotiden, bijvoorbeeld, vormen de bouwstenen voor het leven en worden de primaire metabolieten genoemd. Onder secundaire metabolieten worden alle overige natuurstoffen verstaan. Meestal bieden deze moleculen organismen een voordeel onder specifieke omstandigheden of helpen ze de communicatie met andere soorten. Dit leidt tot een enorme chemische diversiteit aan moleculen. Tegelijkertijd zijn de meeste varianten relatief zeldzaam, vooral vergeleken met primaire metabolieten, aangezien ze alleen in specifieke niches voordeel kunnen bieden [2].

De precieze functie van secundaire metabolieten verschilt erg van molecuul tot molecuul, maar vele ervan werken als moleculaire wapens, die in staat zijn bacteriën, schimmels of virussen te doden of hun groei te remmen [2]. Antivirale en antibacteriële metabolieten kunnen bescherming bieden tegen infecties, en hun productie door hogere organismen is daarmee geen verassing. Toch zijn het de micro-organismen zelf die het grootste deel van deze metabolieten produceren, vermoedelijk als wapens en bescherming tegen andere micro-organismen. Daarnaast kunnen sommige van deze stoffen dienen als immunosuppressiva of de groei van tumoren remmen. Het zijn deze eigenschappen die zulke moleculen interessant maken voor gebruik in de klinische omgeving, met toepassing in de behandeling van onder meer infectieziekten, tumoren of immuunziekten. Sinds de ontdekking van penicilline in de jaren `20 [7] is men erin geslaagd een groot aantal verschillende antibiotica te isoleren uit bacteriën en schimmels, die nog steeds in de kliniek gebruikt worden. Dit was een grote revolutie voor de geneeskunde: bacteriële infecties die eerder een doodsvonnis betekenden, konden opeens effectief worden behandeld met een simpele kuur. Het grote succes van antibiotica heeft geleid tot hun wereldwijde gebruik en heeft talloze mensenlevens gered [227].

Helaas zit er een keerzijde aan het wijdverbreide gebruik van antibiotica. Bacteriën in de natuur worden steeds meer blootgesteld aan antibiotica die in het milieu terecht zijn gekomen. Resistente varianten van deze bacteriën hebben daardoor een evolutionair voordeel. Dit heeft als gevolg dat genen en mutaties die resistentie verschaffen tegen antibiotica zich steeds wijder verspreiden [14]. Multiresistente pathogenen zoals MRSA (methicillin-resistant of multi-resistant *Staphylococcus Aureus*) en MDR-TB (multi-drug resistant tuberculosis) veroorzaken infecties die moeilijk te behandelen zijn en ons technologisch terugbrengen naar de tijd van voor de antibiotica [17]. Een rapport van O'Neill, opgesteld in opdracht van de Britse regering, voorspelt dat in het jaar 2050 meer dan 10 miljoen mensen jaarlijks zullen overlijden aan ziektes veroorzaakt door multiresistente bacteriën, meer dan aan de gevolgen van kanker [21]. Om deze scenario's te voorkomen zijn nu meer dan ooit nieuwe soorten antibiotica nodig, waartegen nog geen resistentie is ontwikkeld.

De meeste antibiotica zijn ontdekt in de jaren `50 en `60 van de vorige eeuw, een periode waar vaak naar wordt verwezen als de Gouden Eeuw voor ontdekking van antibiotica. De methoden waren conceptueel simpel: de potentiële producenten van antibiotica, meestal bacteriën en schimmels, werden op grote schaal en in diverse groeicondities opgekweekt en chemisch geëxtraheerd. Het extract werd daarna toegevoegd aan groeiende bacteriën, om te zien of de groei daarvan beïnvloed werd. Als dit het geval was, konden de verantwoordelijke stoffen gezuiverd worden uit het extract, en onderzocht worden op hun potentie als nieuw antibioticum [158]. Veel van deze stoffen werden geïsoleerd uit Actinomyceten, een fylum van bacteriën, en met name uit het genus *Streptomyces*. De groei van streptomyceten lijkt op die van schimmels: in de aarde vormen ze netwerken van langwerpige hyfen, het bacteriële equivalent van schimmeldraden. Als de voedselbronnen opraken rondom dit netwerk, begint de tweede levensfase van de streptomyceet. Hierin breekt hij zijn eigen netwerk van cellen af en gebruikt de bouwstoffen om sporen te vormen, die weer verspreid kunnen worden, waarna deze tot een nieuw netwerk kunnen uitgroeien [32]. Streptomyceten blijken uitzonderlijke producenten van secundaire metabolieten: meer dan twee derde is afkomstig van dit genus [158].

Hoe succesvol de methode voor het vinden van nieuwe antibiotica andere medicijnen ook was, de laatste decennia is het aantal nieuwe antibiotica dat hiermee gevonden is drastisch gedaald. Dit komt mede door de herontdekking van veel antibiotica: veelvoorkomende antibiotica wordt telkens opnieuw gevonden, wat de ontdekking van nieuwe en meer zeldzame antibiotica belemmert. Onderzoek wordt daardoor steeds minder rendabel voor de farmaceutische industrie, waardoor er minder wordt geïnvesteerd en het aantal ontdekkingen nog sneller daalt. Om het rampscenario van snelle toename van resistentie en een gebrek aan goede medicijnen te voorkomen is duidelijk dat er nieuwe methodes nodig zijn voor de ontdekking van nieuwe antibiotica [24, 25].

De analyse van de genomen van antibioticaproducenten: een nieuwe revolutie

Een grote doorbraak werd bereikt toen de sequentie van het DNA van *Streptomyces coelicolor*, een modelorganisme voor alle streptomyceten, werd bepaald [27]. Uit het genoom bleek dat deze bacterie in staat was tot het produceren van wel dertig verschillende secundaire metabolieten, terwijl er toen nog maar enkele uit deze stam gezuiverd waren. Het aantal producten dat door één stam geproduceerd kan worden, wordt bepaald aan de hand van het aantal biosynthetische genclusters (BGCs) dat gevonden wordt [26]. Een BGC is een verzameling genen die naast elkaar op het genoom liggen, en die vertaald worden in eiwitten die allen betrokken zijn bij de productie van één groep secundaire metabolieten, of zelfs één specifiek metaboliet. Nu het bepalen van de DNA-sequentie aanzienlijk goedkoper is geworden door de ontwikkeling van nieuwe technieken, worden steeds meer genoomsequenties bepaald, en steeds meer BGCs gevonden [264, 275]. Van de meeste hiervan is het product niet bekend. Deze worden ook wel de cryptische BGCs genoemd [31], en zij representeren een potentiële bron van nieuwe secundaire metabolieten en dus nieuwe antibiotica.

Nieuwe BGCs kunnen gevonden door te zoeken naar genen die lijken op genen uit al gekarakteriseerde BGCs. Hoewel de meeste secundaire metabolieten grote chemische verschillen tonen, zit er vaak wel overlap in de

manier waarop ze geproduceerd worden. Deze overlap in productiemethoden leidt ook tot overlap op genetisch niveau, waarvan gebruik gemaakt wordt door software die BGCs in genomen detecteert. De beperking van deze methode is dat er dus altijd een overlap moet zitten tussen bekende BGCs en nieuw gedetecteerde BGCs, en dat compleet nieuwe klassen niet gevonden zullen worden. Als zulke klassen bestaan, zullen we ze alleen kunnen vinden door nieuwe methoden toe te passen om BGCs te detecteren, waarbij we buiten gebaande wegen moeten treden [276].

Een subklasse van metabolieten die goed geschikt lijkt voor het vinden van nieuwe varianten, is die van de ribosomaal gesynthetiseerde en posttranslationaal gemodificeerde peptiden (RiPPs) [42, 48]. RiPPs zijn chemisch divers, maar delen een bepaalde biosynthetische logica [44]. De basis van een RiPP wordt gelegd door een klein eiwit (vaak korter dan 100 aminozuren), dat net als andere eiwitten door het ribosoom wordt geproduceerd. Dit eiwit, ook wel de precursor genoemd, wordt vervolgens uitgebreid gemodificeerd door andere eiwitten waarvan de coderende genen ook in het BGC aanwezig zijn. Na modificatie wordt er een groot deel van afgeknipt, en het complete product, de RiPP, wordt geëxporteerd. Opvallend aan RiPPs is de grote diversiteit van hun producten. Zowel de sequentie van de precursor, als de modificaties kunnen sterk uiteenlopen, en beiden bepalen de structuur van het uiteindelijke product. Bovendien hebben veel RiPPs, net als andere secundaire metabolieten, ook antibacteriële of antivirale eigenschappen.

RiPPs worden onderverdeeld in subklassen, die elk hun eigen specifieke modificaties hebben. Zo bevatten lanthipeptiden thioetherbruggen tussen cysteïnes en serines of threonines [47] en staan lassopeptiden bekend om hun structuur die een knoop vormt [45]. De BGCs die horen bij de verschillende subklassen coderen elk voor verschillende modificerende enzymen en tonen grote verschillen. Dit heeft gevolgen voor de zoekstrategie op genetisch niveau. Over het algemeen geldt dat met informatie van een BGC wel andere BGCs van die subklasse gevonden worden, maar geen BGCs van een andere subklasse. Desalniettemin worden er nog wel vaak nieuwe subklassen ontdekt: in de afgelopen zes jaar is het aantal bekende subklassen verdubbeld van 20 naar 40 [42, 48]. Een zoekstrategie voor BGCs van RiPPs die juist nieuwe subklassen

zoekt lijkt daardoor veelbelovend, maar tot nog toe bestaan zulke strategieën niet. In het kader van Syngenopep, een onderzoeksvoorstel gericht op de ontdekking van nieuwe antimicrobiële peptiden, is hier verder onderzoek naar gedaan. Juist door nieuwe methoden te ontwikkelen voor de detectie van RiPP BGCs, die onafhankelijk zijn van de RiPP subklasse, wordt er gericht op de ontdekking van nieuwe subklassen en dus ook nieuwe soorten peptiden. Hiervoor wordt onder andere gebruik gemaakt van kunstmatige intelligentie en exploratief genetisch onderzoek. Het resultaat van dit onderzoek staat in dit proefschrift gepresenteerd.

Identificatie van nieuwe RiPP-subklassen via detectie van de precursors

De meeste RiPP-subklassen worden gekarakteriseerd door sterk uiteenlopende modificaties en precursoreiwitten. Desalniettemin zijn er elementen die overlappen tussen verschillende RiPP-subklassen, waarvan gebruik gemaakt kan worden voor de detectie van nieuwe subklassen. De belangrijkste hiervan is ongetwijfeld het gen dat codeert voor de precursor, die de basis legt voor het uiteindelijke product. Omdat de sequentie van de precursors zo sterk varieert, wordt er steeds meer gebruik gemaakt van kunstmatige intelligentie om precursors van andere eiwitten te onderscheiden. Hierbij worden niet alleen de sequenties van de precursors vergeleken, maar ook berekende eigenschappen zoals lading, hydrofobiciteit, lengte en frequentie van verschillende aminozuren. Verschillende modellen zijn al eerder gerapporteerd, zoals NeuRiPP en NLPPrecursor [88, 89]. In Hoofdstuk 3 wordt het model van **decRiPPter** (Data-driven Exploratory Class-independent RiPP TrackER) beschreven. Dit model is gebaseerd op een Support Vector Machine (SVM) en kan precursors identificeren van vele verschillende subklassen, soms ook als deze niet in de trainingsset voorkomen. Dit suggereert dat er eigenschappen zijn die deze precursors gemeenschappelijk hebben, ongeacht de subklasse waar ze toe behoren. Dat impliceert dat de precursors van compleet nieuwe subklassen ook met dit model te vinden moeten zijn.

DecRiPPter borduurt voort op het bovenstaande idee om nieuwe RiPP-subklassen te vinden. Het is daarmee het eerste programma dat precursors

gebruikt als basis voor de detectie van RiPP-subklassen in plaats van de omliggende genen die coderen voor eiwitten betrokken bij de verdere productie van de RiPP. Een obstakel aan deze methode is het aantal kleine genen dat mogelijk voor precursors codeert. Uit de analyse van 1.295 *Streptomyces* genomen werden meer dan 71 miljoen kleine genen gevonden, terwijl een ruime schatting (10 per genoom) niet meer dan 13 duizend precursors voorspelt. Zelfs als het model maar in 0,1 procent van de gevallen een valspositief resultaat zou geven, zou dat betekenen dat het aantal valspositieven vele malen groter zou zijn dan het aantal echte precursors (71 duizend tegenover 13 duizend). In werkelijkheid werden iets meer dan 832 duizend mogelijke precursors gevonden, waarvan het merendeel waarschijnlijk valspositief is. Om deze reden moeten de resultaten verder gefilterd worden.

Er zijn vele methoden denkbaar om precursors te filteren, bijvoorbeeld op basis van precursorsequentie. Een eigenschap die de precursors van een aantal subklassen hebben is de aanwezigheid van meerdere kernsequenties [51, 239, 267]. Elk van deze sequenties wordt verwerkt tot een RiPP, terwijl de rest eraf geknipt wordt door een protease. Deze kernsequenties lijken binnen één precursor wel vaak op elkaar en de aanwezigheid hiervan zou een goede aanwijzing zijn dat het eiwit daadwerkelijk een precursor is. Dit kenmerk is alleen niet universeel voor alle RiPPs en het gebruik van deze filter zou dus betekenen dat precursors van veel RiPPs verwijderd zouden worden. Dit is de afweging die typerend is voor het exploratieve onderzoek beschreven in dit proefschrift: geen enkele filter is op zichzelf perfect, maar elke filter kan wel bepaald voordeel bieden en het deel van de data laten zien waar de gebruiker interesse in heeft.

Het gebruik van de genetische context voor prioritering van nieuwe RiPP BGCs

Alleen de precursor gebruiken om nieuwe RiPP-subklassen te vinden is niet nauwkeurig genoeg. DecRiPPter maakt gebruik van de genetische context van de precursorgenen om de resultaten verder te filteren. Hierin wordt gekeken of er genen aanwezig zijn naast de precursors, die typisch zijn voor RiPP BGCs en waarvan de producten betrokken kunnen zijn bij RiPP biosynthese. Dit soort

genen moeten dus coderen voor modificerende eiwitten, transporteiwitten, regulerende eiwitten en peptidases, die de precursor knippen. Daarnaast wordt er een eis gesteld aan de frequentie van de genen: als deze in het merendeel van de geanalyseerde genomen voorkomen, is het waarschijnlijk dat ze betrokken zijn bij het primaire en niet bij het secundaire metabolisme. Prioriteit wordt daarom gegeven aan genclusters die maar zelden voorkomen, maar wel zoveel mogelijk elementen bevatten die typerend zijn voor RiPPs. Een laatste eis is dat de genclusters niet moeten lijken op die van al bekende RiPP-subklassen. Hiervoor worden de genomen ook geanalyseerd met antiSMASH, software die bekende RiPP BGCs detecteert [39].

Een grondige analyse van 1.295 *Streptomyces* genomen resulteerde in grofweg 700.000 potentiële RiPP BGCs, die niet allemaal handmatig geëvalueerd konden worden. Door steeds strengere eisen te stellen aan de eiwitten die door een gencluster gecodeerd moesten worden, kon de lijst kandidaten worden teruggebracht naar een werkbare hoeveelheid. Helaas bevat niet elk RiPP BGC elk kenmerk (bijvoorbeeld dat sommige RiPP BGCs geen gen voor een regulatie-eiwit of voor een peptidase bevatten), dus door de strengere filters werden deze ook gefilterd. Wel nam het percentage bekende RiPP genclusters toe naarmate strengere filters werden gebruikt, een signaal dat onder de overgebleven kandidaten waarschijnlijk steeds meer nieuwe subklassen gevonden kunnen worden. De dataset werd teruggebracht tot een lijst van enkele honderden genclusters, die handmatig geëvalueerd werden. Dit resulteerde in 151 genclusters gegroepeerd in 42 nieuwe RiPP-subklassen. Eén van deze subklassen is experimenteel gevalideerd en hiervan bleek het inderdaad om een nieuwe RiPP-subklasse te gaan, een nieuwe variant van de lanthipeptiden. Een andere lid van deze subklasse werd rond dezelfde tijd ontdekt via meer traditionele methoden. Hieruit bleek dat ook deze subklasse kandidaten bevat die antimicrobiële activiteit hebben [204]. Het is nog onduidelijk hoeveel van de andere kandidaten daadwerkelijk RiPP-subklassen zijn, maar zelfs als dat voor de helft zou gelden, zou dit als nog een significante bijdrage zijn voor het aantal bekende RiPP-subklassen.

Veel andere methoden en invalshoeken zijn denkbaar om de resultaten te filteren aan de hand van de genetische context. Een ander element dat in de

modificerende enzymen van veel RiPP-subklassen voorkomt is een RiPP Recognition Element (RRE). Dit zijn kleine elementen, die onderdeel maken van enzymen en als een grijparm functioneren om de precursors te herkennen en vast te houden, terwijl de modificatie wordt aangebracht [109]. Opvallend genoeg is de secundaire structuur van dit element vaak hetzelfde, ook tussen verschillende RiPP-subklassen. Deze RREs kunnen consistent gevonden worden door HHPred, een programma dat onder andere de secundaire structuur vergelijkt. HHPred is alleen ontworpen voor de vergelijking van secundaire structuur in het algemeen, niet alleen voor RREs, en neemt daarom veel tijd in beslag. Hierdoor kan de methode niet op grote schaal worden toegepast om RREs van mogelijke nieuwe subklassen te vinden. **RRE-Finder**, beschreven in Hoofdstuk 2, is ontwikkeld om de zoekfunctionaliteit van HHPred toe te spitsen op alleen RREs. RRE-Finder kan in twee modi gerund worden: een conservatieve modus, die razendsnel bekende RREs met bekende sequenties vindt; en een exploratieve modus, die meer tijd kost, maar ook RREs kan vinden die qua sequentie minder vergelijkbaar zijn. In Hoofdstuk 2 laten we zien dat met de exploratieve modus van RRE-Finder er RREs gevonden worden in modificerende enzymen die niet eerder geassocieerd werden met RiPP biosynthese en die dus kunnen leiden tot nieuwe RiPP-subklassen. De exploratieve modus levert wel een groter aantal valspositieven op in vergelijking met de conservatieve modus. De meeste valspositieven worden gevonden in regulatie-eiwitten, waarvoor dit domein vermoedelijk dient als grijparm om DNA te herkennen. Aangezien dit geen modificerende enzymen zijn, zijn deze makkelijk te onderscheiden van mogelijke RiPP modificerende enzymen en het ingebouwde filter kan deze verwijderen. Het combineren van RRE-Finder met decRiPPter zou bovendien tot goede resultaten kunnen leiden en het totaal aantal valspositieven nog verder doen dalen. De kans dat een gencluster zowel een valspositief van RRE-Finder als van decRiPPter bevat, is namelijk aanzienlijk kleiner dan dat één ervan dat is.

Nieuwe RiPPs en classificatie ervan

Uit de resultaten van decRiPPter's analyse op de *Streptomyces*-genomen zijn twee kandidaten geselecteerd voor experimentele analyse. Eén van de genclusters bleek inderdaad de machinerie voor een RiPP te coderen. Dit gencluster, dat onder andere werd gevonden in *Streptomyces pristinaespiralis*,

bevat drie precursorgenen en produceert daarmee drie RiPPs, die de **pristinins** zijn genoemd. Van één hiervan, pristinin A3, is de structuur en de modificaties bepaald met behulp van LCMS-MS, NMR en chemische labeling (zie Hoofdstuk 4). Pristinin A3 bevat thioether bruggen, gedehydrateerde aminozuren en verscheidene serines zijn omgebouwd tot alanines. Dit zijn modificaties die eerder gevonden zijn in lanthipeptides, één van de meest veel voorkomende subklassen van RiPPs. Lanthipeptides hebben vaak antimicrobiële activiteit en zijn daarmee een uitgelezen kandidaat voor verder onderzoek als nieuwe antibiotica.

Net zoals bij andere RiPP-subklassen, kan de karakterisatie van dit BGC leiden tot de identificatie van nog meer nieuwe RiPP BGCs van dezelfde klasse. Hiervoor werd onderzocht welke enzymen die gecodeerd worden door het BGC de modificatie aanbrengen die typerend is voor deze subklasse. De modificatie die elke lanthipeptide heeft is een thioether brug tussen een cysteine en een serine of threonine. Tot nu toe waren er vier verschillende sets enzymen bekend die deze thioether brug konden installeren, wat heeft geleid tot de onderverdeling van lanthipeptides in vier subklassen [47]. De genen van deze enzymen konden niet teruggevonden in het BGC van pristinins, wat hoogstwaarschijnlijk betekent dat het hier om een vijfde subklasse van lanthipeptides gaat. De producten van twee genen van het pristinin BGC leken uitgelezen kandidaten om de thioether brug aan te brengen, niet alleen voor pristinins, maar in alle type V lanthipeptides. Het verwijderen van deze genen uit het pristinin BGC stopte de productie van pristinins, wat suggereert dat hun producten inderdaad betrokken zijn bij de productie van deze RiPPs. Bovendien worden deze twee genen altijd samen teruggevonden, wat suggereert dat ze van elkaar afhankelijk zijn voor hun functie. De aanwezigheid van dit genenpaar in diverse genetische contexten en in de genomen van allerlei bacteriën toont aan dat er nog veel type V lanthipeptides te ontdekken en karakteriseren zijn, die eerder altijd uit het zicht lagen.

Een tweede gencluster dat onderzocht is staat beschreven in Hoofdstuk 5. Het gaat hier om een gencluster dat verdeeld is over twee delen, die naast elkaar in tegengestelde richting liggen op het genoom en elk beginnen met een sterk geconserveerd precursorgen. Deze genen bevatten ook meerdere

herhalende patronen met dezelfde sequentie (TTGWQ). Dit soort sequenties wordt vaak teruggevonden in de precursors van andere RiPPs, waarbij elke herhaalde sequentie in een RiPP wordt omgezet na modificaties. Het gencluster codeert daarnaast een scala aan verschillende modificerende enzymen, waarvan sommigen tot families behoren die betrokken kunnen zijn bij RiPP biosynthese, zoals de radical S-adenosyl methionine (rSAM) enzymen en de ATP-grasp ligases. Desalniettemin is deze overlap erg klein en werd dit gencluster daarom ook niet gevonden door RiPP-identificatiesoftware zoals antiSMASH of BAGEL. Activatie van het gencluster leidde tot de identificatie van een paar moleculen in de chemische extracten van de producerende stam, die niet meer aanwezig waren als het gencluster was verwijderd. Helaas kon de structuur van deze moleculen niet worden opgehelderd en is dus nog niet zeker of deze producten daadwerkelijk RiPPs zijn. Verder onderzoek is nog nodig om de structuur te bepalen en om te zien of deze producten inderdaad van de precursors zijn afgeleid.

Deze BGCs en die van andere recent ontdekte RiPP-subklassen brengen een interessante discussie op gang met betrekking tot RiPP classificatie. RiPPs worden meestal geclassificeerd op basis van één typerende modificatie, terwijl de rest als secundaire modificaties worden gezien. Door het toenemende aantal RiPP-subklassen dat wordt ontdekt, wordt het steeds duidelijker dat genen vaak uitgewisseld worden tussen de BGCs van verschillende subklassen, met als gevolg dat veel modificaties overlappen tussen verschillende RiPPs. Of die modificaties worden gezien als primair of als secundair hangt af van de RiPP-subklasse. Zo worden bijvoorbeeld gedehydrateerde aminozuren gezien als een primaire modificatie in linaridins, maar als secundair in lanthipeptides.

Hetzelfde zal mogelijk het geval zijn bij het gencluster dat in Hoofdstuk 5 wordt besproken. Dit gencluster bevat genen die al eerder geassocieerd zijn met RiPP BGCs en mogelijk zal het product van dit gencluster ook al bekende modificaties bevatten. Desalniettemin wordt dit gencluster niet gevonden door RiPP-detectie software, die er vooral op gericht is alleen BGCs van bekende BGCs te vinden met duidelijk gedefinieerde regels. Hoewel deze methoden effectief zijn, is het belangrijk dat de gebruiker begrijpt waarop de methoden zijn gebaseerd en wat voor belemmeringen dat met zich meebrengt. Een bredere

zoektocht, waarbij gezocht wordt naar RiPP modificerende enzymen in diverse genetische contexten, ongeacht primair of secundair, kan leiden tot interessante ontdekkingen en had mogelijk al kunnen leiden tot de ontdekking van de BGCs die in dit proefschrift besproken staan.

Conclusie

BGCs en hun producten komen in vele soorten en maten, wat leidt tot een enorme genetische diversiteit om te onderzoeken. In dit proefschrift zijn verschillende methoden onderzocht die proberen een meer verkennende invalshoek te geven aan deze zoektocht door te zoeken naar nieuwe RiPP BGCs middels kunstmatige intelligentie. Hoewel RiPP BGCs erg verschillend zijn, blijft de biosynthetische logica behouden, waarvan gebruik wordt gemaakt bij de ontwikkeling van decRiPPter en RRE-Finder. Een groot scala aan kandidaten wordt geprioriteerd aan de hand van de precursorgenen en hun genetische context. Twee van deze kandidaten zijn onderzocht en van één is aangetoond dat het inderdaad om een nieuwe RiPP-subklasse gaat. Dit laat zien dat zulke methodes veel potentie hebben, door rekening te houden met een grotere kans op het aantal vals-positieven. Verdere ontwikkeling van de modellering, aangescherpt door nieuwe biologische en chemische kennis, zullen kunnen leiden tot een doorgaande stroom van nieuwe RiPP BGCs en dit zal hopelijk leiden tot de ontdekking van nieuwe antibiotica.

Reference list

1. Kossel, A., *Archiv für Physiologie*, in *Archiv für Physiologie*. 1891, Veit & Comp.
2. Demain, A.L. and A. Fang, *The natural functions of secondary metabolites*, in *History of Modern Biotechnology*. 2000, Springer: Berlin. p. 1-39.
3. Amos, G.C.A., et al., *Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality.* Proc Natl Acad Sci U S A, 2017. **114**(52): p. E11121-E11130.
4. Xu, L., et al., *Comparative Genomic Insights into Secondary Metabolism Biosynthetic Gene Cluster Distributions of Marine Streptomyces.* Mar Drugs, 2019. **17**(9).
5. Choudoir, M.J., C. Pepe-Ranney, and D.H. Buckley, *Diversification of Secondary Metabolite Biosynthetic Gene Clusters Coincides with Lineage Divergence in Streptomyces.* Antibiotics (Basel), 2018. **7**(1).
6. Fleming, A., *On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. 1929.* Bull World Health Organ, 2001. **79**(8): p. 780-90.
7. Ligon, B.L., *Penicillin: its discovery and early development.* Semin Pediatr Infect Dis, 2004. **15**(1): p. 52-7.
8. Woodruff, H.B., *Selman A. Waksman, winner of the 1952 Nobel Prize for physiology or medicine.* Appl Environ Microbiol, 2014. **80**(1): p. 2-8.
9. Schatz, A., E. Bugie, and S.A. Waksman, *Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. 1944.* Clin Orthop Relat Res, 2005(437): p. 3-6.
10. Waksman, S.A. and H.A. Lechevalier, *Neomycin, a New Antibiotic Active against Streptomycin-Resistant Bacteria, including Tuberculosis Organisms.* Science, 1949. **109**(2830): p. 305-7.
11. Durand, G.A., D. Raoult, and G. Dubourg, *Antibiotic discovery: history, methods and perspectives.* Int J Antimicrob Agents, 2019. **53**(4): p. 371-382.
12. Berdy, J., *Bioactive microbial metabolites.* J Antibiot (Tokyo), 2005. **58**(1): p. 1-26.
13. Merillon, J.M. and K.G. Ramawat, *Biotechnology for Medicinal Plants: Research Need*, in *Biotechnology: Secondary Metabolites: Plants and Microbes (2nd edition)*. 2007, CRC Press. p. 3-4.
14. Davies, J. and D. Davies, *Origins and evolution of antibiotic resistance.* Microbiol Mol Biol Rev, 2010. **74**(3): p. 417-33.
15. Davies, J., *What are antibiotics? Archaic functions for modern activities.* Mol Microbiol, 1990. **4**(8): p. 1227-32.
16. Aminov, R.I., *The role of antibiotics and antibiotic resistance in nature.* Environ Microbiol, 2009. **11**(12): p. 2970-88.
17. Rice, L.B., *Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE.* J Infect Dis, 2008. **197**(8): p. 1079-81.

18. Velayati, A.A., et al., *Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in iran.* Chest, 2009. **136**(2): p. 420-425.
19. Sotgiu, G., et al., *Epidemiology and clinical management of XDR-TB: a systematic review by TBNET.* Eur Respir J, 2009. **33**(4): p. 871-81.
20. Migliori, G.B., et al., *MDR-TB and XDR-TB: drug resistance and treatment outcomes.* Eur Respir J, 2009. **34**(3): p. 778-9.
21. Organization, W.H., *Antimicrobial resistance: global report on surveillance*. 2014, France.
22. Payne, D.J., et al., *Drugs for bad bugs: confronting the challenges of antibacterial discovery.* Nat Rev Drug Discov, 2007. **6**(1): p. 29-40.
23. Lewis, K., *Platforms for antibiotic discovery.* Nat Rev Drug Discov, 2013. **12**(5): p. 371-87.
24. Kolter, R. and G.P. van Wezel, *Goodbye to brute force in antibiotic discovery?* Nat Microbiol, 2016. **1**: p. 15020.
25. Cooper, M.A. and D. Shlaes, *Fix the antibiotics pipeline.* Nature, 2011. **472**(7341): p. 32.
26. Medema, M.H. and M.A. Fischbach, *Computational approaches to natural product discovery.* Nat Chem Biol, 2015. **11**(9): p. 639-48.
27. Bentley, S.D., et al., *Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2).* Nature, 2002. **417**(6885): p. 141-7.
28. Kautsar, S.A., et al., *BiG-SCLiCE; A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters.* 2020.
29. Kautsar, S.A., et al., *MIBiG 2.0: a repository for biosynthetic gene clusters of known function.* Nucleic Acids Res, 2020. **48**(D1): p. D454-D458.
30. van der Meij, A., et al., *Chemical ecology of antibiotic production by actinomycetes.* FEMS Microbiol Rev, 2017. **41**(3): p. 392-416.
31. van Bergeijk, D.A., et al., *Ecology and genomics of Actinobacteria: new concepts for natural product discovery.* Nat Rev Microbiol, 2020. **18**(10): p. 546-558.
32. Barka, E.A., et al., *Taxonomy, Physiology, and Natural Products of Actinobacteria.* Microbiol Mol Biol Rev, 2016. **80**(1): p. 1-43.
33. Rutledge, P.J. and G.L. Challis, *Discovery of microbial natural products by activation of silent biosynthetic gene clusters.* Nat Rev Microbiol, 2015. **13**(8): p. 509-23.
34. Mao, D., et al., *Recent advances in activating silent biosynthetic gene clusters in bacteria.* Curr Opin Microbiol, 2018. **45**: p. 156-163.
35. Luo, Y., B. Enghiad, and H. Zhao, *New tools for reconstruction and heterologous expression of natural product biosynthetic gene clusters.* Nat Prod Rep, 2016. **33**(2): p. 174-82.
36. Weissman, K.J., *The structural biology of biosynthetic megaenzymes.* Nat Chem Biol, 2015. **11**(9): p. 660-70.
37. Oldfield, E. and F.Y. Lin, *Terpene biosynthesis: modularity rules.* Angew Chem Int Ed Engl, 2012. **51**(5): p. 1124-37.

38. Hertweck, C., et al., *Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork.* Nat Prod Rep, 2007. **24**(1): p. 162-90.
39. Blin, K., et al., *antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline.* Nucleic Acids Res, 2019. **47**(W1): p. W81-W87.
40. Skinnider, M.A., et al., *PRISM 3: expanded prediction of natural product chemical structures from microbial genomes.* Nucleic Acids Res, 2017. **45**(W1): p. W49-W54.
41. Cimermancic, P., et al., *Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters.* Cell, 2014. **158**(2): p. 412-421.
42. Arnison, P.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): p. 108-60.
43. Oman, T.J. and W.A. van der Donk, *Follow the leader: the use of leader peptides to guide natural product biosynthesis.* Nat Chem Biol, 2010. **6**(1): p. 9-18.
44. Ortega, M.A. and W.A. van der Donk, *New Insights into the Biosynthetic Logic of Ribosomally Synthesized and Post-translationally Modified Peptide Natural Products.* Cell Chem Biol, 2016. **23**(1): p. 31-44.
45. Tietz, J.I., et al., *A new genome-mining tool redefines the lasso peptide biosynthetic landscape.* Nat Chem Biol, 2017. **13**(5): p. 470-478.
46. Kelly, W.L., L. Pan, and C. Li, *Thiostrepton biosynthesis: prototype for a new family of bacteriocins.* J Am Chem Soc, 2009. **131**(12): p. 4327-34.
47. Repka, L.M., et al., *Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes.* Chem Rev, 2017. **117**(8): p. 5457-5520.
48. Montalbán-López, M., *New developments in RiPP discovery, enzymology and engineering.* 2020.
49. Zdouc, M.M., et al., *A biaryl-linked tripeptide from Planomonospora leads to a widespread class of minimal RiPP gene clusters.* Biorxiv, 2020.
50. Berridge, N.J., G.G. Newton, and E.P. Abraham, *Purification and nature of the antibiotic nisin.* Biochem J, 1952. **52**(4): p. 529-35.
51. Kersten, R.D. and J.K. Weng, *Gene-guided discovery and engineering of branched cyclic peptides in plants.* Proc Natl Acad Sci U S A, 2018. **115**(46): p. E10961-E10969.
52. Santos-Aberturas, J., et al., *Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool.* Nucleic Acids Res, 2019. **47**(9): p. 4624-4637.
53. Salomon, R.A. and R.N. Farias, *Microcin 25, a novel antimicrobial peptide produced by Escherichia coli.* J Bacteriol, 1992. **174**(22): p. 7428-35.
54. Lee, H., Y. Park, and S. Kim, *Enzymatic Cross-Linking of Side Chains Generates a Modified Peptide with Four Hairpin-like Bicyclic Repeats.* Biochemistry, 2017. **56**(37): p. 4927-4930.
55. Hudson, G.A., et al., *Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified*

Peptides Identifies New Calpha, Cbeta, and Cgamma-Linked Thioether-Containing Peptides. J Am Chem Soc, 2019. **141**(20): p. 8228-8238.
56. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
57. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
58. Velasquez, J.E. and W.A. van der Donk, *Genome mining for ribosomally synthesized natural products.* Curr Opin Chem Biol, 2011. **15**(1): p. 11-21.
59. McClerren, A.L., et al., *Discovery and in vitro biosynthesis of haloduracin, a two-component lantibiotic.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17243-8.
60. Sudek, S., et al., *Structure of trichamide, a cyclic peptide from the bloom-forming cyanobacterium Trichodesmium erythraeum, predicted from the genome sequence.* Appl Environ Microbiol, 2006. **72**(6): p. 4382-7.
61. Knappe, T.A., et al., *Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of Burkholderia thailandensis E264.* J Am Chem Soc, 2008. **130**(34): p. 11446-54.
62. van Heel, A.J., et al., *BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins.* Nucleic Acids Res, 2018. **46**(W1): p. W278-W281.
63. de Jong, A., et al., *BAGEL: a web-based bacteriocin genome mining tool.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W273-9.
64. Skinnider, M.A., et al., *Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining.* Proc Natl Acad Sci U S A, 2016. **113**(42): p. E6343-E6351.
65. Eddy, S.R., *Accelerated Profile HMM Searches.* PLoS Comput Biol, 2011. **7**(10): p. e1002195.
66. Iftime, D., et al., *Streptocollin, a Type IV Lanthipeptide Produced by Streptomyces collinus Tu 365.* Chembiochem, 2015. **16**(18): p. 2615-23.
67. Poorinmohammad, N., R. Bagheban-Shemirani, and J. Hamedi, *Genome mining for ribosomally synthesised and post-translationally modified peptides (RiPPs) reveals undiscovered bioactive potentials of actinobacteria.* Antonie Van Leeuwenhoek, 2019. **112**(10): p. 1477-1499.
68. Zhang, Q., et al., *Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in actinobacteria.* Appl Environ Microbiol, 2015. **81**(13): p. 4339-50.
69. Delcher, A.L., et al., *Identifying bacterial genes and endosymbiont DNA with Glimmer.* Bioinformatics, 2007. **23**(6): p. 673-9.
70. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Res, 1999. **27**(23): p. 4636-41.
71. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC Bioinformatics, 2010. **11**: p. 119.
72. Schwalen, C.J., et al., *Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics.* J Am Chem Soc, 2018. **140**(30): p. 9494-9501.

73. Walker, M.C., et al., *Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family.* BMC Genomics, 2020. **21**(1): p. 387.
74. Georgiou, M.A., et al., *Bioinformatic and Reactivity-Based Discovery of Linaridins.* bioRxiv, 2020: p. 2020.07.09.196543.
75. El-Gebali, S., et al., *The Pfam protein families database in 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D427-D432.
76. Haft, D.H., et al., *TIGRFAMs: a protein family resource for the functional identification of proteins.* Nucleic Acids Res, 2001. **29**(1): p. 41-3.
77. Blin, K., et al., *antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification.* Nucleic Acids Res, 2017. **45**(W1): p. W36-W41.
78. Zong, C., et al., *Albusnodin: an acetylated lasso peptide from Streptomyces albus.* Chem Commun (Camb), 2018. **54**(11): p. 1339-1342.
79. Iorio, M., et al., *A glycosylated, labionin-containing lanthipeptide with marked antinociceptive activity.* ACS Chem Biol, 2014. **9**(2): p. 398-404.
80. Wiebach, V., et al., *The anti-staphylococcal lipolanthines are ribosomally synthesized lipopeptides.* Nat Chem Biol, 2018. **14**(7): p. 652-654.
81. Burkhart, B.J., et al., *Chimeric Leader Peptides for the Generation of Non-Natural Hybrid RiPP Products.* ACS Cent Sci, 2017. **3**(6): p. 629-638.
82. Ziemert, N., et al., *Ribosomal synthesis of tricyclic depsipeptides in bloom-forming cyanobacteria.* Angew Chem Int Ed Engl, 2008. **47**(40): p. 7756-9.
83. Ziemert, N., et al., *Exploiting the natural diversity of microviridin gene clusters for discovery of novel tricyclic depsipeptides.* Appl Environ Microbiol, 2010. **76**(11): p. 3568-74.
84. Roh, H., et al., *A Topologically Distinct Modified Peptide with Multiple Bicyclic Core Motifs Expands the Diversity of Microviridin-Like Peptides.* Chembiochem, 2019. **20**(8): p. 1051-1059.
85. Lee, H., et al., *Genome Mining Reveals High Topological Diversity of omega-Ester-Containing Peptides and Divergent Evolution of ATP-Grasp Macrocyclases.* J Am Chem Soc, 2020. **142**(6): p. 3013-3023.
86. DiCaprio, A.J., et al., *Enzymatic Reconstitution and Biosynthetic Investigation of the Lasso Peptide Fusilassin.* J Am Chem Soc, 2019. **141**(1): p. 290-297.
87. Agrawal, P., et al., *RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links.* Nucleic Acids Res, 2017. **45**(W1): p. W80-W88.
88. de Los Santos, E.L.C., *NeuRiPP: Neural network identification of RiPP precursor peptides.* Sci Rep, 2019. **9**(1): p. 13406.
89. Merwin, N.J., 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): p. 371-380.
90. Mohimani, H., et al., *Dereplication of peptidic natural products through database search of mass spectra.* Nat Chem Biol, 2017. **13**(1): p. 30-37.
91. Gurevich, A., et al., *Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra.* Nat Microbiol, 2018. **3**(3): p. 319-327.

92. Cao, L., et al., *MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities.* Cell Syst, 2019. **9**(6): p. 600-608 e4.
93. Hayakawa, Y., et al., *Thioviridamide, a novel apoptosis inducer in transformed cells from Streptomyces olivoviridis.* J Antibiot (Tokyo), 2006. **59**(1): p. 1-5.
94. Tang, J., et al., *Discovery and biosynthesis of thioviridamide-like compounds.* Chinese Chemical Letters, 2018. **29**(7): p. 1022-1028.
95. Burkhart, B.J., et al., *YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function.* Chem Rev, 2017. **117**(8): p. 5389-5456.
96. Benjdia, A., C. Balty, and O. Berteau, *Radical SAM Enzymes in the Biosynthesis of Ribosomally Synthesized and Post-translationally Modified Peptides (RiPPs).* Front Chem, 2017. **5**: p. 87.
97. Gomez-Escribano, J.P., et al., *Posttranslational beta-methylation and macrolactamidination in the biosynthesis of the bottromycin complex of ribosomal peptide antibiotics.* Chemical Science, 2012. **3**(12): p. 3522-3525.
98. Freeman, M.F., et al., *Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides.* Science, 2012. **338**(6105): p. 387-90.
99. Grove, T.L., et al., *Structural Insights into Thioether Bond Formation in the Biosynthesis of Sactipeptides.* J Am Chem Soc, 2017. **139**(34): p. 11734-11744.
100. Holliday, G.L., et al., *Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a "Plug and Play" Domain.* Methods Enzymol, 2018. **606**: p. 1-71.
101. Fluhe, L., et al., *Two [4Fe-4S] clusters containing radical SAM enzyme SkfB catalyze thioether bond formation during the maturation of the sporulation killing factor.* J Am Chem Soc, 2013. **135**(3): p. 959-62.
102. Haft, D.H., *Bioinformatic evidence for a widely distributed, ribosomally produced electron carrier precursor, its maturation proteins, and its nicotinoprotein redox partners.* BMC Genomics, 2011. **12**: p. 21.
103. Morinaka, B.I., et al., *Natural noncanonical protein splicing yields products with diverse beta-amino acid residues.* Science, 2018. **359**(6377): p. 779-782.
104. Bushin, L.B., et al., *Charting an Unexplored Streptococcal Biosynthetic Landscape Reveals a Unique Peptide Cyclization Motif.* J Am Chem Soc, 2018. **140**(50): p. 17674-17684.
105. Schramma, K.R., L.B. Bushin, and M.R. Seyedsayamdost, *Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink.* Nat Chem, 2015. **7**(5): p. 431-437.
106. Caruso, A., et al., *Radical Approach to Enzymatic beta-Thioether Bond Formation.* J Am Chem Soc, 2019. **141**(2): p. 990-997.
107. Caruso, A., et al., *Macrocyclization via an Arginine-Tyrosine Crosslink Broadens the Reaction Scope of Radical S-Adenosylmethionine Enzymes.* J Am Chem Soc, 2019. **141**(42): p. 16610-16614.
108. Clark, K.A., L.B. Bushin, and M.R. Seyedsayamdost, *Aliphatic Ether Bond Formation Expands the Scope of Radical SAM Enzymes in Natural Product Biosynthesis.* J Am Chem Soc, 2019. **141**(27): p. 10610-10615.

109. Burkhart, B.J., et al., *A prevalent peptide-binding domain guides ribosomal natural product biosynthesis.* Nat Chem Biol, 2015. **11**(8): p. 564-70.
110. Zhang, Z., et al., *Biosynthetic Timing and Substrate Specificity for the Thiopeptide Thiomuracin.* J Am Chem Soc, 2016. **138**(48): p. 15511-15514.
111. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
112. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J Mol Biol, 1999. **292**(2): p. 195-202.
113. Ting, C.P., et al., *Use of a scaffold peptide in the biosynthesis of amino acid-derived natural products.* Science, 2019. **365**(6450): p. 280-284.
114. Vignolle, G.A., et al., *Novel approach in whole genome mining and transcriptome analysis reveal conserved RiPPs in Trichoderma spp.* BMC Genomics, 2020. **21**(1): p. 258.
115. Tagirdzhanov, A.M., A. Shlemov, and A. Gurevich, *NPS: scoring and evaluating the statistical significance of peptidic natural product-spectrum matches.* Bioinformatics, 2019. **35**(14): p. i315-i323.
116. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2020.
117. Hudson, G.A. and D.A. Mitchell, *RiPP antibiotics: biosynthesis and engineering potential.* Curr Opin Microbiol, 2018. **45**: p. 61-69.
118. Cox, C.L., J.R. Doroghazi, and D.A. Mitchell, *The genomic landscape of ribosomal peptides containing thiazole and oxazole heterocycles.* BMC Genomics, 2015. **16**: p. 778.
119. Zhang, Q., et al., *Evolution of lanthipeptide synthetases.* Proc Natl Acad Sci U S A, 2012. **109**(45): p. 18361-6.
120. Davis, K.M., et al., *Structures of the peptide-modifying radical SAM enzyme SuiB elucidate the basis of substrate recognition.* Proc Natl Acad Sci U S A, 2017. **114**(39): p. 10420-10425.
121. Ortega, M.A., et al., *Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB.* Nature, 2015. **517**(7535): p. 509-12.
122. Koehnke, J., et al., *Structural analysis of leader peptide binding enables leader-free cyanobactin processing.* Nat Chem Biol, 2015. **11**(8): p. 558-563.
123. Latham, J.A., et al., *PqqD is a novel peptide chaperone that forms a ternary complex with the radical S-adenosylmethionine protein PqqE in the pyrroloquinoline quinone biosynthetic pathway.* J Biol Chem, 2015. **290**(20): p. 12908-18.
124. Mavaro, A., et al., *Substrate recognition and specificity of the NisB protein, the lantibiotic dehydratase involved in nisin biosynthesis.* J Biol Chem, 2011. **286**(35): p. 30552-60.
125. Mitchell, D.A., et al., *Structural and functional dissection of the heterocyclic peptide cytotoxin streptolysin S.* J Biol Chem, 2009. **284**(19): p. 13004-12.
126. Regni, C.A., et al., *How the MccB bacterial ancestor of ubiquitin E1 initiates biosynthesis of the microcin C7 antibiotic.* EMBO J, 2009. **28**(13): p. 1953-64.

127. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W5-9.
128. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.
129. Klinman, J.P. and F. Bonnot, *Intrigues and intricacies of the biosynthetic pathways for the enzymatic quinocofactors: PQQ, TTQ, CTQ, TPQ, and LTQ.* Chem Rev, 2014. **114**(8): p. 4343-65.
130. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res, 2016. **44**(D1): p. D279-85.
131. Evans, R.L., 3rd, et al., *Nuclear Magnetic Resonance Structure and Binding Studies of PqqD, a Chaperone Required in the Biosynthesis of the Bacterial Dehydrogenase Cofactor Pyrroloquinoline Quinone.* Biochemistry, 2017. **56**(21): p. 2735-2746.
132. Zallot, R., N. Oberg, and J.A. Gerlt, *The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.* Biochemistry, 2019. **58**(41): p. 4169-4182.
133. Su, G., et al., *Biological network exploration with Cytoscape 3.* Curr Protoc Bioinformatics, 2014. **47**: p. 8 13 1-24.
134. Finn, R.D., et al., *HMMER web server: 2015 update.* Nucleic Acids Res, 2015. **43**(W1): p. W30-8.
135. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
136. Mirdita, M., et al., *Uniclust databases of clustered and deeply annotated protein sequences and alignments.* Nucleic Acids Res, 2017. **45**(D1): p. D170-D176.
137. Blin, K., et al., *The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters.* Nucleic Acids Res, 2017. **45**(D1): p. D555-D559.
138. Steinegger, M. and J. Soding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nat Biotechnol, 2017. **35**(11): p. 1026-1028.
139. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nat Methods, 2011. **9**(2): p. 173-5.
140. Epstein, S.C., L.K. Charkoudian, and M.H. Medema, *A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: prospects for research-based educational experiences.* Stand Genomic Sci, 2018. **13**: p. 16.
141. van der Donk, W.A. and S.K. Nair, *Structure and mechanism of lanthipeptide biosynthetic enzymes.* Curr Opin Struct Biol, 2014. **29**: p. 58-66.
142. Cheung, W.L., et al., *Lasso Peptide Biosynthetic Protein LarB1 Binds Both Leader and Core Peptide Regions of the Precursor Protein LarA.* ACS Cent Sci, 2016. **2**(10): p. 702-709.

143. Dunbar, K.L., et al., *Identification of an Auxiliary Leader Peptide-Binding Protein Required for Azoline Formation in Ribosomal Natural Products.* J Am Chem Soc, 2015. **137**(24): p. 7672-7.
144. Haft, D.H., et al., *TIGRFAMs and Genome Properties in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D387-95.
145. Khusainov, R., G.N. Moll, and O.P. Kuipers, *Identification of distinct nisin leader peptide regions that determine interactions with the modification enzymes NisB and NisC.* FEBS Open Bio, 2013. **3**: p. 237-42.
146. Haft, D.H. and M.K. Basu, *Biological systems discovery in silico: radical S-adenosylmethionine protein families and their target peptides for posttranslational modification.* J Bacteriol, 2011. **193**(11): p. 2745-55.
147. Zhang, Z. and W.A. van der Donk, *Nonribosomal Peptide Extension by a Peptide Amino-Acyl tRNA Ligase.* J Am Chem Soc, 2019. **141**(50): p. 19625-19633.
148. Hannigan, G.D., et al., *A deep learning genome-mining strategy for biosynthetic gene cluster prediction.* Nucleic Acids Res, 2019. **47**(18): p. e110.
149. Sardar, D., et al., *Recognition sequences and substrate evolution in cyanobactin biosynthesis.* ACS Synth Biol, 2015. **4**(2): p. 167-76.
150. Schwalen, C.J., et al., *In Vitro Biosynthetic Studies of Bottromycin Expand the Enzymatic Capabilities of the YcaO Superfamily.* J Am Chem Soc, 2017. **139**(50): p. 18154-18157.
151. Ghodge, S.V., et al., *Post-translational Claisen Condensation and Decarboxylation en Route to the Bicyclic Core of Pantocin A.* J Am Chem Soc, 2016. **138**(17): p. 5487-90.
152. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: improvements in performance and usability.* Mol Biol Evol, 2013. **30**(4): p. 772-80.
153. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
154. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Mol Biol Evol, 2009. **26**(7): p. 1641-50.
155. Miyanaga, A., et al., *Discovery and assembly-line biosynthesis of the lymphostin pyrroloquinoline alkaloid family of mTOR inhibitors in Salinispora bacteria.* J Am Chem Soc, 2011. **133**(34): p. 13311-3.
156. Haft, D.H., M.K. Basu, and D.A. Mitchell, *Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family.* BMC Biol, 2010. **8**: p. 70.
157. Davies, J., *Origins and evolution of antibiotic resistance.* Microbiologia, 1996. **12**(1): p. 9-16.
158. Berdy, J., *Thoughts and facts about antibiotics: Where we are now and where we are heading.* J Antibiot (Tokyo), 2012. **65**(8): p. 441.
159. van der Aart, L.T., et al., *Polyphasic classification of the gifted natural product producer Streptomyces roseifaciens sp. nov.* Int J Syst Evol Microbiol, 2019. **69**(4): p. 899-908.

160. Cruz-Morales, P., et al., *Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes.* Genome Biol Evol, 2016. **8**(6): p. 1906-16.
161. Selem-Mojica, N., et al., *EvoMining reveals the origin and fate of natural product biosynthetic enzymes.* Microb Genom, 2019. **5**(12).
162. Noike, M., et al., *A peptide ligase and the ribosome cooperate to synthesize the peptide pheganomycin.* Nat Chem Biol, 2015. **11**(1): p. 71-6.
163. Ogasawara, Y., et al., *Exploring Peptide Ligase Orthologs in Actinobacteria-Discovery of Pseudopeptide Natural Products, Ketomemicins.* ACS Chem Biol, 2016. **11**(6): p. 1686-92.
164. Singh, M. and D. Sareen, *Novel LanT associated lantibiotic clusters identified by genome database mining.* PLoS One, 2014. **9**(3): p. e91352.
165. Mitchell, A., et al., *The InterPro protein families database: the classification resource after 15 years.* Nucleic Acids Res, 2015. **43**(Database issue): p. D213-21.
166. Medema, M.H., et al., *A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis.* PLoS Comput Biol, 2014. **10**(12): p. e1004016.
167. Wolf, Y.I. and E.V. Koonin, *A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes.* Genome Biol Evol, 2012. **4**(12): p. 1286-94.
168. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
169. Dalquen, D.A. and C. Dessimoz, *Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals.* Genome Biol Evol, 2013. **5**(10): p. 1800-6.
170. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, *An efficient algorithm for large-scale detection of protein families.* Nucleic Acids Res, 2002. **30**(7): p. 1575-84.
171. Van Dongen, S., *Graph clustering by Flow Simulation*. 2000, University of Utrecht.
172. Kriventseva, E.V., 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): p. D807-D811.
173. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.
174. Waterhouse, R.M., et al., *BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.* Mol Biol Evol, 2018. **35**(3): p. 543-548.
175. Atkinson, H.J., et al., *Using sequence similarity networks for visualization of relationships across diverse protein superfamilies.* PLoS One, 2009. **4**(2): p. e4345.
176. Kersten, R.D., et al., *A mass spectrometry-guided genome mining approach for natural product peptidogenomics.* Nat Chem Biol, 2011. **7**(11): p. 794-802.

177. Onaka, H., et al., *Cloning and characterization of the goadsporin biosynthetic gene cluster from Streptomyces sp. TP-A0584.* Microbiology, 2005. **151**(Pt 12): p. 3923-3933.
178. Crone, W.J.K., F.J. Leeper, and A.W. Truman, *Identification and characterisation of the gene cluster for the anti-MRSA antibiotic bottromycin: expanding the biosynthetic diversity of ribosomal peptides.* Chemical Science, 2012. **3**(12): p. 3516-3521.
179. Hou, Y., et al., *Structure and biosynthesis of the antibiotic bottromycin D.* Org Lett, 2012. **14**(19): p. 5050-3.
180. Horie, A., et al., *Discovery of proteinaceous N-modification in lysine biosynthesis of Thermus thermophilus.* Nat Chem Biol, 2009. **5**(9): p. 673-9.
181. Fawaz, M.V., M.E. Topper, and S.M. Firestine, *The ATP-grasp enzymes.* Bioorg Chem, 2011. **39**(5-6): p. 185-91.
182. van der Palen, C.J., et al., *MauE and MauD proteins are essential in methylamine metabolism of Paracoccus denitrificans.* Antonie Van Leeuwenhoek, 1997. **72**(3): p. 219-28.
183. Jacobi, A., R. Rossmann, and A. Bock, *The hyp operon gene products are required for the maturation of catalytically active hydrogenase isoenzymes in Escherichia coli.* Arch Microbiol, 1992. **158**(6): p. 444-51.
184. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2019. **47**(D1): p. D23-D28.
185. Navarro-Munoz, J.C., et al., *A computational framework to explore large-scale biosynthetic diversity.* Nat Chem Biol, 2020. **16**(1): p. 60-68.
186. van der Heul, H.U., et al., *Regulation of antibiotic production in Actinobacteria: new perspectives from the post-genomic era.* Nat Prod Rep, 2018. **35**(6): p. 575-604.
187. Bode, H.B., et al., *Big effects from small changes: possible ways to explore nature's chemical diversity.* Chembiochem, 2002. **3**(7): p. 619-27.
188. Wu, C., et al., *Lugdunomycin, an angucycline-derived molecule with unprecedented chemical architecture.* Angew Chem Int Ed Engl, 2019. **58**(9): p. 2809-2814.
189. van Bergeijk, D.A., et al., *Ecology and genomics of Actinobacteria: new concepts for natural product discovery.* Nat Rev Microbiol, 2020.
190. Zhu, H., et al., *Eliciting antibiotics active against the ESKAPE pathogens in a collection of actinomycetes isolated from mountain soils.* Microbiology, 2014. **160**: p. 1714-1725.
191. Folcher, M., et al., *A transcriptional regulator of a pristinamycin resistance gene in Streptomyces coelicolor.* J Biol Chem, 2001. **276**(2): p. 1479-85.
192. Mast, Y., et al., *Characterization of the 'pristinamycin supercluster' of Streptomyces pristinaespiralis.* Microb Biotechnol, 2011. **4**(2): p. 192-206.
193. Garneau, S., N.I. Martin, and J.C. Vederas, *Two-peptide bacteriocins produced by lactic acid bacteria.* Biochimie, 2002. **84**(5-6): p. 577-92.
194. Sit, C.S., S. Yoganathan, and J.C. Vederas, *Biosynthesis of aminovinyl-cysteine-containing peptides and its application in the production of potential drug candidates.* Acc Chem Res, 2011. **44**(4): p. 261-8.

195. Clausen, M., et al., *PAD1 encodes phenylacrylic acid decarboxylase which confers resistance to cinnamic acid in Saccharomyces cerevisiae.* Gene, 1994. **142**(1): p. 107-12.
196. Medema, M.H., et al., *Minimum Information about a Biosynthetic Gene cluster.* Nat Chem Biol, 2015. **11**(9): p. 625-31.
197. Izawa, M., et al., *Identification of essential biosynthetic genes and a true biosynthetic product for thioviridamide.* J Gen Appl Microbiol, 2018. **64**(1): p. 50-53.
198. Vara, J., et al., *Cloning of genes governing the deoxysugar portion of the erythromycin biosynthesis pathway in Saccharopolyspora erythraea (Streptomyces erythreus).* J Bacteriol, 1989. **171**(11): p. 5872-81.
199. Bierman, M., et al., *Plasmid cloning vectors for the conjugal transfer of DNA from Escherichia coli to Streptomyces spp.* Gene, 1992. **116**(1): p. 43-9.
200. Fedoryshyn, M., et al., *Functional expression of the Cre recombinase in actinomycetes.* Appl Microbiol Biotechnol, 2008. **78**(6): p. 1065-70.
201. Larson, J.L. and C.L. Hershberger, *The minimal replicon of a streptomycete plasmid produces an ultrahigh level of plasmid DNA.* Plasmid, 1986. **15**(3): p. 199-209.
202. Yang, X. and W.A. van der Donk, *Post-translational Introduction of D-Alanine into Ribosomally Synthesized Peptides by the Dehydroalanine Reductase NpnJ.* J Am Chem Soc, 2015. **137**(39): p. 12426-9.
203. Zhao, X. and W.A. van der Donk, *Structural Characterization and Bioactivity Analysis of the Two-Component Lantibiotic Flv System from a Ruminant Bacterium.* Cell Chem Biol, 2016. **23**(2): p. 246-256.
204. Ortiz-Lopez, F.J., et al., *Cacaoidin, First Member of the New Lanthidin RiPP Family.* Angew Chem Int Ed Engl, 2020. **59**(31): p. 12654-12658.
205. Ross, A.C., et al., *Synthesis of the lantibiotic lactocin S using peptide cyclizations on solid phase.* J Am Chem Soc, 2010. **132**(2): p. 462-3.
206. Frattaruolo, L., et al., *A Genomics-Based Approach Identifies a Thioviridamide-Like Compound with Selective Anticancer Activity.* ACS Chem Biol, 2017. **12**(11): p. 2815-2822.
207. Kenney, G.E., et al., *The biosynthesis of methanobactin.* Science, 2018. **359**(6382): p. 1411-1416.
208. Giacomini, A., A. Squartini, and M.P. Nuti, *Nucleotide sequence and analysis of plasmid pMD136 from Pediococcus pentosaceus FBB61 (ATCC43200) involved in pediocin A production.* Plasmid, 2000. **43**(2): p. 111-22.
209. Cotter, P.D., 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): p. 18584-9.
210. Kieser, T., et al., *Practical Streptomyces genetics*. 2000: John Innes Foundation.
211. Swiatek, M.A., et al., *Functional analysis of the N-acetylglucosamine metabolic genes of Streptomyces coelicolor and role in the control of development and antibiotic production.* J Bacteriol, 2012. **194**(5): p. 1136-1144.

212. Zacchetti, B., et al., *Aggregation of germlings is a major contributing factor towards mycelial heterogeneity of Streptomyces.* Sci Rep, 2016. **6**: p. 27045.
213. Zacchetti, B., P. Smits, and D. Claessen, *Dynamics of Pellet Fragmentation and Aggregation in Liquid-Grown Cultures of Streptomyces lividans.* Front Microbiol, 2018. **9**: p. 943.
214. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* BMC Bioinformatics, 2010. **11**: p. 395.
215. Myers, O.D., et al., *One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.* Anal Chem, 2017. **89**(17): p. 8696-8703.
216. Chong, J., D.S. Wishart, and J. Xia, *Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis.* Curr Protoc Bioinformatics, 2019. **68**(1): p. e86.
217. Wessel, D. and U.I. Flugge, *A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids.* Anal Biochem, 1984. **138**(1): p. 141-3.
218. van Rooden, E.J., et al., *Mapping in vivo target interaction profiles of covalent inhibitors using chemical proteomics with label-free quantification.* Nat Protoc, 2018. **13**(4): p. 752-767.
219. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
220. Distler, U., et al., *Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics.* Nat Methods, 2014. **11**(2): p. 167-70.
221. Elsayed, S.S., et al., *Chaxapeptin, a Lasso Peptide from Extremotolerant Streptomyces leeuwenhoekii Strain C58 from the Hyperarid Atacama Desert.* J Org Chem, 2015. **80**(20): p. 10252-60.
222. Katz, L. and R.H. Baltz, *Natural product discovery: past, present, and future.* J Ind Microbiol Biotechnol, 2016. **43**(2-3): p. 155-76.
223. Silver, L.L., *Challenges of antibacterial discovery.* Clin Microbiol Rev, 2011. **24**(1): p. 71-109.
224. Doroghazi, J.R., et al., *A roadmap for natural product discovery based on large-scale genomics and metabolomics.* Nat Chem Biol, 2014. **10**(11): p. 963-8.
225. Cruz-Morales, P., et al., *The genome sequence of Streptomyces lividans 66 reveals a novel tRNA-dependent peptide biosynthetic system within a metal-related genomic island.* Genome Biol Evol, 2013. **5**(6): p. 1165-75.
226. Ohnishi, Y., et al., *Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350.* J Bacteriol, 2008. **190**(11): p. 4050-60.
227. Hutchings, M.I., A.W. Truman, and B. Wilkinson, *Antibiotics: past, present and future.* Curr Opin Microbiol, 2019. **51**: p. 72-80.

228. Blin, K., et al., *Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters.* Brief Bioinform, 2019. **20**(4): p. 1103-1113.
229. Zhu, H., et al., *Eliciting antibiotics active against the ESKAPE pathogens in a collection of actinomycetes isolated from mountain soils.* Microbiology (Reading), 2014. **160**(Pt 8): p. 1714-1725.
230. Medema, M.H., E. Takano, and R. Breitling, *Detecting sequence homology at the gene cluster level with MultiGeneBlast.* Mol Biol Evol, 2013. **30**(5): p. 1218-23.
231. Gu, W., et al., *The Biochemistry and Structural Biology of Cyanobactin Pathways: Enabling Combinatorial Biosynthesis.* Methods Enzymol, 2018. **604**: p. 113-163.
232. Leikoski, N., et al., *Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides.* Chem Biol, 2013. **20**(8): p. 1033-43.
233. Chekan, J.R., et al., *Characterization of the macrocyclase involved in the biosynthesis of RiPP cyclic peptides in plants.* Proc Natl Acad Sci U S A, 2017. **114**(25): p. 6551-6556.
234. Craik, D.J. and U. Malik, *Cyclotide biosynthesis.* Curr Opin Chem Biol, 2013. **17**(4): p. 546-54.
235. Zhang, Y., et al., *A distributive peptide cyclase processes multiple microviridin core peptides within a single polypeptide substrate.* Nat Commun, 2018. **9**(1): p. 1780.
236. Umemura, M., et al., *Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus.* Fungal Genet Biol, 2014. **68**: p. 23-30.
237. Nagano, N., et al., *Class of cyclic ribosomal peptide synthetic genes in filamentous fungi.* Fungal Genet Biol, 2016. **86**: p. 58-70.
238. Ding, W., et al., *Biosynthetic investigation of phomopsins reveals a widespread pathway for ribosomal natural products in Ascomycetes.* Proc Natl Acad Sci U S A, 2016. **113**(13): p. 3521-6.
239. Quijano, M.R., et al., *Distinct Autocatalytic alpha- N-Methylating Precursors Expand the Borosin RiPP Family of Peptide Natural Products.* J Am Chem Soc, 2019. **141**(24): p. 9637-9644.
240. Gondry, M., et al., *Cyclodipeptide synthases are a family of tRNA-dependent peptide bond-forming enzymes.* Nat Chem Biol, 2009. **5**(6): p. 414-20.
241. Francklyn, C.S. and A. Minajigi, *tRNA as an active chemical scaffold for diverse chemical transformations.* FEBS Lett, 2010. **584**(2): p. 366-75.
242. Garg, R.P., et al., *Molecular characterization and analysis of the biosynthetic gene cluster for the azoxy antibiotic valanimycin.* Mol Microbiol, 2002. **46**(2): p. 505-17.
243. Garg, R.P., et al., *Identification, characterization, and bioconversion of a new intermediate in valanimycin biosynthesis.* J Am Chem Soc, 2009. **131**(28): p. 9608-9.

244. Olano, C., et al., *Biosynthesis of the angiogenesis inhibitor borrelidin by Streptomyces parvulus Tu4055: cluster analysis and assignment of functions.* Chem Biol, 2004. **11**(1): p. 87-97.
245. Bibb, M.J., *Regulation of secondary metabolism in streptomycetes.* Curr Opin Microbiol, 2005. **8**(2): p. 208-15.
246. Wu, C., et al., *Discovery of C-Glycosylpyranonaphthoquinones in Streptomyces sp. MBT76 by a Combined NMR-Based Metabolomics and Bioinformatics Workflow.* J Nat Prod, 2017. **80**(2): p. 269-277.
247. Santos, C.L., et al., *A walk into the LuxR regulators of Actinobacteria: phylogenomic distribution and functional diversity.* PLoS One, 2012. **7**(10): p. e46758.
248. Swiatek, M.A., et al., *Functional analysis of the N-acetylglucosamine metabolic genes of Streptomyces coelicolor and role in control of development and antibiotic production.* J Bacteriol, 2012. **194**(5): p. 1136-44.
249. Gubbens, J., et al., *Natural product proteomining, a quantitative proteomics platform, allows rapid discovery of biosynthetic gene clusters for different classes of natural products.* Chem Biol, 2014. **21**(6): p. 707-18.
250. Du, C. and G.P. van Wezel, *Mining for Microbial Gems: Integrating Proteomics in the Postgenomic Natural Product Discovery Pipeline.* Proteomics, 2018. **18**(18): p. e1700332.
251. Bartholomae, M., et al., *Major gene-regulatory mechanisms operating in ribosomally synthesized and post-translationally modified peptide (RiPP) biosynthesis.* Mol Microbiol, 2017. **106**(2): p. 186-206.
252. Li, M.H., et al., *Automated genome mining for natural products.* BMC Bioinformatics, 2009. **10**: p. 185.
253. Culp, E.J., et al., *Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics.* Nat Biotechnol, 2019. **37**(10): p. 1149-1154.
254. Wang, M., et al., *Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking.* Nat Biotechnol, 2016. **34**(8): p. 828-837.
255. Ernst, M., et al., *MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools.* Metabolites, 2019. **9**(7).
256. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v4: recent updates and new developments.* Nucleic Acids Res, 2019. **47**(W1): p. W256-W259.
257. Blin, K., et al., *The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters.* Nucleic Acids Res, 2019. **47**(D1): p. D625-D630.
258. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
259. Xie, H.F., et al., *Feature-Based Molecular Networking Analysis of the Metabolites Produced by In Vitro Solid-State Fermentation Reveals Pathways for the Bioconversion of Epigallocatechin Gallate.* J Agric Food Chem, 2020. **68**(30): p. 7995-8007.

260. van der Hooft, J.J., et al., *Topic modeling for untargeted substructure exploration in metabolomics.* Proc Natl Acad Sci U S A, 2016. **113**(48): p. 13738-13743.
261. Li, Y. and S. Rebuffat, *The manifold roles of microbial ribosomal peptide-based natural products in physiology and ecology.* J Biol Chem, 2020. **295**(1): p. 34-54.
262. Nett, M., H. Ikeda, and B.S. Moore, *Genomic basis for natural product biosynthetic diversity in the actinomycetes.* Nat Prod Rep, 2009. **26**(11): p. 1362-84.
263. van Santen, J.A., et al., *Microbial natural product databases: moving forward in the multi-omics era.* Nat Prod Rep, 2020.
264. Kautsar, S.A., et al., *BiG-FAM: the biosynthetic gene cluster families database.* Nucleic Acids Res, 2020.
265. Umemura, M., H. Koike, and M. Machida, *Motif-independent de novo detection of secondary metabolite gene clusters-toward identification from filamentous fungi.* Front Microbiol, 2015. **6**: p. 371.
266. Takeda, I., et al., *Motif-independent prediction of a secondary metabolism gene cluster using comparative genomics: application to sequenced genomes of Aspergillus and ten other filamentous fungal species.* DNA Res, 2014. **21**(4): p. 447-57.
267. Luo, S. and S.H. Dong, *Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products.* Molecules, 2019. **24**(8).
268. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
269. Hillenmeyer, M.E., et al., *Evolution of chemical diversity by coordinated gene swaps in type II polyketide gene clusters.* Proc Natl Acad Sci U S A, 2015. **112**(45): p. 13952-7.
270. Nakai, T., et al., *The Radical S-Adenosyl-L-methionine Enzyme QhpD Catalyzes Sequential Formation of Intra-protein Sulfur-to-Methylene Carbon Thioether Bonds.* J Biol Chem, 2015. **290**(17): p. 11144-66.
271. Ma, S. and Q. Zhang, *Linaridin natural products.* Nat Prod Rep, 2020.
272. Kudo, K., et al., *Comprehensive Derivatization of Thioviridamides by Heterologous Expression.* ACS Chem Biol, 2019. **14**(6): p. 1135-1140.
273. Dischinger, J., S. Basi Chipalu, and G. Bierbaum, *Lantibiotics: promising candidates for future applications in health care.* Int J Med Microbiol, 2014. **304**(1): p. 51-62.
274. van Heel, A.J., M. Montalban-Lopez, and O.P. Kuipers, *Evaluating the feasibility of lantibiotics as an alternative therapy against bacterial infections in humans.* Expert Opin Drug Metab Toxicol, 2011. **7**(6): p. 675-80.
275. van Santen, J.A., et al., *The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery.* ACS Cent Sci, 2019. **5**(11): p. 1824-1833.
276. Donia, M.S., et al., *A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics.* Cell, 2014. **158**(6): p. 1402-1414.

Curriculum vitae

Alexander Kloosterman was born in Leiden on December 28th, 1990. After finishing high school at the Stedelijk Gymnasium in Leiden in 2009, he started a Bachelor Life Science and Technology, a shared programme between Leiden University and the Technical University of Delft. While in college, Alexander participated in the Honour's College programme in Philosophy in Leiden, and was awarded the Young Talent Encouragement Award in his first year. After joining Marcellus Ubbink's lab for an internship on protein-protein interactions and protein NMR, Alexander obtained his Bachelor's degree in July 2012. Afterwards, he continued his studies by starting a research Master Life Science and Technology in Leiden. During this time, he joined Gilles van Wezel's lab for his first research internship, under the supervision of Kasia Celler, focusing on flotillins in streptomycetes. His second internship took place in Gregory Schneider's lab, under supervision of Hadi Arjmandi Tash, where he studied the biaxial compression of graphene using lipids, before obtaining his Master's on February 2016. In December 2015, he started his PhD study at Leiden University, on the project Syngenopep, in collaboration with the Leiden University Medical Center (LUMC), the University of Groningen (RUG), supported by BaseClear, Dupont and EnzyPep, and funded by the Dutch Research Organization (NWO). Under the joint supervision of Gilles van Wezel and Marnix Medema, he worked on the discovery of leads for novel antimicrobials, specifically focusing on post-translationally modified peptides combining bioinformatics and machine learning with chemical and molecular biological tools. The work on this topic is presented in this thesis. After his PhD, Alexander has started working as a postdoctoral researcher in the lab of Björn Högberg at the Karolinska Institute in Stockholm, Sweden, on the topic of DNA sequencing microscopy.

List of publications

Alexander M. Kloosterman, Kyle Shelton, Gilles van Wezel, Marnix Medema, and Douglas Mitchell. *RRE-Finder: a genome-mining tool for class-independent RiPP discovery.* mSystems. 2020. **5**(5): e00267-20.

Alexander M. Kloosterman. Peter Cimermancic, Somayah S. Elsayed, Chao Du, Michalis Hadjithomas, Mohamed S. Donia, Michael A. Fischbach, Gilles P. van Wezel, and Marnix H. Medema. *Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides.* PLoS. Biol., 2020. **18**(12): e3002016.

Alexander M. Kloosterman, Marnix H. Medema, Gilles P. van Wezel. *Omics-based strategies to discover novel classes of RiPP natural products*. Curr. Opin. Biotechnol., 2021. **6**, p: 60-67.

Sanne Westhoff, **Alexander M. Kloosterman**, Stephan F.A. van Hoesel, Gilles P. van Wezel, Daniel Rozen. *Competition sensing changes antibiotic production in* Streptomyces. mSystems., 2021. **12**(1): e02729-20.

Lizah van der Aart, Imen Nouioui, **Alexander M. Kloosterman**, José-Mariano Igual, Joost Willemse, Michael Goodfellow, Gilles van Wezel. *Polyphasic classification of the gifted natural product producer* Streptomyces roseifaciens *sp. nov.* Int. J. Syst. Evol. Microbiol., 2019. **69**(4): p. 899-908.

Yoshitaka Hiruma, Ankur Gupta, **Alexander M. Kloosterman**, Caroline Olijve, Dr. Mathias A. S. Hass, Prof. Dr. Marcellus Ubbink. *Hot-spot residues in the Cytochrome P450cam-Putidaredoxin binding interface.* Chembiochem., 2014. **15**(1): p. 80-6.