MetaRiPPquest: A Peptidogenomics Approach for the Discovery of Ribosomally Synthesized and Post-translationally Modified Peptides

Hosein Mohimani[1,2], Alexey Gurevich[3], Kelsey L. Alexander[4,5], C. Benjamin Naman[4], Tiago Leão[4], Evgenia Glukhov[4], Nathan A. Moss[4], Tal Luzzatto-Knaan[6], Fernando Vargas[6], Louis-Felix Nothias[6], Nitin K. Singh[7], Jon G. Sanders[8], Rodolfo A. S. Benitez[8], Luke R. Thompson[8], Md-Nafiz Hamid[9], James T. Morton[1,8], Alla Mikheenko[3], Alexander Shlemov[3], Anton Korobeynikov[3,10], Iddo Friedberg[9], Rob Knight[1,8,11], Kasthuri Venkateswaran[7], William Gerwick[4], Lena Gerwick[4], Pieter C. Dorrestein[6,11], and Pavel A. Pevzner[1,10,11]

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA,
[2]Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA,
[3]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia,
[4]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA,
[5]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, USA,
[6]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA,
[7]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA,
[8]Department of Pediatrics, University of California San Diego, School of Medicine, La Jolla, California, USA,
[9]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, Iowa, USA,
[10]Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia,
[11]Center for Microbiome Innovation, Jacobs School of Engineering, University of California, San Diego, La Jolla, California, USA.

Abstract

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important class of natural products that include many antibiotics and a variety of other bioactive compounds. While recent breakthroughs in RiPP discovery raised the challenge of developing new algorithms for their analysis, peptidogenomic-based identification of RiPPs by combining genome/metagenome mining with analysis of tandem mass spectra remains an open problem. We present here MetaRiPPquest, a software tool for addressing this challenge that is compatible with large-scale screening platforms for natural product discovery. After searching millions of spectra in the Global Natural Products Social (GNPS) molecular networking infrastructure against just six genomic and metagenomic datasets, MetaRiPPquest identified 27 known and discovered 5 novel RiPP natural products.

Introduction

Natural products are at the center of attention as new pharmaceutical leads, as exemplified by the recent discoveries of novel classes of bioactive natural product drugs[1-4]. Complementing this, the recent launch of the Global Natural Products Social (GNPS) molecular networking infrastructure[5] brought together over a thousand laboratories worldwide that have already generated an unprecedented amount of tandem mass

spectra of natural products. However, to transform natural product discovery into a high-throughput technology and to fully realize the promise of the GNPS project, new algorithms are needed for natural products discovery[6-10]. Indeed, while spectra in the GNPS molecular network represent a gold mine for future chemical discoveries, their interpretation remains a bottleneck due to the large volume of data produced by modern mass spectrometers and unavailability of computational platforms for data processing.

The efforts present herein focus on *Ribosomally synthesized and Post-translationally modified Peptides (RiPPs)*, a rapidly expanding group of natural products with applications in pharmaceutical and food industries[11]. RiPPs are produced by *RiPP Synthetases (RiPPS)* through the *Post Ribosomal Peptide Synthesis (PRPS)* pathway[11]. RiPPs are initially synthesized as *precursor peptides*, encoded by RiPP *structural genes*. The RiPP structural genes are often quite short, making their annotation difficult[12]. A precursor peptide consists of a prefix *leader peptide* appended to a suffix *core peptide*. A leader peptide is important for recognition by the *RiPP post-translational modification enzymes* and for exporting the RiPP out of the cell. The core peptide is post-translationally modified by the RiPP biosynthetic machinery, proteolytically cleaved from the leader peptide to yield the *mature RiPP*, and exported out of the cell by transporters. The precursor peptide and the enzymes responsible for post-translational modifications (PTMs), proteolytic cleavage, and transportation usually appear in a contiguous *biosynthetic gene cluster (BGC)* of a RiPP within a microbial genome. The length of the microbial RiPP-encoding BGCs typically varies from 1,000 to 40,000 bp (average length 10,000 bp), larger than the current length of short reads generated by next generation sequencing (350bp), and making DNA assembly a critical part of any short read based RiPP discovery method.

Genome mining refers to the informatics-based structural interpretation of a natural product BGC to infer information about the natural product itself. The discoveries of coelichelin in *Streptomyces coelicolor*[13,14] and orfamide in *Pseudomonas fluorescens* Pf-5[14b,14c] were the first examples of genome mining[13,14] that were followed by discoveries of various bioactive RiPPs in microbial samples. Donia et al.[15] discovered lactocillin, a thiopeptide antibiotic from human vaginal isolates, that showed activity against vaginal pathogens. Zhao et al.[16] discovered eight novel lanthipeptides with antibiotic activity from a ruminant bacterium. Freeman et al. and Wilson et al.[17,18] used metagenome mining of a sponge to assign a BGC to the known RiPP polytheonamide, with post-translational modifications distributed across 49 residues. Thus, large-scale metagenomics projects, such as Earth Microbiome Project[19,19b], American Gut Project[20], and Human Microbiome Project[21,22,22b], have the potential to contribute to RiPP discovery, provided that improved bioinformatics tools for the enhanced identification of novel RiPPs are available. However, discovery of lactocillin and other recently identified RiPPs were not achieved by an automated process, but rather used time-consuming manual technologies that required information about individual (isolated) bacterial genomes. Our goal is to discover RiPPs directly from metagenomic information using a fully automated approach.

While recent analysis of thousands of bacterial and fungal genomes has already resulted in the discovery of many putative BGCs, including *ca.* 20,000 RiPP-encoding BGCs in the Integrated Microbial Genome Atlas of Biosynthetic Gene Clusters (IMG-ABC), connecting these BGCs to their metabolites has not kept pace with the speed of microbial genome sequencing[23]. Currently, only 35 out of these roughly 20,000 RiPP-encoding BGCs in IMG-ABC have been experimentally connected to their RiPPs[23,24]. Linking this impressive number of RiPP-encoding BGCs to unknown RiPPs requires the development of novel computational tools.

Kersten et al.[25] introduced the peptidogenomics approach to RiPP discovery, which refers to finding sequential amino acid tags from tandem mass spectra (peptidomics) and mining them in assembled DNA reads obtained from the same sample. Mohimani et al.[12] introduced RiPPquest, the first automated approach to RiPP discovery by mass spectrometry and genome mining. This tool is based on *Peptide-Spectrum Matches (PSMs)*, which are generated by aligning predicted spectra of putative RiPPs discovered by genome mining to mass spectra. If a *PSM* between a candidate RiPP from the assembled genome and a spectrum is statistically significant, then RiPPquest reports it as a putative annotation of the spectrum. RiPPquest resulted in identification of the lanthipeptide 'informatipeptin', the first natural product discovered in a fully automatic fashion by a computer. However, RiPPquest has a number of limitations: (a) RiPPquest is limited to lanthipeptides which constitutes only one of 19 classes of RiPPs[11], (b) RiPPquest is designed for small genomes and small spectral datasets, making it rather slow in the case of large metagenomic datasets and the

entire GNPS infrastructure, (c) RiPPquest does not report the statistical significance of identified RiPPs, the key requirement for any high-throughput peptide identification tool, and (d) RiPPquest is limited to searches for a predefined set of post-translational modification (PTMs) and does not enable blind searches for new PTMs.

Here we present MetaRiPPquest for high-throughput RiPP discovery that improves on RiPPquest in several critical aspects: (a) MetaRiPPquest identifies several important classes of lanthipeptides, lassopeptides, linear azole containing peptides (LAPs), linaridins, glycocins, cyanobactins, and proteusins (versus only lanthipeptides by RiPPquest), (b) algorithmic improvements in MetaRiPPquest increased speed by two orders of magnitude compared to RiPPquest, thus enabling searches of the entire GNPS database against metagenomes, (c) MetaRiPPquest implements a new approach for estimating statistical significance of identified PSMs, (d) MetaRiPPquest is capable of searching for RiPPs with unusual modifications in a blind mode, (e) MetaRiPPquest, in contrast to RiPPquest (which was designed for analyzing low-resolution spectra), utilizes the power of high-resolution mass spectrometry, and (f) MetaRiPPquest features a web-based interface through the GNPS infrastructure.

Results

Brief description of MetaRiPPquest. Figure 1 shows the MetaRiPPquest pipeline, which includes the following steps (see Methods section): (a) selecting a biological sample (an isolated microbe or a bacterial/fungal community), (b) generating DNA sequence data, (c) assembly of DNA sequence data with SPAdes[26] or MetaSPAdes[27], (d) identifying putative RiPP precursor peptides using antiSMASH[28] and BOA[28a] (Bacteriocin gene block and Operon Associator), and constructing a database of putative RiPP structures (as well as a decoy database), (e) generating tandem mass spectra for samples, (f) matching spectra against the constructed RiPP structure database using Dereplicator[29], and (g) enlarging the set of described RiPPs via spectral networking[30,31].

Datasets. We analyzed the following paired datasets of spectra and genome/metagenome data (all datasets, with the exception of the BACIL dataset, contain high-resolution spectra):

• *Standard dataset (STANDARD*). This small dataset consists of 18 spectra of known RiPPs that were used for benchmarking MetaRiPPquest (GNPS datasets MSV000079506 and MSV000079622). Spectra were collected from purified RiPPs from *Prochlorococcus marinus* MIT 9313 (four analogs of prochlorosins), *Geobacillus thermodenitrificans* NG80 (geobacillin), *Bacillus subtilis* NCIB 3610 (sublancin), *Bacillus halodurans* C-125 (haloduracin), *Lactococcus lactis* (lacticin), *Bacillus cereus* SJ1 (two analogs of bicereusins) and *Ruminococcus flavefaciens* FD-1 (eight analogs of flavecins). For these strains, we used genome sequence information available from the NCBI RefSeq database. AntiSMASH identified 70 BGCs in these genomes, including 29 RiPP-encoding BGCs. Since the genome sequence of the lacticin producer is not available, we searched its spectrum against its described biosynthetic gene cluster [31x].

• *Actinomycetes dataset (ACTI)*. This dataset consists of 473,135 spectra from bacterial extracts of 36 *Actinomycetales* strains with sequenced genomes[32,33] (GNPS datasets MSV000078839 and MSV000078604). We downloaded sequence information for these 36 genomes from the NCBI RefSeq database. AntiSMASH identified 1,140 BGCs in these genomes, including 168 RiPP-encoding BGCs. Furthermore, we downloaded and mixed the short reads from 21 out of 36 strains that were available from the NCBI Short Reads Archive (read length 150 bp, insert sizes varying between 200bp to 300bp). We randomly down-sampled each dataset to 10 million reads (resulting in an approximate 300-fold coverage), and mixed all the reads to simulate a metagenomic dataset for this sample. Running MetaRiPPquest on the separate genomes from this dataset resulted in the same set of identified RiPPs as obtained from the simulated metagenome.

• *Bacillus dataset (BACIL).* This dataset consists of 40,051 low-resolution spectra from bacterial extracts of two *Bacillus* strains with known genomes[34] (MSV000078552). We downloaded genome sequence information for these isolates from the NCBI RefSeq dataset.

• *Space station dataset (SPACE).* This dataset consists of 58,422 spectra from bacterial extracts of 21 isolated strains from the international space station (MSV000080102). Among these strains, twelve are *Staphylococcus*, six *Bacillus*, four *Enterobacteria* and one *Acinetobacter* strain. The complete genomes are available for all of these strains.[35,36]

• *Sponge dataset (SPONGE).* This dataset contains 223,135 spectra from bacterial extracts of *Theonella swinhoei* (GNPS dataset MSV000078670). Wilson et al.[18] used the SPONGE dataset to analyze for the RiPP polytheonamide. We searched spectra from the SPONGE dataset against the genome of *Theonella swinhoei* symbiont *Candidatus Entotheonella* sp. TSY1. AntiSMASH identified 27 BGCs in the symbiont genome, including one RiPP-encoding BGC.

• *Cyanobacteria dataset (CYANO).* This dataset consists of 11,921,457 spectra from the extracts of 317 cyanobacterial samples[37] (GNPS dataset MSV000078568). Each sample represents a mini-metagenome[27,38] with one or a few highly abundant strains. The metagenomic reads were collected from 195 of these samples.

Genome mining. MetaRiPPquest uses antiSMASH and BOA for identification of RiPP-encoding BGCs and has two genome mining modes for selecting Open Reading Frames (ORFs), a slow all-ORF mode introduced in RiPPquest[12], and a new fast motif-ORF mode. The all-ORF approach analyzes all short ORFs within a BGC, while the motif-ORF approach relies on RiPP motif finding[39] to narrow the set of putative RiPP-encoding ORFs.

We illustrate positive and negative features of these approaches through genome mining of the *Streptomyces roseosporous* NRRL 11379 genome obtained from the ACTI dataset. AntiSMASH found 30 BGCs in this genome, including six RiPP-encoding BGCs in this genome. Within these six BGCs, the motif-ORF approach identified only two short ORFs matching core RiPP motifs, while the all-ORF approach identified 14,694 short ORFs. When analyzing all 36 of the ACTI strains, antiSMASH discovered 1,140 BGCs, including 168 RiPP-encoding BGCs. MetaRiPPquest in the motif-ORF and all-ORF modes identified 67 and 565,138 short ORFs, respectively. This example illustrates that the motif-ORF mode may result in a four order of magnitude reduction in the number of ORF candidates as compared to the all-ORF mode. However, antiSMASH predictions are based on searching for a set of known motifs, therefore the motif-ORF mode misses some ORFs with novel RiPP motifs. BOA is based on identifying known proximal genes ("context genes") that reside next to the RiPP, rather than by the RiPP sequence itself. In that manner, BOA identifies non-orthologous RiPP replacements if those RiPPs maintain homologous context genes. However, if the RiPPs do not have context genes, BOA may not detect those RiPPs. Also, since BOA is trained on bacteriocin context genes only, it is especially suited for that type of RiPPs.

Although the all-ORF mode searches a larger set of ORFs than the motif-ORF mode, it does not necessarily result in an increased number of identified RiPPs after matching ORFs against the spectral dataset. Indeed, the PSMs that are statistically significant in the motif-ORF mode may become statistically insignificant in the all-ORF mode because the search space in the all-ORF mode is orders of magnitude larger than in the motif-ORF mode resulting in an increased false discovery rate (FDR). Because MetaRiPPquest only reports statistically significant PSMs, the all-ORF mode may miss some peptides identified in the motif-ORF mode. Conversely, because MetaRiPPquest searches more ORFs in the all-ORF mode than in the motif-ORF mode, the motif-ORF mode may miss some peptides identified in the all-ORF mode.

RiPP identification. Below we describe applications of MetaRiPPquest to various datasets:

STANDARD. For STANDARD dataset in the all-ORF mode, MetaRiPPquest identified 18 RiPPs including prochlorosin[40], geobacillin[41], sublancin[42], haloduracin[43], lacticin 481[44], bicereucin[45], flavecin[16] with p-values ranging from $8•10^{-14}$ to $5•10^{-55}$ (Table 1). In contrast, MetaRiPPquest in the motif-ORF mode identified only five out of 18 RiPPs since antiSMASH failed to predict 13 out of 18 RiPP-encoding ORFs.

ACTI. Figure 2 shows a comparison of performance of MetaRiPPquest with all-ORF and motif-ORF genome mining approaches on the ACTI dataset. At the extremely conservative 0% FDR, MetaRiPPquest in the motif-ORF mode identified three novel RiPPs and five known RiPPs. The five known RiPPs include the linaridin grisemycin at p-value of $3•10^{-36}$ (from *Streptomyces griseus* IFO 13350[25,46]), lantibiotics AmfS and SRO-3108 at p-value of $4•10^{-12}$ and $9•10^{-12}$ (from *Streptomyces roseosporous*[25,47]), lantibiotic informatipeptin at p-value $5•10^{-12}$ (from *Streptomyces viridochromogenes*[12]), and the lantibiotic SapB at p-value of $4•10^{-12}$ (from *Streptomyces coelicolor* A3(2)[48]). The three novel RiPPs include a class II lantibiotic (referred to as Compound X) from *Streptomyces sp.* CNT360 with p-value $6•10^{-13}$, an informatipeptin-like lantibiotic (referred to as informatipeptin B) from *Streptomyces cattleya* with p-value $5•10^{-12}$, and a lassopeptide (refered to as Compound Y) from *Streptomyces viridochromogenes* with p-value $5•10^{-31}$. MetaRiPPquest in the all-ORF mode identified only two known RiPPs at 0% FDR (grisemycin and AmfS). Note that while the all-ORF mode improves on the motif-ORF mode for the STANDARD dataset, the motif-ORF mode improves on the all-ORF mode for the ACTI dataset. MetaSPAdes assembled the simulated ACTI metagenome into 6,204 contigs with lengths greater than 1000 bp, with total length of 99.9 Mb and N50 of around 60 kb. The longest contig was approximately 756 kb. AntiSMASH identified 353 BGCs (66 RiPPs) in the assembled metagenome. All of the identified RiPPs were re-identified using the simulated metagenome rather than individual genomes.

BACIL. AntiSMASH identified 12 BGCs (one RiPP-encoding BGCs) in *B. amyloliquefaciens* FZB42 and 11 BGCs (four RiPP-encoding BGCs) in *B. licheniformis* ES-221. MetaRiPPquest identified two known RiPPs from the BACIL dataset. Lichenicidin A, a class II lantibiotic[51], was identified from a novel producer *Bacillus licheniformis* ES-221. Plantzocilin A, a linear azole containing peptide[49,50], was identified from the known producer *Bacillus amyloliquefaciens* FZB42.

SPACE. AntiSMASH identified 119 BGCs, including 27 RiPP-encoding BGCs, in this dataset. MetaRiPPquest in the all-ORF mode identified one novel lantibiotic named 'compound Z' from three *Bacillus* strains (*Bacillus* sp. ISSFR-3F, *Bacillus* sp. S1-R2-T1 and *Bacillus* sp. S1-R3-J1).

SPONGE. AntiSMASH identified 27 BGCs (including four RiPP-encoding BGCs) in the SPONGE metagenome. MetaRiPPquest identified the known proteusin RiPP polytheonamide[17,52] with a p-value 10^{-20} encoded by one of these four BGCs.

CYANO. AntiSMASH identified 2,898 BGCs in the 195 cyanobacterial metagenomes, including 491 RiPP-encoding BGCs. MetaRiPPquest identified the known RiPP wewakazole[53] and a novel RiPP named cyanobactin X in all-ORF mode.

Novel RiPPs. Below we describe five novel RiPPs and wewakazole, a known RiPP with a novel gene cluster, identified by MetaRiPPquest.

Informatipeptin B (**N**GGGAST**V**SLLSCVSA**G**SV**I**LC**V**) is a novel lantibiotic identified with a p-value of $5•10^{-12}$ (Figure 3) in the ACTI dataset. Informatipeptin B differs from informatipeptin[12] in only 5 amino acids (shown in bold). The most abundant analog of informatipeptin B (mass of 1870.95 Da) corresponds to six dehydrations (unmodified peptide mass of 1979.01). The BGC of informatipeptin B is similar to the BGCs encoding class III lantibiotics (AmfS and SapB). Three out of 10 nodes in the spectral network of informatipeptin B have been identified as its analogs: informatipeptin B1 (addition of N-terminal amino acid N), informatipeptin B2 (truncation by one amino acid), and informatipeptin B3 (truncation by two amino acids). These analogs of informatipeptin B (with stepwise N-terminal leader processing) provide additional evidence that informatipeptin B is a novel RiPP. Four other nodes in the spectral network have been identified as sodium and potassium adducts (+22 Da and +38 Da mass shifts).

Compound X (DTGGCSGLCTVLVCTVIVC) is a novel lantibiotic that was identified with p-value $6•10^{-13}$ in the ACTI dataset and that shows no homology to any of the known RiPPs (Figure 4). One of the genes within its BGC shows similarity to genes encoding class I lantibiotics (e.g., subtilin). The most abundant analog of Compound X (mass 1769.80 Da) corresponds to four dehydrations (unmodified peptide mass of 1841.84). The spectral network of Compound X revealed an additional Compound X1 analog (truncation by one amino acid).

Compound Y (LLGRHGNDRLILSKN) is a novel lassopeptide that was identified with a p-value of $5•10^{-31}$ in the ACTI dataset and that shows distant homology to microcin J25 (Figure 5). Two of the genes in its BGC are similar to microcin J25 hypothetical genes McjB and McjC from *S. avertimilis* with a 78% identity for both. The most abundant analog of Compound Y (mass 1645.962 Da) corresponds to a -59 Da modification (unmodified peptide mass of 1704.964 Da).

Compound Z (DATITTVTVTSTSIWASTVSNHC) is a new RiPP identified in the SPACE dataset that shows no similarity to any known RiPP. It was not identified in the motif-ORF mode since antiSMASH failed to identify the ORF encoding this RiPP. The gene cluster of Compound Z shows similarity to enterotoxin genes. While enterotoxin gene clusters from *Escherichia coli* are known to produce RiPPs[9,54], this is the first evidence for production of a RiPP by a *Bacillus* enterotoxin BGC.

The most abundant analog of Compound Z (mass 2109.064 Da) corresponds to eleven dehydrations and a -87Da N-terminal modification (unmodified peptide mass 2394.126 Da). MetaRiPPquest assigned a p-value of $3•10^{-25}$ to the PSM formed by Compound Z (Figure 6). There are multiple identical ORFs encoding Compound Z in the lantibiotic gene clusters of strains ISSFR-3F (13 copies), S1-R2-T1 (3 copies) and S1-R3-J1 (3 copies). Using spectral networks, MetaRiPPquest detected two less abundant analogs of Compound Z with -1 Da and -15 Da N-terminal modifications, instead of the -87 Da N-terminal modification in the most abundant analog. Moreover, there are analogs corresponding to the sequence DATITTVTVT with five dehydrations and a -87 Da or -15 Da N-terminal modification, and a analog corresponding to the alternative ORF DATITTVTVTSTSIWASTVSNYC in the same lantibiotic gene cluster with a single H to Y mutation.

Cyanobactin X is a cyclic peptide ISNGYLIP (mass 857.47 Da) with p-value $2•10^{-17}$ identified in strain PNG22APR06-1 of the CYANO dataset (Figure 7). The cyanobactin X core peptide has no similarity to any of the known cyanobactins. Its BGC, encoded in an 8 kb contig, is missed by antiSMASH. The spectrum of cyanobactin X does not cluster with any other spectra in the spectral network.

Wewakazole is a cyclic dodecapeptide IS−20APPGVT−20FS−20FP with mass 1140.54 Da, originally discovered in *Lyngbya majuscula* from Papua New Guinea[52] (S−20 and T−20 stand for oxazole and methyl-oxazole, respectively). MetaRiPPquest identified wewakazole and its BGC with p-value $2•10^{-22}$ in strain PNG19MAY05-2/7 (Figure 8). AntiSMASH failed to predict any precursor peptide for this gene cluster.

Confirmation of wewakazole identification. MetaRiPPquest identified wewakazole in a polar fraction from the extract of strain PNG19MAY05-2/7, a marine cyanobacterium collected at Kape Point, Papua New Guinea. Wewakazole was first reported by one of our groups from another Papua New Guinea collection of *Lyngbya majuscula* (revised to *Moorea producens*) obtained from Wewak Bay[52]. Subsequently a related compound, wewakazole B was isolated from a Red Sea collection of this cyanobacterium[55]. To confirm and to validate the accuracy of MetaRiPPquest to find and identify new compounds from strain PNG19MAY05-2/7, reverse phase C_{18} column chromatography and preparative HPLC separations were successful in the isolation of 31.2 µg of this compound. The compound possessed the same molecular formula as wewakazole, $C_{59}H_{72}N_{12}O_{12}$, based on the molecular ion sodium adduct $[M+Na]^+$ in the HRESIMS *(m/z* 1163.5282, Supplementary Figure 1). Its chemical identity was further confirmed utilizing ^{1}H, HSQC and HMBC NMR data, which allowed for direct comparison with data previously reported for wewakazole (Supplementary Figure 2, 3, 4)[52]. Moreover, the tandem mass spectrum and retention time of the isolated compound matched the data previously reported for wewakazole (Supplementary Figures 5, 6)[52]. Furthermore, the ECCD spectrum resembled that of wewakazole B[53], and the specific rotation showed the same sign as previously reported for wewakazole[52], excluding the possibility of an enantiomeric relationship of this isolate to that of wewakazole (Supplementary Figure 7). Thus, the compound identified by MetaRiPPquest was isolated and its identity was confirmed as wewakazole.

Comparison of extraction/fractionation strategies. MetaRiPPquest can assist in determining the optimal extraction strategy for novel RiPP discovery. As an example, nine strategies were used for fractionation of the samples, and among them strategy H, 25% Methanol and 75% Ethyl acetate, is the only strategy capable of detecting both wewakazole and Cyanobactin X.

Discussion

While recent genome mining efforts have revealed over 20,000 hypothetical RiPP-encoding BGCs[23], only 35 RiPPs have been identified that match to these BGCs. To keep pace with the speed of microbial genome sequencing, high-throughput methods for structure elucidation of RiPPs by combining metagenomics, genome mining, and peptidomics are needed. MetaRiPPquest extends our previous RiPPquest tool (limited to lanthipeptides) to lassopeptides, LAPs, linaridins, glycocins, proteusins, and cyanobactins, and enables the blind search for RiPPs with unusual modifications.

Articles describing RiPPs are usually limited to the analysis of a single peptide or a few related peptides. The first application of MetaRiPPquest revealed many known RiPPs, as well as their unknown analogs, and five novel RiPPs (three lantibiotics, one lassopeptide, and one cyanobactin) along with their numerous analogs, from only six spectral datasets. This result provides optimism that MetaRiPPquest can potentially make RiPP identification as robust as peptide identification in traditional proteomics. The increased robustness was validated by the isolation of the RiPP metabolite wewakezole, and its structure was confirmed by orthogonal approaches, confirming that the MetaRiPPquest prediction was correct. In contrast to the existing genome mining approaches that rely on known BGC motifs[28], MetaRiPPquest in the all-ORF mode has the ability discover new BGCs (with previously unknown motifs) that encode novel RiPPs (e.g. compound Z and cyanobactin X) that are very different from currently known RiPPs and thus are not captured by the existing genome mining tools.

Finally, algorithmic improvements in MetaRiPPquest have resulted in a two orders of magnitude increase in speed compared to RiPPquest, thus enabling searches of the entire GNPS infrastructure against metagenomic information.

Code availability. MetaRiPPquest is available both as a command line tool and as a web application at http://metarippquest.metabologenomics.org .

References

1. J.W. Li and J.C. Vederas. Drug discovery and natural products: end of an era or an endless frontier? Science, 325:161–165, 2009.

2. M.A. Fischbach and C.T. Walsh. Antibiotics for emerging pathogens. Science, 325:1089–93, 2009.

3. L.L. Ling, T. Schneider, A.J. Peoples, A.L. Spoering, I. Engels, B.P. Conlon, A. Mueller, T.F. Schberle, D.E. Hughes, S. Epstein, M. Jones, L. Lazarides, V.A. Steadman, D.R. Cohen, C.R. Felix, K.A. Fetterman, W.P. Millett, A.G. Nitti, A.M. Zullo, C. Chen, and K. Lewis. A new antibiotic kills pathogens without detectable resistance. Nature, 517:455–459, 2015.

4. A.L. Harvey, R. Edrada-Ebel, and R.J. Quinn. The re-emergence of natural products for drug discovery in the genomics era. Nat. Rev. Drug Discov., 14:111–129, 2015.

5. M. Wang et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol., 34:828–837, 2016.

6. A. Vaniya and O. Fiehn. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. Trends Anal. Chem., 69:52–61, 2015.

7. H. Mohimani and P.A. Pevzner. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectal networks. Nat. Prod. Rep., 33:73–86, 2016.

8. M. H. Medema and M. A. Fischbach. Computational approaches to natural product discovery. Nat. Chem. Biol., 11:639–648, 2015.

9. M. S. Donia and M. A. Fischbach. Small molecules from the human microbiota. Science, 349:1254766, 2015.

10. C. T. Walsh. A chemocentric view of the natural product inventory. Nat. Chem. Biol., 11:620–624, 2015.

11. P.G. Arnison, Bibb M.J., G. Bierbaum, Bowers A.A., Bulaj G., J.A. Camarero, D.J. Campopiano, J. Clardy, P.D. Cotter, Craik D.J., E. Dittmann, S. Donadio, P.C. Dorrestein, K.D. Entian, M.A. Fischbach, J.S. Garavelli, U. Gransson, C.W. Gruber, D.H. Haft, T.K. Hemscheidt, C. Hertweck, C. Hill, A.R. Horswill, M. Jaspars, W.L. Kelly, Klinman J.P., Kuipers O.P., A.J. Link, W. Liu, M.A. Marahiel, D.A. Mitchell, G.L. Moll, B.S. Moore, S.K. Nair, I.F. Nes, G.E. Norris, B.M. Olivera, H. Onaka, M.L. Patchett, M.J.T. Reaney, S. Rebuffat, R.P. Ross, Sahl H.G., E.W. Schmidt, M.E. Selsted, K. Severinov, B. Shen, K. Sivonen, L. Smith, T. Stein, R.E. Sssmuth, J.R. Tagg, G.L. Tang, J.C. Vederas, C.T. Walsh, J.D. Walton, J.M. Willey, and W.A. van der Donk. Ribosomally synthesized and post-translationally modified peptide natural products: Overview and recommendations for a universal nomenclature. Natural Product Reports, 30:108–160, 2013.

12. H. Mohimani, R. Kersten, W.T. Liu, M. Wang, S.O. Purvine, S. Wu, H.M. Brewer, L. Pasa-Tolic, B.S. Moore, P.A. Pevzner, and P.C. Dorrestein. Automated genome mining of ribosomal peptide natural products. ACS Chemical Biology, 9:1545–1551, 2014.

13. G. Challis and J. Ravel. Coelichelin, a new peptide siderophore encoded by the streptomyces coelicolor genome: structure prediction from the sequence of its non ribosomal peptide synthetase. FEMS Microbiology Letter, 187:111–114, 2000.

14. S. Lautru, R. Deeth, L. Bailey, and G. Challis. Discovery of a new peptide natural product by streptomyces coelicolor genome mining. Nat. Chem. Biol., 1:265–269, 2005.

14a. C.M. Paulsen, et al. Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. Nat. Biotechnol., 23:873-878, 2005.

14b. H. Gross, V.O. Stockwell, M.D. Henkels, B. Nowak-Thompson, J.E. Loper, W.H. Gerwick. The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Biol. 14:53-63, 2007.

15. M. S. Donia, P. Cimermancic, C. J. Schulze, L. C. Wieland Brown, J. Martin, M. Mitreva, J. Clardy, R. G. Linington, and M. A. Fischbach. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell, 158:1402–1414,
2014.

16. X. Zhao 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., 23:246–56, 2016.

17. M.F. Freeman, C. Gurgui, M.J. Helf, B.I. Morinaka, A.R. Uria, N.J. Oldham, H.G. Sahl, S. Matsunaga, and J. Piel. Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science, 338:387–90, 2012.

18. M.C. Wilson et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature, 506:58–62, 2014.

19. J.A. Gilbert, J.K. Jansson, and R. Knight. The earth microbiome project: successes and aspirations. BMC Biol., 12:69, 2014.

19b. L.R. Thompson *et al.*, A communal catalogue reveals Earth's multiscale microbial diversity. Nature, DOI: 10.1038/nature24621.

20. D. McDonald, M. Hornig, C. Lozupone, J. Debelius, J.A. Gilbert, and R. Knight. Towards large-cohort comparative studies to define the factors influencing the gut microbial community structure of asd patients. Microb Ecol Health Dis., 9:26555, 2015.

21. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature, 486:207–14, 2012.

22. Human Microbiome Project Consortium. A framework for human microbiome research. Nature, 486:215–21, 2012.

22b. J. Lloyd-Price, A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A.B. Hall, A. Brady, H.H. Creasy, C. McCracken, M.G. Giglio, D. McDonald, E.A. Franzosa, R. Knight, O. White, C. Huttenhower. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature, 550:61-66, 2017.

23. M. Hadjithomas, I.A. Chen, K. Chu, A. Ratner, K. Palaniappan, E. Szeto, J. Huang, T.B.K. Reddy, P. Cimermani, M.A. Fischbach, N.N. Ivanova, V.M. Markowitz, and N.C. Kyrpides. IMG-ABC: A knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. Mbio., 6:e00932–15, 2015.

24. M. H. Medema et al. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol., 11:625–631, 2015.

25. R. Kersten, Y. Yang, P. Cimermancic, S. Nam, W. Fenical, M. Fischbach, B. Moore, and P.C. Dorrestein. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol., 7:794–802, 2011.

26. A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P.A. Pevzner. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol., 19:455–77, 2012.

27. S. Nurk, D. Meleshko, A. Korobeynikov, and P. Pevzner. MetaSPAdes: a new versatile metagenomics assembler. Genome Res., 27:824–834, 2017.

28. T. Weber, K. Blin, S. Duddela, D. Krug, H.U. Kim, R. Bruccoleri, S.Y. Lee, M.A. Fischbach, R. Muller, W. Wohlleben, R. Breitling, E. Takano, and M.H. Medema. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res., 43:W237–43, 2015.

28a. J.T. Morton, S.D. Freed, S.W. Lee, and I. Friedberg. A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. Bioinformatics, 16:381, 2015.

29. H. Mohimani, A. Gurevich, A. Mikheenko, N. Garg, L. F. Nothias, A. Ninomiya, K. Takada, P. C. Dorrestein, and P. A. Pevzner. Dereplication of peptidic natural products through database search of mass spectra. Nat. Chem. Biol., 13(1):30–37, Jan 2017.

30. N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein identification by spectral networks analysis. Proc. Natl. Acad. Sci., 104:6140–5, 2007.

31. J. Watrous, P. Roach, T. Alexandrov, B. Heath, J. Yang, R. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. Raaijmakers, B. Moore, J. Laskin, N. Bandeira, and P. Dorrestein. Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci., 109:E1743–1752, 2012.

31x. A. Rincé, A. Dufour, P. Uguen, J.P. Le Pennec and D. Haras, Characterization of the lacticin 481 operon: the Lactococcus lactis genes lctF, lctE, and lctG encode a putative ABC transporter involved in bacteriocin immunity. Appl. Environ. Microbiol. 63:4252–4260, 1997.

32. K.R. Duncan, M. Crsemann, A. Lechner, A. Sarkar, J. Li, N. Ziemert, M. Wang, N. Bandeira, B.S. Moore, P.C. Dorrestein, and P.R. Jensen. Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from salinispora species. Chem Biol., 22:460–471, 2015.

33. H. Mohimani, W.T. Liu, R. Kersten, B.S. Moore, P.C. Dorrestein, and P.A. Pevzner. Nrpquest: Coupling mass-spectrometry and genome mining for non ribosomal peptide discovery. J Nat. Prod., 77:1902–1909, 2014.

34. D.D. Nguyen, C.H. Wu, W.J. Moree, A. Lamsa, M.H. Medema, X. Zhao, R.G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, J. Ballesteros, J. S anchez, J.D. Watrous, V.V. Phelan, C. van de Wiel, R.D. Kersten, S. Mehnaz, R. De Mot, E.A. Shank, P. Charusanti, H. Nagarajan, B.M. Duggan, B.S. M oore, N. Bandeira, K. Palsson, B.. nd

Pogliano, M. Gutierrez, and P.C. Dorrestein. MS/MS networking guided analysis of molecule and gene cluster families. Proc. Natl. Acad. Sci., 110:E2611–20, 2013.

35. N.K. Singh, A. Blachowicz, A. Checinska, C. Wang, and K. Venkateswaran. Draft Genome Sequences of Two Aspergillus fumigatus Strains, Isolated from the International Space Station. *Genome Announc,* 4:e00553-16, 2016.

36. K. Venkateswaran, S. Checinska, A. Ratnayake, R.K. Pope, T.E. Blank, V.G. Stepanov, G.E. Fox, S.P. van Tongeren, C. Torres, J. Allen et al. : Draft Genome Sequences from a Novel Clade of Bacillus cereus Sensu Lato Strains, Isolated from the International Space Station. *Genome Announc,* 5:e00680-17, 2017.

37. T. Luzzatto-Knaan, N. Garg, M. Wang, E. Glukhov, Y. Peng, G. Ackermann, A. Amir, B.M. Duggan, S. Ryazanov, L. Gerwick, R. Knight, T. Alexandrov, N. Bandeira, W.H. Gerwick, and P.C. Dorrestein. Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. Elife, 11:e24214, 2017.

38. S. Nurk , A. Bankevich, D. Antipov, A.A. Gurevich, A. Korobeynikov, A. Lapidus, A.D. Prjibelski, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, S.R. Clingenpeel, T. Woyke, J.S. McLean, R. Lasken, G. Tesler, M.A. Alekseyev, and P.A. Pevzner. Assembling single-cell genomes and mini-metagenomes from chimeric mda products. J. Comput. Biol., 20:714–37,
2013.

39. K. Blin, D. Kazempour, W. Wohlleben, and T. Weber. Improved lanthipeptide detection and prediction for antiSMASH. PLoS ONE, 9:e89420, 2014.

40. W. Tang and W.A. can der Donk. Structural characterization of four prochlorosins: a novel class of lantipeptides produced by planktonic marine cyanobacteria. Biochemistry, 51:4271–9, 2012.

41. N. Garg, W. Tang, Y. Goto, S.K. Nair, and W.A. van der Donk. Lantibiotics from geobacillus thermodenitrificans. Proc Natl Acad Sci, 109:5241–6, 2012.

42. S.H. Paik, A. Chakicherla, and J.N. Hansen. Identification and characterization of the structural and transporter genes for, and the chemical and biological properties of, sublancin 168, a novel lantibiotic produced by bacillus subtilis 168. J Biol Chem., 273:23134–42, 1998.

43. A. McClerren, L. Cooper, C. Quan, P. Thomas, N. Kelleher, and W. van der Donk. Discovery and in vitro biosynthesis of haloduracin, a two-component lantibiotic. Proc. Natl. Acad. Sci., 103:17243–8, 2006.

44. A. Rince, A. Dufour, S. Le Pogam, D. Thuault, C.M. Bourgeois, and J.P. Le Pennec. Cloning, expression, and nucleotide sequence of genes involved in production of lactococcin dr, a bacteriocin from lactococcus lactis subsp. lactis. Appl Environ Microbiol., 60:1652–7, 1994.

45. L. Huo and W.A. van der Donk. Discovery and characterization of bicereucin, an unusual d- amino acid-containing mixed two-component lantibiotic. J Am Chem Soc., 138:5254–7, 2016.

46. J. Claesen and M. Bibb. Biosynthesis and regulation of grisemycin, a new member of the linaridin family of ribosomally synthesized peptides produced by streptomyces griseus ifo 13350. J. Bacteriol., 193:2510–6, 2011.

47. K. Ueda, K. Oinuma, G. Ikeda, K. Hosono, Y. Ohnishi, S. Horinouchi, and T. Beppu. Amfs, an extracellular peptidic morphogen in streptomyces griseus. J. Bacteriol., 184:1488–1492,
2002.

48. S. Kodani, M.E. Hudson, M.C. Durrant, M.J. Buttner, J.R. Nodwell, and Willey J.M. The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene rams in streptomyces coelicolor. Proc. Natl. Acad. Sci., 101:11448–11453, 2004.

49. K.J. Molohon, J.O. Melby, J. Lee, B.S. Evans, K.L. Dunbar, S.B. Bumpus, N.L. Kelleher, and D.A. Mitchell. Structure determination and interception of biosynthetic intermediates for the plantazolicin class of highly discriminating antibiotics. ACS Chem. Biol., 6:1307–13, 2011.

50. R. Scholz, K. Molohon, J. Nachtigall, J. Vater, A. Markley, R. Sussmuth, D. Mitchell, and R. Borriss. Plantazolicin, a novel microcin b17/streptolysin S-like natural product from bacillus amyloliquefaciens fzb42. Bioelectrochemistry, 193:215–224, 2011.

51. M. Begley, P.D. Cotter, C. Hill, and R.P. Ross. Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. Appl. Environ. Microbiol.,
75:5451–60, 2009.

52. L.M. Nogle, B.L. Marquez, and W.H. Gerwick. Wewakazole, a novel cyclic dodecapeptide from a papua new guinea lyngbya majuscula. Org. Lett., 5:3–6, 2003.

53. T. Hamada, S. Matsunaga, G. Yano, and N. Fusetani. Polytheonamides A and B, highly cytotoxic, linear polypeptides with unprecedented structural features, from the marine sponge, theonella swinhoei. J Am Chem Soc., 127:110–8, 2005.

54. H. Ozaki, T. Sato, H. Kubota, Y. Hata, Y. Katsube, and Y. Shimonishi. Molecular structure of the toxin domain of heat-stable enterotoxin produced by a pathogenic strain of escherichia coli. a putative binding site for a binding protein on rat intestinal epithelial cell membranes. J. Biol. Chem., 266:5934–5941, 1991.

55. J.A.V. Lopez, S.S. Al-Lihaibi, W.M. Alarif, A. Abdel-Lateff, Y. Nogata, K. Washio, M. Morikawa, T. Okino, Wewakazole B, a Cytotoxic Cyanobactin from the Cyanobacterium Moorea producens Collected in the Red Sea, *J. Nat. Prod.*, 79:1213−1218, 2016.

Acknowledgement

The work of H.M., P.D. and P.A.P. was supported by NIH 2-P41-GM103484. P.D. is supported by GM097509. A.G., A.M., A.S., A.K. and P.A.P. were supported by Russian Science Foundation 14-50-00069. T.L.K., P.D., L.G. and W.H.G. were supported by NIH 2R01GM107550. L.G. and W.H.G. were supported by NIH R01GM118815. JTM was funded by NSF GRFP DGE-1144086 .We thank the implementation team of the Microbial Observatory (Microbial Tracking-1) project at NASA Ames Research Center and sample processing/isolation of microbes by Aleksandra Checinska Sielaff, JPL. Part of the research described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. The work of N.K.S. and K.V. was funded by NASA 19-12829- 26 and 19-12829-27. C.B.N. was supported by postdoctoral fellowship from NCI/NIH Training Program in the Biochemistry of Growth Regulation and Oncogenesi (T32 CA009523). T.L.K was supported by Vaadia-BARD Postdoctoral Fellowship Award no. FI-494-13. T.L. was supported by CAPES Foundation for Research Fellowship (13425-13-7). The work of I.F. was supported, in part, by NSF awards ABI-1551363 and ABI-1458359.

METHODS

MetaRiPPQuest algorithm. Below we describe various steps of MetaRiPPquest.

(a) Selecting a biological sample. MetaRiPPquest works both with DNA sequencing datasets from isolate microbes and bacterial/fungi communities.

(b) Generating DNA sequencing data. MetaRiPPquest works with short Illumina reads or pre-assembled genomes/metagenomes.

(c) Assembly of DNA sequencing data. MetaRiPPquest assembles reads using SPAdes[26] or MetaSPAdes[27].

(d) Identifying putative RiPP precursor peptides and constructing a database of putative RiPP structures. MetaRiPPquest uses antiSMASH[28,56] for genome mining. AntiSMASH identifies putative RiPP-encoding BGCs, searches for known precursor peptide motifs in all ORFs within these BGCs, and constructs the set of putative core peptides. It also identifies modification enzymes in each RiPP-encoding BGC and uses the observation that RiPPs are typically encoded in a 10 kb window centered at a modification enzyme gene.

This approach usually results in a small number of putative core peptides per BGC, resulting in a fast and specific but non-sensitive peptidogenomics approach. As an alternative to this motif-ORF approach, the all-ORF approach (i) starts with all RiPP-encoding BGCs identified by antiSMASH, (ii) constructs a 10 kb windows centered at each modification enzyme of the identified RiPP-encoding BGC, and (iii) identifies all ORFs shorter than a pre-defined threshold (the default is 200 amino acids) as the putative precursor peptide[12]. This approach is sensitive but less specific than the motif-ORF approach.

After all modification enzymes in each RiPP-encoding BGC are identified, MetaRiPPquest considers the corresponding modifications for all suffixes of the identified ORFs within these BGC to construct the target RiPP structure database by considering all possible combinations of modifications (consistent with the modification enzyme) on all possible residues. The decoy RiPP structure database is similarly constructed, starting from decoy ORFs. MetaRiPPquest constructs a decoy database of precursor peptides by random shuffling of the target precursor peptide[57].

(e) Generating tandem mass spectra. Samples were extracted using various fractionations and spectra were collected using various instruments described in supplementary material.

(f) Matching spectra against the constructed RiPP structure database. MetaRiPPquest uses a modified version of Dereplicator[29] for searching spectra against the database of putative RiPP structures as follows (i) theoretical spectra for all peptides in the target and decoy RiPP structure database are constructed, (ii) PSMs are generated and scored, (iii) p-values of PSMs are computed using MS-DPR[58], (iv) false discovery rates are computed using the decoy database, (v) statistically significant PSMs are output as putative RiPP identifications.

While exhaustive generation of candidate RiPPs and scoring by Dereplicator is feasible when a small number of modifications are considered, the running time rapidly increases with the increase in the number of modifications. We use the spectral alignment technique to efficiently find modifications of the core peptide that best matches the spectrum[59-61,12]. Moreover, to handle blind modifications we adapted the unrestrictive modification search algorithm[61]. This dynamic programming approach restricts the number of modifications and penalizes high score matches with more than one modification.

While the dynamic programming approach from RiPPquest[12] can handle modifications in linear peptides, it is not applicable to cyclic peptides. MetaRiPPquest uses a brute-force approach to search all RiPP modifications of each candidate cyclic peptide against all spectra. We do not currently perform blind modification searches for cyclic peptides due to the inherent computational complexity.

(g) Enlarging the set of identified RiPPs via spectral networking. The set of found RiPPs is enlarged via spectral networks[30,31].

Confirmation of wewakazole structure. HESIRMS data was collected using an Agilent 6230 Accurate-Mass TOFMS in positive ion mode by the UCSD Chemistry and Biochemistry Mass Spectrometry Facility. UV-Vis data were recorded on a Beckman Coulter DU 800 spectrophotometer at room temperature in MeOH (λ_{max} at 214 nm and 217 nm). The ECCD spectrum was measured in MeOH using an Aviv 215 CD spectrometer. Optical rotation was measured at 25 °C using a JASCO P-2000 polarimeter ($[\alpha]^{25}$ -3.9 (*c* 0.022, MeOH) (lit.,[52] $[\alpha]^{21}_{D}$ -46.8 (*c* 0.41, MeOH)). A Bruker AVANCE III 600 MHz NMR with a 1.7 mm dual tune TCI cryoprobe was used to record ^{1}H, HMBC and HSQC NMR data at 298 °K with standard Bruker pulse sequences. A Varian Vx 500 NMR with a cold probe and z-gradients was used to record ^{1}H NMR data at 298 K with standard pulse sequences. NMR data were recorded in $CDCl_3$ and calibrated using residual solvent peaks (δ_H 7.26 and δ_C 77.16).

For LC-MS analysis, a Thermo Finnigan Surveyor HPLC System was used with a Phenomenex Kinetex 5 µm C18 100 x 4.6 mm column coupled to a Thermo-Finnigan LCQ Advantage Max Mass Spectrometer. Samples were separated using a linear gradient with (A) H_2O + 0.1% FA to (B) CH_3CN + 0.1% FA at a flow rate of 0.6 mL/min. The gradient started with a 5 min isocratic step at 30% B followed by an increase to 99% B over 17 min, which was held at 99% B for 5 min and then moved to 30% B in 1 min, and then held for 4 min. Mass spectra were acquired with an ESI source ranging from *m/z* 200-1600.

Preparative HPLC was done using a Kinetex 5 µm C18 150 x 10.0 mm semi-preparative column coupled to a Thermo Dionex Ultimate 3000 pump, RS autosampler, RS diode array detector, and automated fraction collector.

Isolation of wewakazole. The fraction in which metaRiPPquest identified wewakazole from sample PNG19MAY05-2/7 is identified here as 1648H. This fraction (26.5 mg) was initially separated using a 500 mg/8mL Xpertek® C_{18} SPE cartridge with 100% CH_3CN to yield 10.7 mg after concentration under vacuum. The compound was isolated from this eluent by semi-preparative HPLC using a linear gradient with (A) H_2O + 0.1% FA to (B) CH_3CN at a flow rate of 4 mL/min, and the chromatogram was monitored at 218 nm. The gradient started with a 5 min isocratic step at 40% B followed by an increase to 95% B in 25 min. Approximately 2.5 mg of the sample were injected per run to yield 31.2 µg of wewakazole (t_R=13.0 min).

56. M.H. Medema, K. Blin, P. Cimermancic, Jager V., P. Zakrzewski, M.A. Fischbach, T. Weber, E. Takan, and R. Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res., 39:W339–W346, 2011.

57. J.E. Elias and S.P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods, 4:207–214, 2007.

58. H. Mohimani, S. Kim, and P.A. Pevzner. A new approach to evaluating statistical significance of spectral identifications. J. Prot. Res., 12:1560–1568, 2013.

59. P. Pevzner, V. Dancik, and C. Tang. Mutation and modification-tolerant protein identification via tandem mass-spectrometry. J. Comput. Biol., 7:777–787, 2000.

60. P. Pevzner, Z. Mulyukov, V. Dancik, and C.L. Tang. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome. Res., 11:290–299, 2001.

61. D. Tsur, S. Tanner, E. Zandi, V. Bafna, and P. Pevzner. Identification of post-translational modifications via blind search of mass-spectra. Nature Biotechnology, 23:1562–1567, 2005.

Table 1: MetaRiPPquest identified 26 known and five novel RiPPs in STANDARD, BACIL, SPACE, ACTI, and SPONGE datasets. All identified RiPPs are linear except for cyclic cyanobactins. The novel RiPPs are shown in bold. The mode stands for the mode of discovery. A stands for RiPPs discovered by antiSMASH motif search, B stands for RiPPs discovered by BOA search, and E stands for RiPPs discovered by exhaustive search. Overall, 11 out of 31 RiPPs are predicted by antiSMASH motif search , 8 out of them by BOA search, and 17 only with exhaustive search.

Dataset	RiPP class	core peptide	mode	p-value	strain	reference
STANDRAD	lantibiotic	SVAGGGRIDTCPAGGGTSEQTGTCC	EA	$3 \cdot 10^{-47}$	*P. marinus*	prochlorosin[40]
STANDRAD	lantibiotic	LEAASGGGDTGIQAVLHTAGCYGGTKMCRA	EA	$1 \cdot 10^{-40}$	*P. marinus*	Prochlorosin[40]
STANDRAD	lantibiotic	GAAGGCCITGESPGSAPTNDYKCTKGRGPGGCY	EA	$1 \cdot 10^{-30}$	*P. marinus*	prochlorosin[40]
STANDRAD	lantibiotic	GVAGGGGGCDGIRITDKQTVADNTIVPCSCFHQ	EA	$1 \cdot 10^{-26}$	*P. marinus*	prochlorosin[40]
STANDRAD	lantibiotic	VSPQSTIVCVSLRICNWSLRFCPSFKVRCPM	E	$3 \cdot 10^{-24}$	*G. thermodenitrificans*	geobacillin[41]
STANDRAD	glycocin	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR	EB	$1 \cdot 10^{-17}$	*B. subtilis*	sublancin[42]
STANDRAD	lantibiotic	TTWPCATVGVSVALCPTTKCTSQC	EB	$2 \cdot 10^{-17}$	*B. halodurans*	haloduracin[43]
STANDRAD	lantibiotic	KGGSGVIHTISHECNMNSWQFVFTCCS	EB	$1 \cdot 10^{-14}$	*L. lactis*	lacticin 481[44]
STANDRAD	lantibiotic	AVEQRATPATPATPWLIKASYVVSGAGVSFVASYITVN	E	$6 \cdot 10^{-21}$	*B. cereus*	bicereucin alpha[45]
STANDRAD	lantibiotic	AVEQRATPTLATPLTPHTPYATYVVSGGVVSAISGIFSNNKTCLG	EA	$8 \cdot 10^{-14}$	*B. cereus*	bicereucin beta[45]
STANDRAD	lantibiotic	NITGAGSTIQCVNTTIGTILSVVFDCCPTSACTPPCRF	E	$5 \cdot 10^{-55}$	*R. flavefaciens*	flavecin A2.d[16]
STANDRAD	lantibiotic	NMAGAGSPLTVTITGLIVAATTGFDWCPTGACTYSCRV	E	$9 \cdot 10^{-44}$	*R. flavefaciens*	flavecin A2.b[16]
STANDRAD	lantibiotic	TTGGASTVNTVGIHTTYLISKGLQNCPLKPTTILPILPRK	E	$2 \cdot 10^{-39}$	*R. flavefaciens*	flvecin A2.a[16]
STANDRAD	lantibiotic	TTVGAASTLPCAEVVVTVTGIIVKATTGFDWCPTGACTHSCRF	E	$2 \cdot 10^{-27}$	*R. flavefaciens*	flavecin A2.g[16]
STANDRAD	lantibiotic	TTVGAGSSNDCADLILKITGVVVSATSKFDWCPTGACTTSCRF	E	$3 \cdot 10^{-25}$	*R. flavefaciens*	flavecin A2.c[16]
STANDRAD	lantibiotic	KQTIVCTIAQGTVGCLVSYGLGNGGYCCTYTVECSKTCNK	EB	$8 \cdot 10^{-25}$	*R. flavefaciens*	flavecin A1[16]
STANDRAD	lantibiotic	MFDDSVVGAVGYTTYWGILPLVTKNPQICPVSENTVKCRLL	E	$1 \cdot 10^{-17}$	*R. flavefaciens*	flavecin A2.f[16]
STANDRAD	lantibiotic	SNVIGGTSSIDCVRLASNTPEGTVNLTVRIEFCPSAACTYSCRL	E	$4 \cdot 10^{-17}$	*R. flavefaciens*	flavecin A2.h[16]
ACTI	linaridin	ATPAVAQFVIQGSTICLVC	EB	$3 \cdot 10^{-36}$	*S. griseus*	grisemycin[25,46]
ACTI	lantibiotic	ASTVSLLSCISAASVLLCL(-64Da)	EA	$2 \cdot 10^{-18}$	*S. viridochromogenes*	informatipeptin[12]
ACTI	lantibiotic	**DTGGCSGLCTVLVCTVIVC**	EA	$6 \cdot 10^{-13}$	*Streptomyces* sp.	**Compound X**
ACTI	lassopeptide	**LLGRHGNDRLILSKN**(-59Da)	E	$5 \cdot 10^{-31}$	*S. viridochromogenes*	**Compound Y**
ACTI	lantibiotic	TGSQVSLLVCEYSSLSVVLCTP	EAB	$4 \cdot 10^{-12}$	*S. griseus*	AmfS[47]
ACTI	lantibiotic	TVTVCSPTGTLCGSCSMGTRGCC	EA	$9 \cdot 10^{-12}$	*S. roseosporous*	SRO-3108[25]
ACTI	lantibiotic	**GGGASTVSLLSCVSAGSVILCV**	EA	$5 \cdot 10^{-12}$	*S. cattlya*	**informatipeptin B**
ACTI	lantibiotic	TGSRASLLLCGDSSLSITTCN	EAB	$6 \cdot 10^{-11}$	S. coelicolor	SapB[48]
BACIL	lantibiotic	TTPATTSSWTCITAGVTVSASLCPTTKCTSRC	EB	$2 \cdot 10^{-14}$	*B. licheniforis ES-221*	lichenicidin B49
SPACE	lantibiotic	**DATITTVTVTSTSIWASTVSNYC**(-87Da)	E	$3 \cdot 10^{-25}$	*Bacillus* sp. ISSFR-3F	**Compound Z**
CYANO	cyanobactin	Cyclic(**ISNGYLIP**)	E	$2 \cdot 10^{-17}$	PNG22APR06-1	**Cyanobactin X**
CYANO	cyanobactin	Cyclic(**ISAPPGVTFSFP**)	E	$2 \cdot 9^{-22}$	PNG19MAY05-2/7	Wewakazole[52]
SPONGE	proteusin	TGIGVVVAVVAGAVANTGAGVNQVAGGNINVVGNINVNANVSVNMNQTT	E	$1 \cdot 10^{-20}$	*T. swinhoei* symbiont	polytheonamide[17,53]

Figure 1. The MetaRiPPQuest pipeline includes the following steps: (a) selecting a biological sample (an isolated microbe or a bacterial/fungal community), (b) generating DNA sequence data, (c) assembly of DNA sequence data with SPAdes[26] or metaSPAdes[27], (d) identifying putative RiPP precursor peptides using antiSMASH[28] and constructing a database of putative RiPPs structures (as well as a decoy database), (e) generating tandem mass spectra, (f) matching spectra against the constructed RiPP structure database using Dereplicator[29], and (g) enlarging the set of described RiPPs via spectral networking[30,31].

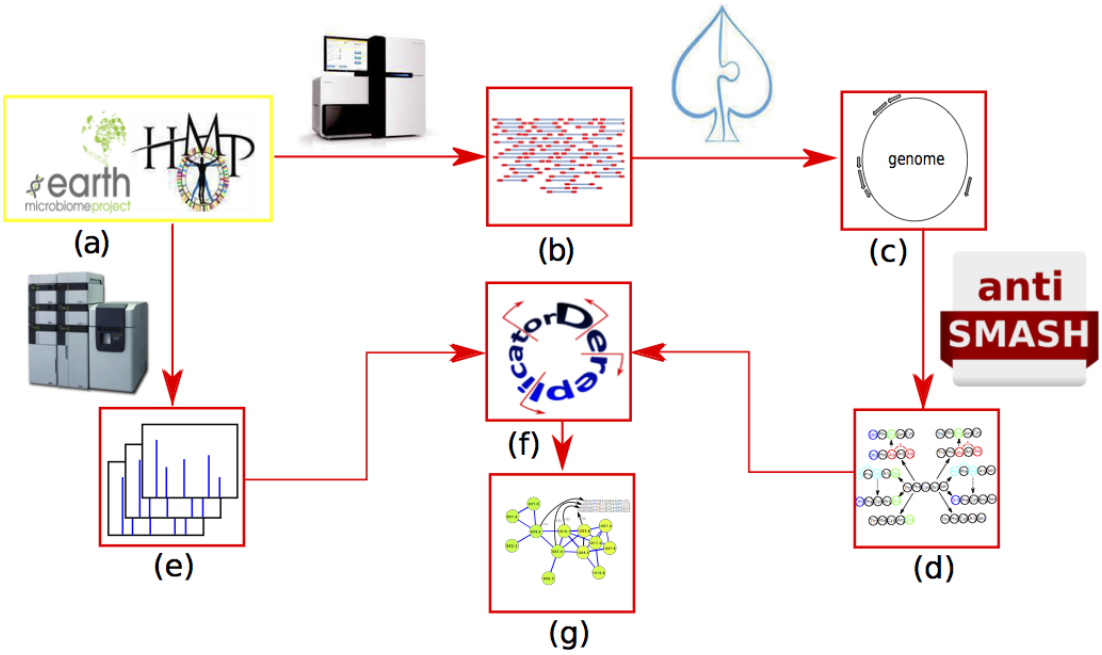


Figure 2. The histograms of p-values of PSMs/Peptides identified by MetaRiPPquest for the target and decoy databases in the search of ACTI dataset in the all-ORF (a,b) and motif-ORF (c,d) modes.

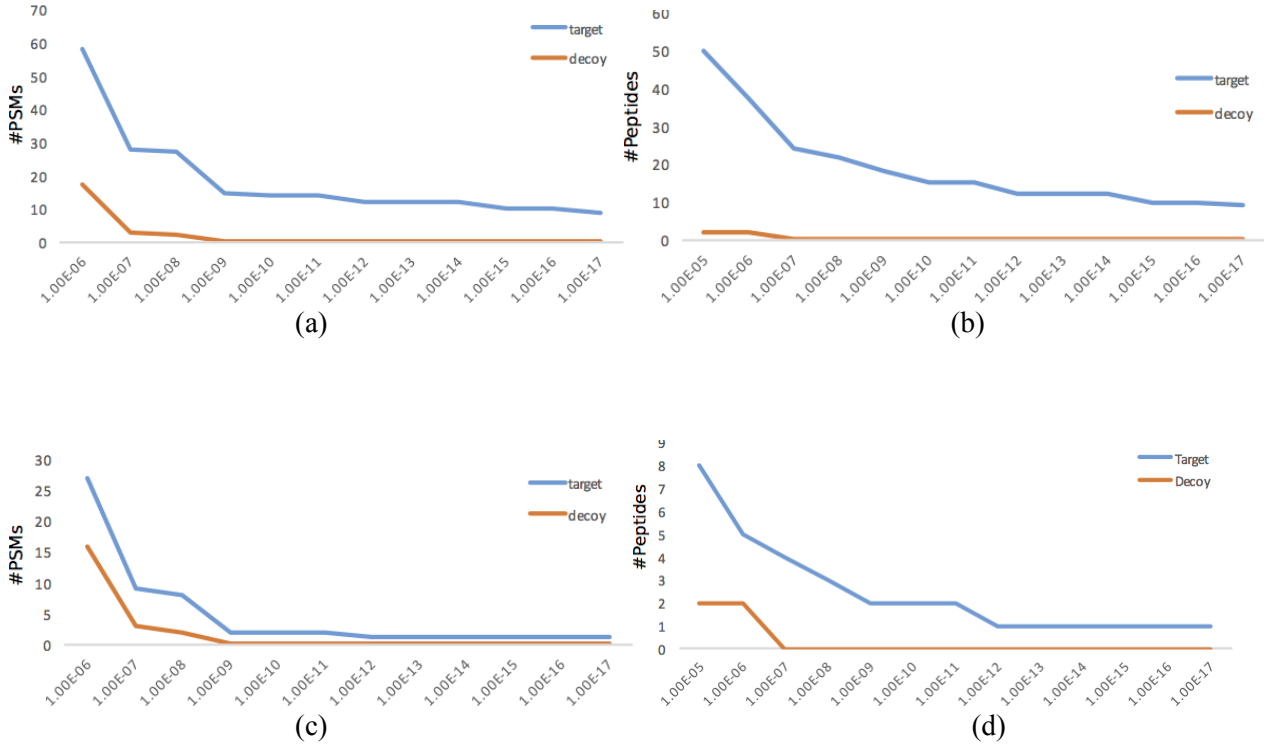


Figure 3. Identification of informatipeptin B. (a) Biosynthetic gene cluster of informatipeptin B from *S. cattleya* and the precursor peptide correctly predicted by antiSMASH. (b) The BGC of informatipeptin B has all essential class III lanthipeptide genes, a lanM-like enzyme, a transporter, and a regulator. (c) Spectral network revealed a plethora of compounds similar to informatipeptin B. Masses shown here are in charge +2 state. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) Tandem mass spectrum of informatipeptin B (score 9, p-value $1 \cdot 10^{-12}$).

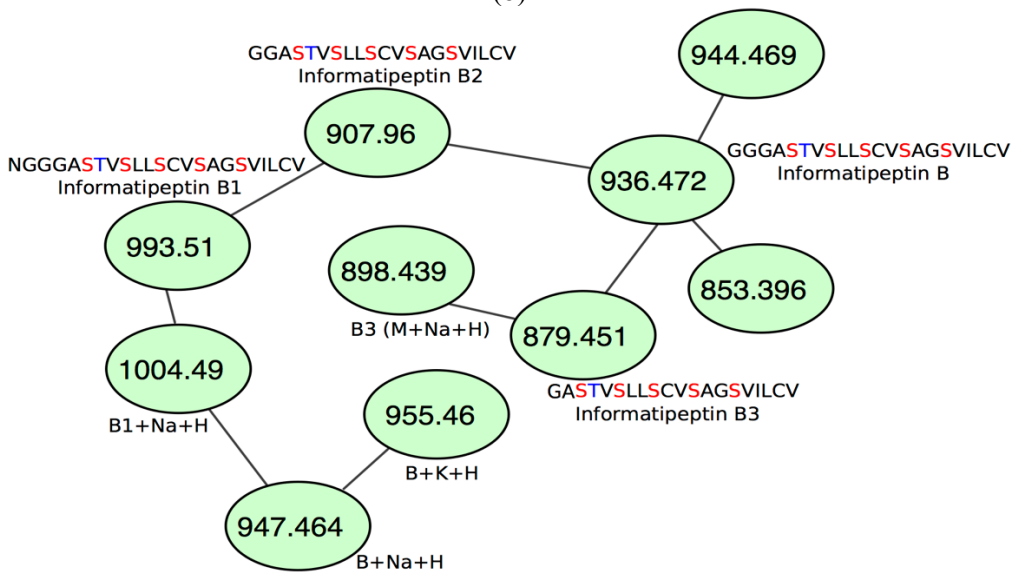
ctg1_orf05561 leader / core peptide, putative Class III

MALLDLQNMESEELNGGGAS - DhbVDhaLLDhaCVDhaAGDhaVILCV

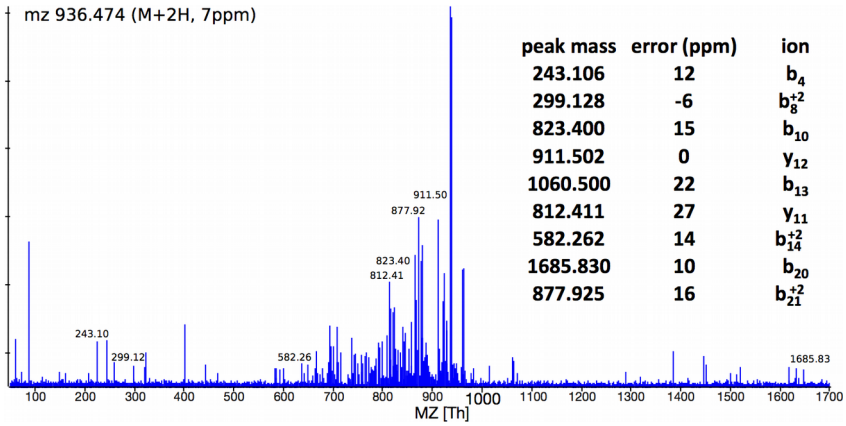
(a)

Gene	Predicted function	Length
orf05554	glycosyl transferase family 1	468aa
orf05556	TetR family transcriptional regulator	208aa
orf05558	helix-turn-helix transcriptional regulator	239aa
orf05559	ABC transporter ATP-binding protein	591aa
orf05560	ABC transporter, AmfB	737aa
orf05561	AmfS protein	37aa
orf05563	serine/threonine protein kinase	882aa
orf05565	aldehyde dehydrogenase	488aa
orf05567	metal dependent hydrolase	351aa
orf05568	hypothetical protein	410aa

(b)

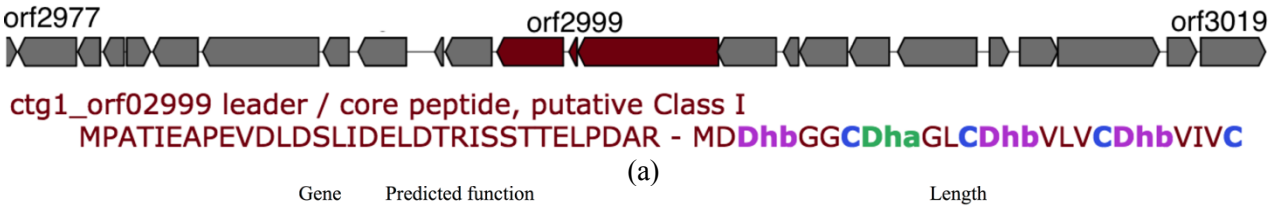


(c)



(d)

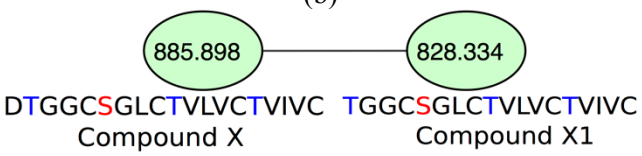
Figure 4. Identification of Compound X. (a) Biosynthetic gene cluster of Compound X from *Streptomyces* sp. and the precursor peptide as correctly predicted by antiSMASH. (b) The gene cluster of Compound X has several genes with *ca.* 60% similarity to genes from BGCs encoding subtilin and other class I lanthipeptide, suggesting that Compound X is a class I lanthipeptide. (c) Spectral networking revealed a analog of Compound X, called Compound X1, lacking the N-terminal aspartic acid residue. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) Tandem mass spectrum of Compound X (score 8, p-value $6.0 \bullet 10^{-13}$).



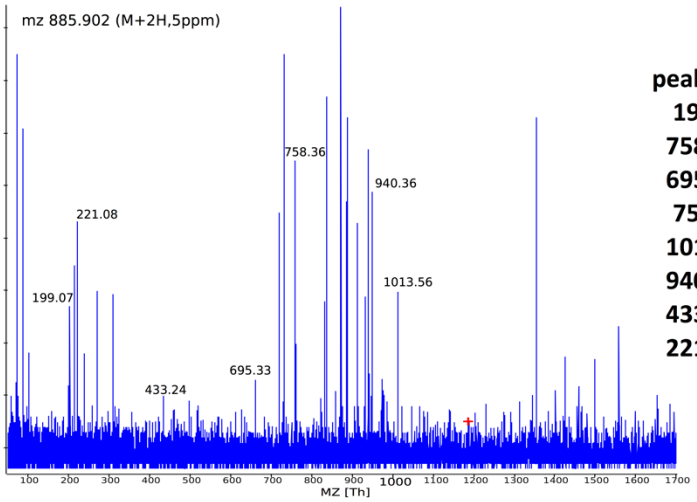
(a)

Gene	Predicted function	Length
orf02991	TetR family transcriptional regulator	175aa
orf02994	Select seq ref|WP_033173212.1| NADPH:quinone reductase	327aa
orf02996	hypothetical protein TGGT1_239580	57aa
orf02997	urea transporter	314aa
orf02998	subtilin biosynthesis protein spaC	450aa
orf02999	hypothetical protein K530_51195	53aa
orf03000	lantibiotic dehydratase	954aa
orf03002	O-methyltransferase clustered with LanBC	400aa
orf03003	NAD(P)H dehydrogenase	102aa
orf03004	oxetanocin A resistance protein	324aa
orf03007	translation initiation factor IF-2	269aa
orf03009	transcriptional regulator, PucR family protein	536aa
orf03010	MerR family transcriptional regulator	131aa

(b)



(c)



peak mass	error (ppm)	ion
199.07	7	b_2
758.365	2	y_{16}^{+2}
695.332	14	y_{14}^{+2}
758.25	12	b_9
1013.56	3	y_{10}
940.361	4	b_{11}
433.244	10	y_4
221.087	36	y_2

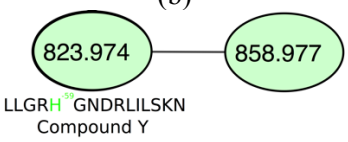
(d)

Figure 5. Identification of Compound Y. (a) Biosynthetic gene cluster of Compound Y from *S. viridochromogenes*. AntiSMASH failed to predict any precursor peptide for this gene cluster. (b) The biosynthetic gene cluster of Compound Y has several genes with similarity to the biosynthetic gene cluster of lassopeptide microcin J25. (c) Spectral network of Compound Y (d) Tandem mass spectrum of Compound Y (score 18, p-value $5 \cdot 10^{-31}$). MetaRiPPquest discovered Compound Y in blind modification search mode, and assigned a modification of -59Da to the histidine residue, shown in green. While *m/z* 858.97 clusters with Compound Y in the spectral network, MetaRiPPquest does not identify it with a low p-value.

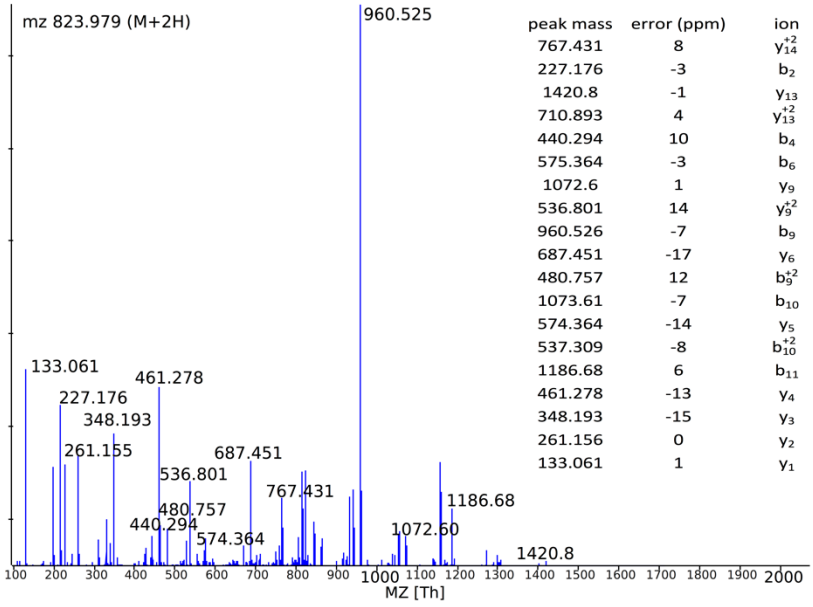
(a)

Gene	Predicted function	Length
orf01529	Transcriptional regulator	253aa
orf01531	Endo-1,4-beta-xylanase	1620aa
orf01533	Dihydrolipoyl dehydrogenase	470aa
orf01534	Putative microcin mature peptide	36aa
orf01536	Asparagine Synthetase B/C	617aa
orf01537	PQQ synthetase	84aa
orf01538	Microcin J25 C	145aa
orf01539	Lipid A export permease/ATP-binding protein	613aa
orf01540	Calcium-binding protein	205aa
orf01543	2-keto-3-deoxy-gluconate kinase	331aa
orf01544	adolase	260aa
orf01546	3-oxoacyl-[acyl-carrier-protein] reductase	170aa
orf01547	TetR family transcriptional regulator	218aa
orf01550	Cytochrome P450 monooxygenase	437aa
orf01551	Dihydroxy-acid dehydratase	571aa
orf01552	Regulatory protein	275aa
orf01553	Alcohol dehydrogenase	336aa

(b)



(c)



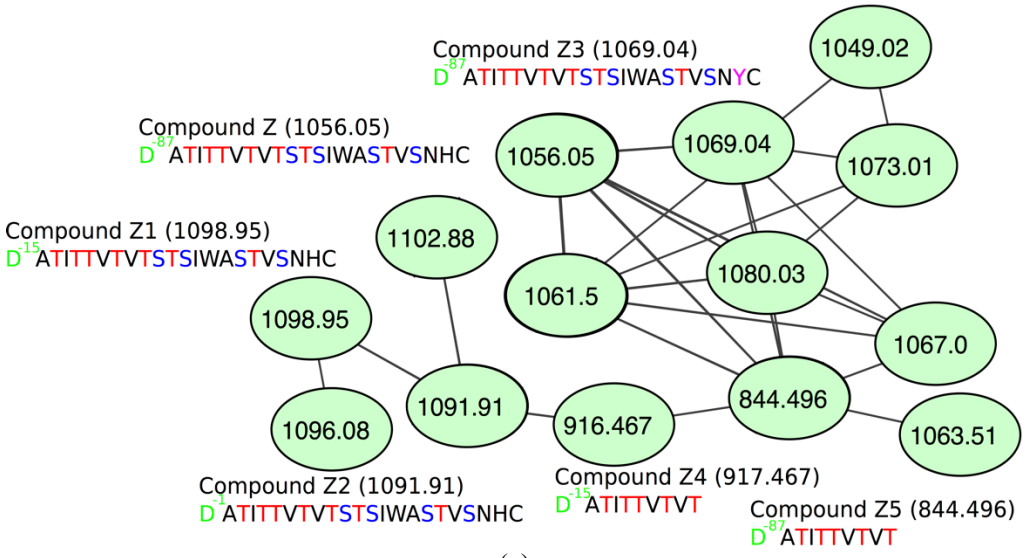
(d)

Figure 6. Identification of Compound Z. (a) Biosynthetic gene cluster of Compound Z in *Bacillus* sp. ISSFR-3F. AntiSMASH failed to predict any precursor peptide for this gene cluster. The precursor gene is repeated fourteen times, with thirteen being exactly the same and one differing from the rest by a His to Tyr replacement (b) The BGC of Compound Z (c) Spectral networking revealed analogs of Compound Z with various N-terminal and C-terminal modifications. Dehydrated serines are shown in red and dehydrated threonines are shown in blue (d) The tandem mass spectrum of Compound Z (score 16, p-value $3 \cdot 10^{-25}$). MetaRiPPquest discovered compound Z in blind modification search mode, and assigned -87Da, -15Da, and -1Da modifications to the aspartic acid residue, shown in green residue. Compound Z4 and Z5 are discovered in charge +1, while other compounds are charge +2.

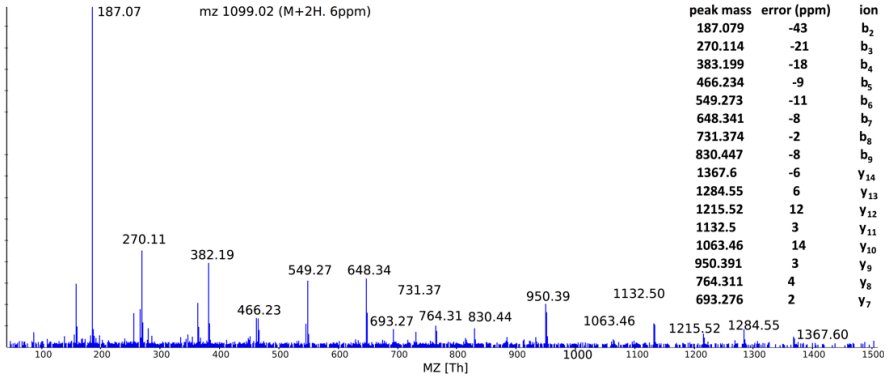
(a)

Gene	Predicted function	Length
orf05700	transcriptional regulator	117aa
orf05702	enterotoxin	421aa
orf05707	multidrug ABC transporter permease	555aa
orf05708	Putative SAM-dependent methyltransferase	277aa
orf05723	serine/threonine protein kinase	872aa
orf05724	ABC-2 transporter family protein	197aa
orf05725	multidrug ABC transporter ATP-binding protein	284aa
orf05726	peptidase M16	427aa
orf05729	multidrug ABC transporter permease	201aa
orf05730	sodium ABC transporter ATP-binding protein	290aa

(b)



(c)

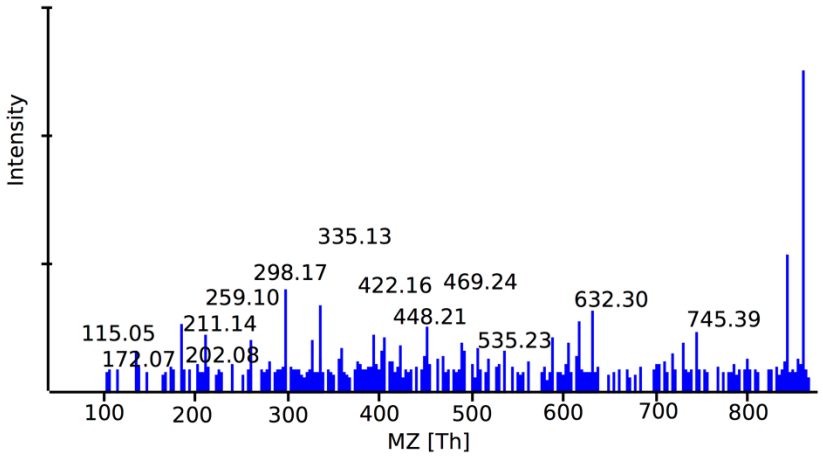


(d)

Figure 7. Identification of Cyanobactin X. (a) The biosynthetic gene cluster of cyanobactin X in the Cyanobacterium PNG22APR06-1. AntiSMASH failed to predict this gene cluster. (b) Annotation of the tandem mass spectrum of Cyanobactin X (score 13, p-value $2 \cdot 10^{-17}$).

Predicted function	Length
glyoxalase	354 aa
Aspartyl/Asparaginyl hydroxylase	471 aa
GMC oxydoreductase	870 aa
GMC oxidoreductase C	345 aa
AMp Binding	1335 aa
Glycosyltransferase	675aa
Succinylglutamate desuccinylase/Aspartoacylase	849aa

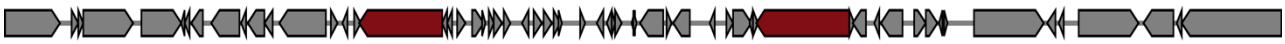
(a)



peak mass	error (ppm)	ion
745.393	-7	SNGYLIP
211.143	8	PI
298.175	3	PIS
535.236	28	SNGYL
448.213	16	NGYL
422.168	-3	SNGY
469.24	2	PISNG
335.134	2	NGY
259.103	4	SNG
632.299	8	PISNGY
202.086	-18	SN
172.077	-32	NG
115.052	-16	N

(b)

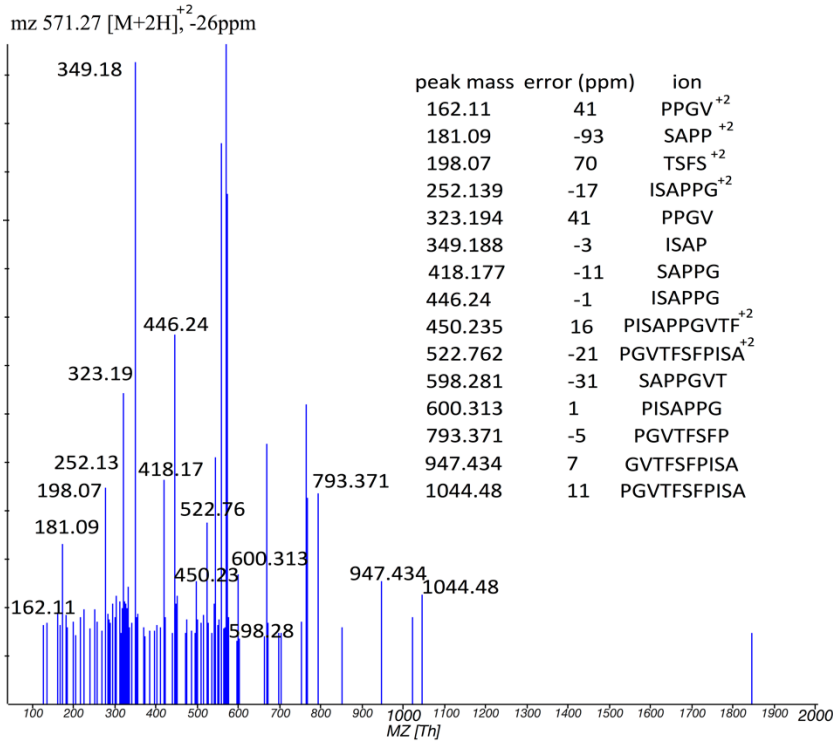
Figure 8: Identification of wewakazole. (a) Biosynthetic gene cluster of wewakazole in the Cyanobacterium PNG19MAY05-2/7. AntiSMASH failed to predict the precursor peptide for this gene cluster by motif search; however, MetaRiPPquest was able to discover this peptide in all-ORF mode. (b) Annotation of the BGC of wewakazole, (c) Annotation of the tandem mass spectrum of wewakazole (score 15, p-value $2 \cdot 10^{-22}$).



(a)

Gene	Predicted function	Length
orf11591	bacterial pre-peptidase	493aa
orf11597	fructosamine kinase family protein	357aa
orf11599	peptidase M16	438aa
orf11603	sporulation protein SpoOM	256aa
orf11605	PIN domain nuclease	147aa
orf11609	methylglyoxal synthase	424aa
orf11615	Peptidase S8	737aa
orf11648	Uma2 family endonuclease	214aa
orf11652	XisI protein	138aa
orf11657	DNA-binding protein	150aa
orf11660	adenylate cyclase	826aa
orf11668	hypothetical protein	201aa
orf11675	alpha-amlyase	624aa
orf11679	dynamin	544aa
orf11680	methyltransferase	277aa
orf11682	ABC-type branched-chain amino acid transport system	890aa

(b)



(c)