Title

MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities

Journal

Cell Systems, 9(6)

ISSN

2405-4712

Authors

Cao, Liu
Gurevich, Alexey
Alexander, Kelsey L
et al.

Publication Date

2019-12-01

DOI

10.1016/j.cels.2019.09.004

Peer reviewed

HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2020 December 18.

Published in final edited form as:

Cell Syst. 2019 December 18; 9(6): 600–608.e4. doi:10.1016/j.cels.2019.09.004.

MetaMiner: A Scalable Peptidogenomics Approach for Discovery of Ribosomal Peptide Natural Products with Blind Modifications from Microbial Communities

Liu Cao[1], **Alexey Gurevich**[2], **Kelsey L. Alexander**[3,4], **C. Benjamin Naman**[3,5], **Tiago Leão**[3], **Evgenia Glukhov**[3], **Tal Luzzatto-Knaan**[6], **Fernando Vargas**[6], **Robby Quinn**[6], **Amina Bouslimani**[6], **Louis Felix Nothias**[6], **Nitin K. Singh**[7], **Jon G. Sanders**[8], **Rodolfo A. S. Benitez**[8], **Luke R. Thompson**[9,10], **Md-Nafiz Hamid**[11,12], **James T. Morton**[8,13], **Alla Mikheenko**[2], **Alexander Shlemov**[2], **Anton Korobeynikov**[2,14], **Iddo Friedberg**[11,12], **Rob Knight**[8,13,15,16], **Kasthuri Venkateswaran**[7], **William H. Gerwick**[3], **Lena Gerwick**[3], **Pieter C. Dorrestein**[6,15], **Pavel A. Pevzner**[13,15], **Hosein Mohimani**[1,13,17,*]

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[2]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia

[3]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

[4]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, California, USA

[5]Li Dak Sum Yip Yio Chin Kenneth Li Marine Biopharmaceutical Research Center, Department of Marine Pharmacy, College of Food and Pharmaceutical Sciences, Ningbo University, Ningbo, Zhejiang, China

[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, Pasadetana, California, USA

[8]Department of Pediatrics, University of California San Diego, School of Medicine, La Jolla, California, USA

[9]Department of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, Mississippi, USA

[*]Corresponding author, hoseinm@andrew.cmu.edu.

Author Contributions

H.M., P.A.P., P.C.D., L.G., W.H.G., K.V., R.K., I.F. designed and directed the study. H.M., R.A.P. and L.C. wrote the paper. L.C., A.G., A.M., A.S., A.K., M.H. and J.T.M. wrote the codes. K.L.A and C.B.N. conducted NMR experiment. T.L., E.G., T.L.K., F.V., R.Q., A.B., L.F.N., N.K.S., J.G.S., R.A.S.B. and L.R.T. collected samples, LC-MS/MS and metagenomics data.

[10]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, La Jolla, California, USA

[11]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, Iowa, USA

[12]Interdepartmental program in Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA

[13]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

[14]Department of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia

[15]Center for Microbiome Innovation, Jacobs School of Engineering, University of California, San Diego, La Jolla, California, USA.

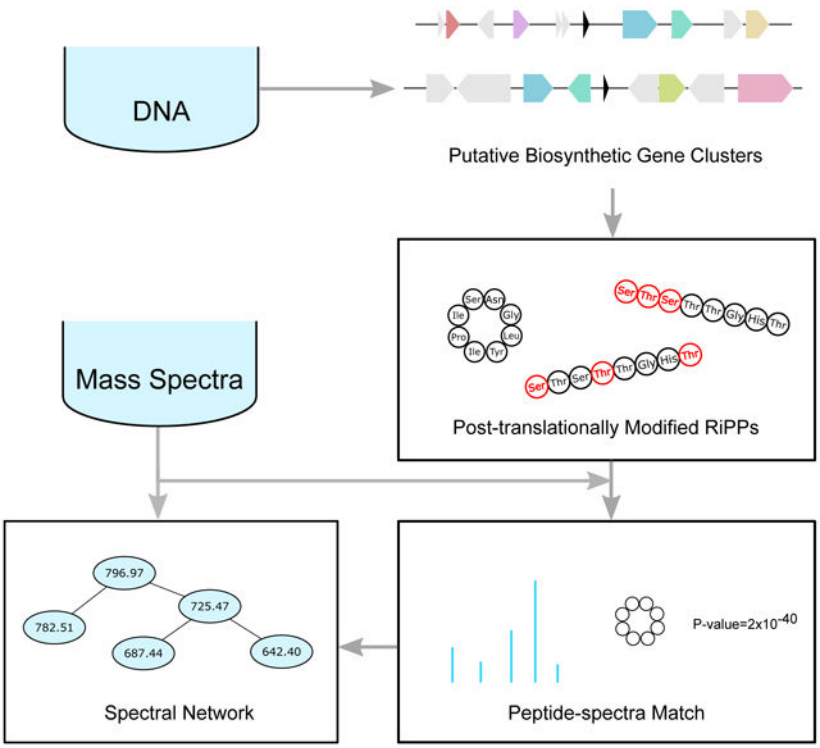
[16]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA

[17]Lead Contact

Summary

Ribosomally synthesized and post-translationally modified peptides (RiPPs) are an important class of natural products that contain antibiotics and a variety of other bioactive compounds. The existing methods for discovery of RiPPs by combining genome mining and computational mass spectrometry are limited to discovering specific classes of RiPPs from small datasets, and they fail to handle unknown post-translational modifications. We present MetaMiner, a software tool for addressing these challenges 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 eight genomic and metagenomic datasets, MetaMiner discovered 31 known and seven unknown RiPPs from diverse microbial communities including human microbiome, lichen microbiome, and micro-organisms isolated from the International Space Station.

Graphical Abstract



eTOC Blurb

We have developed an efficient method for discovering ribosomally synthesized and post-translationally modified peptides with unknown post-translational modifications from microbial communities through the database search of mass spectra against peptides predicted by genome/metagenome mining. Searching eight datasets collected from various environments including human microbiome, marine cyanobacteria and international space station resulted in the discovery of seven unknown RiPPs.

Keywords

Ribosomally synthesized and post-translationally modified peptides; computational mass spectrometry; metagenomics; natural products discovery; microbial metabolites

Introduction

Natural products are back at the center of attention as pharmaceutical leads, as exemplified by the recent discoveries of bioactive natural product drugs (Fischbach and Walsh, 2009; Harvey et al., 2015; Li and Vederas, 2009; Ling et al., 2015). Recent advances in metagenomics are transforming the field of natural product discovery by enabling the recovery of microbial genomes directly from the environmental samples. This has revolutionized our understanding about the microbial composition of various communities and their biosynthetic gene clusters. Biosynthetic gene cluster are sets of genes that synthesize microbial small molecules from simple building blocks. The metagenomes of microbial communities contain thousands of biosynthetic gene clusters with unknown small

molecule products, making them an untapped resource for the future antimicrobial drug discovery (Charlop-Powers et al., 2016; Donia et al., 2014; Hadjithomas et al., 2015).

The biosynthetic gene clusters of natural products in microbial genomes can be identified by sequence similarity searches (Li et al., 2009). Moreover, in the case of peptide natural products (PNPs), it is possible to predict the corresponding putative structures based on the genes present in their biosynthetic gene clusters (Oman and van der Donk, 2010; Stachelhaus et al., 1999). However, the structure of PNPs usually differs from these predictions due to post-translational modifications applied by the enzymes in the biosynthetic gene cluster. For example, in the case of polytheonamides, the PNP has 49 residues, with 21 post-translational modifications, making it nearly impossible to predict the PNP structure solely based on the genomic data. Therefore, in addition to genome mining, metabolomics methods including mass spectrometry and nuclear magnetic resonance are routinely used for determining the structure of the molecular product of biosynthetic gene cluster (Doroghazi et al., 2014; Medema et al., 2014; Mohimani et al., 2014a; Mohimani et al., 2014b).

The recent launch of the Global Natural Products Social (GNPS) molecular networking infrastructure (Wang et al., 2016) brought together over a thousand laboratories worldwide that have already generated an unprecedented amount of tandem mass spectra of natural products. Computational mass spectrometry methods have revealed thousands of known small molecules and their unknown variants in various microbiome datasets from the GNPS molecular networking (Gurevich et al., 2018; Mohimani et al., 2017; Mohimani et al., 2018). However, the majority of the spectra in these datasets remains unannotated, which is often referred as the ‘dark matter of metabolomics’. A portion of these spectra represent unknown small molecule products of biosynthetic gene clusters encoded in the microbial genomes, and computational algorithms are needed for illuminating this dark matter (Donia and Fischbach, 2015; Medema and Fischbach, 2015; Mohimani and Pevzner, 2016; Vaniya and Fiehn, 2015; Walsh, 2015).

This paper focuses on the integration of computational mass spectrometry and genome mining for discovering unknown Ribosomally synthesized and Post-translationally modified Peptides (RiPPs). RiPPs are a rapidly expanding group of natural products with applications in pharmaceutical and food industries (Arnison et al., 2013). RiPPs are produced through the Post Ribosomal Peptide Synthesis (PRPS) pathway (Arnison et al., 2013). Initially, RiPPs are synthesized as precursor peptides encoded by RiPP structural genes. The RiPP structural genes are often short, making their annotations difficult (Mohimani et al., 2014a). The precursor peptide consists of a prefix leader peptide appended to a suffix core peptide. The leader peptide is important for the 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 its post-translational modifications (PTMs), proteolytic cleavage, and transportation are usually located within a contiguous biosynthetic gene cluster of the RiPP. The length of a microbial RiPP-encoding biosynthetic gene cluster typically varies from 1,000 to 40,000 bp (average length 10,000 bp). Since RiPP-encoding

biosynthetic gene clusters are much longer than the current length of short reads generated by next generation sequencing, DNA assembly is a critical part of any RiPP discovery approach based on short reads.

Genome mining refers to the interpretation of a natural product biosynthetic gene cluster to infer information about the natural product itself. The discoveries of coelichelin in *Streptomyces coelicolor* (Challis and Ravel, 2000; Lautru et al., 2005) and orfamide in *Pseudomonas fluorescens* Pf-5 (Gross et al., 2007; Paulsen et al., 2005) were the first examples of genome mining that were followed by discoveries of various bioactive RiPPs in microbial samples. Donia et al. discovered lactocillin, a thiopeptide antibiotic from the human vaginal isolates that showed activity against vaginal pathogens (Donia et al., 2014). Zhao et al. discovered eight novel lanthipeptides with antibiotic activity from a ruminant bacterium (Zhao and van der Donk, 2016). Freeman et al. and Wilson et al. used metagenome mining of a sponge to assign a biosynthetic gene cluster to the known RiPP polytheonamide, with post-translational modifications distributed across 49 residues (Freeman et al., 2012; Wilson et al., 2014). Thus, large-scale metagenomics projects, such as Earth Microbiome Project (Gilbert et al., 2014; Thompson et al., 2017), American Gut Project (McDonald et al., 2018), and Human Microbiome Project (Human Microbiome Project, 2012a, b; Lloyd-Price et al., 2017), 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 analysis that required the isolation of microbes, and the purification of microbial metabolites. Our goal is to discover the RiPPs directly from the mass spectrometry and metagenomics 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 biosynthetic gene clusters, including 20,000 RiPP-encoding biosynthetic gene clusters in the Integrated Microbial Genome Atlas of biosynthetic Gene Clusters (IMG-ABC), connecting these biosynthetic gene clusters to their metabolites has not kept pace with the speed of microbial genome sequencing (Hadjithomas et al., 2015). Currently, only 35 out of these roughly 20,000 RiPP-encoding biosynthetic gene clusters in IMG-ABC have been experimentally connected to their RiPPs (Hadjithomas et al., 2015; Medema et al., 2015). Linking RiPP-encoding biosynthetic gene clusters to unknown RiPPs requires the development of computational tools.

Kersten et al. introduced the peptidogenomics approach to RiPP discovery, which refers to finding sequential amino acid tags from the tandem mass spectra (peptidomics) and mining them in the assembled DNA reads obtained from the same sample (Kersten et al., 2011). Mohimani et al. introduced RiPPquest, the first automated approach to RiPP discovery by combining mass spectrometry and genome mining (Mohimani et al., 2014a). This tool is based on *Peptide-Spectrum Matches*, which are generated by aligning predicted spectra of putative RiPPs annotated by genome mining. If a *peptide-spectrum match* between a candidate RiPP predicted from the assembled genome and a spectrum is statistically significant, then RiPPquest reports it as a putative annotation of the spectrum. RiPPquest resulted in the 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) it is limited to lanthipeptides which constitutes only one of 19 classes of RiPPs (Arnison et al., 2013), (b) it 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) it does not report the statistical significance of the identified RiPPs, a key requirement for any high-throughput peptide identification tool, and (d) it is limited to searches for a predefined set of post-translational modification (PTMs) and does not enable blind searches for unknown PTMs. Since RiPPquest, other tools have been developed that combine genomics with mass spectrometry based discovery (Medema et al., 2014; Skinnider et al., 2015). However, these tools are limited to the analysis of a single or few spectra from isolated genomes and cannot scale to search billions of spectra from GNPS against metagenomics datasets.

This paper describes MetaMiner, a tool that enables searching mass spectrometry databases against metagenomics short reads sequenced from microbiome samples for rapid discovery of RiPPs. Application of MetaMiner to mass spectrometry and metagenomics data from the human microbiome resulted in the identification of known and unknown peptides from the human microbiota, including autoinducer peptide (AIP), Mec-PSM, delta-toxin, and their unknown variants.

Results

Brief description of MetaMiner.

MetaMiner pipeline (Figure 1) analyzes the paired genome/metagenome assemblies and tandem mass spectra from isolated microbes or bacterial/fungal communities. Starting from the genome assemblies, MetaMiner (a) identifies putative biosynthetic gene clusters and the corresponding precursor peptides, (b) constructs target and decoy putative RiPP structure databases (c) matches tandem mass spectra against the constructed RiPP structure databases using Dereplicator, and (d) enlarges the set of described RiPPs via spectral networking (Bandeira et al., 2007; Watrous et al., 2012).

Advances of MetaMiner.

In step (a), MetaMiner searches for diverse classes of RiPPs, including lanthipeptides, lassopeptides, linear azole containing peptides (LAPs), linaridins, glycocins, cyanobactins, phenol-soluble modulins, AIP, and proteusins (versus only lanthipeptides by RiPPquest). Moreover, MetaMiner supports searching for user-defined classes of RiPPs. In step (b)-(c), MetaMiner implement an approach to estimate false discovery rate through target decoy analysis by searching mass spectra against decoy RiPP structures generated by random shuffling. In step (c), MetaMiner uses an efficient algorithm for searching sparse vectors corresponding to mass spectra and RiPP structures, increasing the speed by two orders of magnitude compared to RiPPquest, thus enabling searches of the entire GNPS databases against metagenomes. Unusual modifications are handled through blind post-translational modification searching. In addition, in contrast to RiPPquest (which was designed for analyzing low-resolution spectra), MetaMiner enables searching high-resolution mass spectra, and allows user-adjustable precursor and product ion thresholds.

Genome mining.

MetaMiner uses antiSMASH, and BOA for the identification of RiPP-encoding biosynthetic gene clusters and has two genome mining modes for selecting Open Reading Frames (ORFs), a slow all-ORF mode introduced in RiPPquest (Mohimani et al., 2014a), and a fast motif-ORF mode. The all-ORF approach analyzes all short ORFs within a biosynthetic gene cluster, while the motif-ORF approach relies on RiPP motif finding (Blin et al., 2014) 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 (see STAR methods for details of all the datasets). AntiSMASH found 30 biosynthetic gene clusters in this genome, including six RiPP-encoding biosynthetic gene clusters. Within these six biosynthetic gene clusters, 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 the 36 strains from the ACTI strains, antiSMASH discovered 1,140 biosynthetic gene clusters, including 168 RiPP-encoding biosynthetic gene clusters. MetaMiner 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 a RiPP, rather than by the RiPP sequence itself. Therefore, BOA has a capability to identify 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 most 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 peptide-spectrum matches 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 MetaMiner only reports statistically significant peptide-spectrum matches, the all-ORF mode may miss some peptides identified in the motif-ORF mode. Conversely, because MetaMiner 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.

Figure S1 shows a comparison of the performance of MetaMiner with all-ORF and motif-ORF genome mining approaches on the ACTI dataset. At the extremely conservative 0% FDR, MetaMiner in the motif-ORF mode identified three unknown RiPPs and five known RiPPs. MetaMiner in the all-ORF mode identified only two known RiPPs at 0% FDR. 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.

RiPP Discovery.

MetaMiner identified 31 known RiPPs and discovered seven unknown RiPPs in various datasets, including Actinomyces strains, Bacillus strains, Cyanobacteria strains, sponge microbiome, microbial isolates from the International Space Station, and human microbiome (see STAR methods for details). Table S1 provides information about all the RiPPs identified at 1% FDR by MetaMiner. Among the 31 known RiPPs, 23 were identified in the strains identical to the previous reports, three were identified in strains with 99% or higher 16S rRNA similarity, two were identified in the same species, one was identified in the same genus, and two were identified in the same samples (Table S2, Figures S7-S11, S13). The seven unknown RiPPs belong to various classes, including lanthipeptides, lassopeptides, cynobactins, and phenol-soluble modulins classes (Figure 2). Their putative biosynthetic gene clusters contain all the essential genes responsible for the modifications (Figures 2, S2-S6, S12, and S14).

Confirmation of wewakazole identification.

MetaMiner identified wewakazole in a polar fraction from the extract of the strain PNG26APR06-4, a marine cyanobacterium collected at Kape Point, Papua New Guinea. Wewakazole was first reported by the co-authors of this paper (W.H.G.) from another Papua New Guinea collection of *Lyngbya majuscula* (revised to *Moorea producens*) obtained from Wewak Bay (Nogle et al., 2003). Subsequently a related compound, wewakazole B was isolated from a Red Sea collection of this cyanobacterium (Lopez et al., 2016). To validate the MetaMiner's identification of wewakazole from strain PNG26APR06-4, 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 HR-ESI-MS *(m/z* 1163.5282, Figure S15). Its chemical identity was further confirmed utilizing ^1H, HSQC and HMBC NMR data, which allowed for direct comparison with data previously reported for wewakazole (Supplementary Figure S16-S18) (Nogle et al., 2003). Moreover, the tandem mass spectrum and retention time of the isolated compound matched the data previously reported for wewakazole (Figures S19 and S20) (Nogle et al., 2003). Furthermore, the ECCD spectrum resembled that of wewakazole B (Lopez et al., 2016), and the specific rotation showed the same sign as previously reported for wewakazole (Nogle et al., 2003), excluding the possibility of an enantiomeric relationship of this isolate to that of wewakazole (Figure S21). Thus, the compound identified by MetaMiner was isolated and its identity was confirmed as wewakazole.

Discussion

While recent genome mining efforts have revealed over 20,000 hypothetical RiPP-encoding biosynthetic gene clusters (Hadjithomas et al., 2015), only 35 RiPPs matching these biosynthetic gene clusters have been identified so far. To keep pace with the speed of microbial genome sequencing, high-throughput methods for structure elucidation of RiPPs are needed that combine metagenomics, genome mining, and peptidomics. MetaMiner extends our previous RiPPquest tool (limited to lanthipeptides) to lassopeptides, LAPs,

linaridins, glycocins, cyanobactins, and proteusins, and enables the blind search for RiPPs with unusual modifications.

Studies describing RiPPs are usually limited to the analysis of a single peptide or a few related peptides. The first application of MetaMiner revealed many known RiPPs, as well as their unknown analogs, and seven novel RiPPs (three lanthipeptide, one lassopeptide, two peptide-spectrum matches and one cyclic cyanobactin) along with their numerous analogs, from only eight spectral datasets. MetaMiner identifications were validated by the isolation of the RiPP metabolite wewakezole and confirmation of its structure by orthogonal approaches, confirming that the MetaMiner prediction was correct. In contrast to the existing genome mining approaches that rely on known biosynthetic gene cluster motifs (Weber et al., 2015), MetaMiner in the all-ORF mode has the ability to discover unknown biosynthetic gene clusters (with previously unknown motifs) that encode novel RiPPs (e.g. Compound Bac-ISS-2196 and cyanobactin X) that are very different from all the currently known RiPPs and thus are not captured by the existing genome mining tools. MetaMiner can potentially make RiPP identification as robust as peptide identification in the traditional proteomics.

We further evaluated the performance of MetaMiner on eight paired mass spectral and genomics/metagenomics datasets. In a positive control dataset collected on various isolated RiPPs, MetaMiner correctly identified all the 18 known RiPPs. In a dataset collected on various *Actinomyces* strains, MetaMiner identified eight RiPPs, among which five have been previously reported in similar strains. In a dataset collected on sponge microbiome, MetaMiner successfully discovered a known compound polytheonamide previously reported in the same sample. In a dataset of *Bacillus* strains, a known RiPP lichenicidin is discovered in an unknown producer. In a dataset collected on strains from the human microbiome, MetaMiner discovered an interesting known quorum sensing autoinducer peptide from a *Staphylococcus* strain. Moreover, MetaMiner identified multiple phenol-soluble modulins, a class of secreted staphylococcal peptides that have the ability to lyse human neutrophils, the main cellular defense line against *Staphylococcus aureus* infection. The production of AIP and phenol-soluble modulins have been previously reported in related Staphylococcus strains, but this is the first time these molecules are identified in the human microbiome. MetaMiner also identified a known RiPP, wewakezole, in a cyanobacterial strain, which was confirmed by subsequent isolation and identification by nuclear magnetic resonance.

In this paper we used a target-decoy strategy to control the false discovery rate of MetaMiner. For each dataset, we reported all the RiPPs identified at 1% FDR threshold. This resulted in the identification of 38 RiPPs after analyzing 10 million spectra. While the discovery of only 38 RiPPs might look pessimistic at first glance, our spectral networking analysis shows that these 10 million spectra cluster into only 8071 families. Moreover, searching these families against known chemical structures using Dereplicator+ shows that many of these families belong to non-ribosomal peptides, polyketides, terpenes and other classes of natural products. While the discovery of 38 RiPPs by searching 10 million spectra provides a proof of concept for the MetaMiner method, the presented method is applicable to any mass spectral / genomics / metagenomics data collected on the isolated microbes/ microbial communities.

Currently, only 1% of the spectra from GNPS dataset has been searched against the genomic/metagenomic references using MetaMiner. The other 99% of the spectra in GNPS are not accompanied with the genomics/metagenomics data, making it impossible to search them. Recent genome/metagenome mining studies have revealed hundreds of thousands of biosynthetic gene clusters with uncharacterized small molecules from the publicly available genomic / metagenomic data in the National Center for Biotechnology Information (NCBI) and the Joint Genome Institute (JGI) repositories. A portion of the unannotated spectra in GNPS datasets are likely formed by the small molecule products of biosynthetic gene clusters from these publicly available genomes/metagenomes.

MetaMiner enables rapid search of billions of mass spectra from GNPS infrastructure against the metagenomics/reference genome datasets collected on the microbial communities. MetaMiner is capable of searching large metagenomics and mass spectral datasets in order to construct catalogues of unknown antimicrobial molecules that can be used as drug leads in high-throughput screening efforts.

STAR Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Hosein Mohimani (hoseinm@andrew.cmu.edu). This study did not generate new unique reagents.

METHOD DETAILS

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 MetaMiner (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 biosynthetic gene clusters in these genomes, including 29 RiPP-encoding clusters. Since the genome sequence of the lacticin producer is not available, we searched its spectrum against its described biosynthetic gene clusters (Rince et al., 1997).

Actinomycetes dataset (ACTI).: This dataset consists of 473,135 spectra from bacterial extracts of 36 *Actinomycetales* strains with sequenced genomes (Duncan et al., 2015; Mohimani et al., 2014b) (GNPS datasets MSV000078839 and MSV000078604). We downloaded sequence information for these 36 genomes from the NCBI RefSeq database. AntiSMASH identified 1,140 biosynthetic gene clusters in these genomes, including 168 RiPP-encoding clusters. 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 200 bp to 300 bp). 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 MetaMiner 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 (Nguyen et al., 2013) (MSV000078552). We downloaded genome sequence information for these isolates from the NCBI RefSeq dataset. AntiSMASH identified 12 biosynthetic gene clusters (one RiPP-encoding cluster) in *B. amyloliquefaciens* FZB42 and 11 biosynthetic gene clusters (four RiPP-encoding clusters) in *B. licheniformis* ES-221.

Space station dataset (SPACE).: This dataset consists of 58,422 spectra from bacterial extracts of 21 isolated strains collected at 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 (Singh et al., 2016; Venkateswaran et al., 2017). AntiSMASH identified 119 biosynthetic gene clusters, including 27 RiPP-encoding clusters.

Sponge dataset (SPONGE).: This dataset contains 223,135 spectra from bacterial extracts of *Theonella swinhoei* (GNPS dataset MSV000078670). Wilson et al. (Wilson et al., 2014) used the SPONGE dataset to analyze the RiPP polytheonamide. We searched spectra from the SPONGE dataset against the genome of *Theonella swinhoei* symbiont *Candidatus Entotheonella* sp. TSY1. In this dataset, AntiSMASH identified 27 biosynthetic gene clusters, including four RiPP-encoding clusters.

Cyanobacteria dataset (CYANO).: This dataset consists of 11,921,457 spectra from the extracts of 317 cyanobacterial samples (Luzzatto-Knaan et al., 2017) (GNPS dataset MSV000078568). Each sample represents a mini-metagenome (Nurk et al., 2013; Nurk et al., 2017) with one or a few highly abundant strains. The metagenomic reads were collected from 195 of these samples. AntiSMASH identified 2,898 biosynthetic gene clusters in the 195 cyanobacterial metagenomes, including 491 RiPP-encoding clusters.

Reference Human Microbiome isolates (HUMAN-iso).: This dataset contains 137,556 spectra from 17 human microbiome isolates (GNPS dataset MSV000078556). AntiSMASH identified 55 biosynthetic gene clusters in the 17 human microbiome isolate references, including 12 RiPP-encoding clusters.

Human Microbiome isolates from Cystic Fibrosis patients.: *(HUMAN-CF).* This dataset contains 7,554,646 spectra from 276 microbial isolates from the sputum culture of cystic fibrosis patients, (GNPS dataset MSV000080251). The short reads were collected on these 276 samples, and each sample contains a mixture of few (from one to eleven) bacteria. AntiSMASH identified 1,111 biosynthetic gene clusters, including 171 RiPP-encoding clusters.

MetaMiner pipeline.—MetaMiner works with paired genomic/metagenomic and tandem mass spectral data collected on isolated microbes or bacterial/fungal communities. The genomic input to MetaMiner could be short-read data, genome assemblies or extracted biosynthetic gene clusters, in fastq or fasta format. The mass spectral input to MetaMiner is in MGF or mzXML format. If the input is short read data, MetaMiner first assembles the reads into contigs. The pipeline for MetaMiner is described below:

(a) Constructing the database of putative RiPP precursor peptides. MetaMiner uses two strategies to search for precursor peptides, motif-ORF and all-ORF strategy. The motif-ORF strategy is based on the annotations from antiSMASH (Medema et al., 2011; Weber et al., 2015), and BOA (Morton et al., 2015). In this mode, MetaMiner first extracts genes related to secondary metabolites and their neighboring fragments from antiSMASH and BOA annotations. Then, putative peptides are constructed based on the ORFs within these clusters that are sequentially similar to known RiPPs. This strategy usually identifies a small number of putative precursor peptides per biosynthetic gene cluster, resulting in a fast and specific peptidogenomics approach. This strategy is not sensitive, as the peptides that do not have sequence similarity to the known RiPPs are missed.

In all-ORF mode, MetaMiner (i) translates DNA sequence into protein sequence using 6-frame translation, (ii) runs HMMer (Eddy, 2011) on the resulting sequences to detect all the modification enzymes, (iii) constructs a 10kbp window centered at each modification enzyme (this window is called the putative biosynthetic gene cluster), (iv) identifies all the ORFs shorter than a pre-defined threshold (default is 200 amino acids) in each putative biosynthetic gene cluster. This approach is capable of discovering RiPPs that do not have any sequence similarity to the known RiPPs.

(b) Constructing target and decoy databases of post-translantionally modified RiPPs. To construct the target RiPP structure database, MetaMiner first searches the biosynthetic gene clusters for all the modification enzymes previously reported in RiPPs (Arnison et al., 2013) using HMMer. If a specific modification enzyme is found in a biosynthetic gene cluster, MetaMiner considers the corresponding modification for the identified precursor peptide/peptides. Table S3 lists the modifications currently considered by MetaMiner, along with the corresponding amino acid residues and mass shifts.

To construct the database of decoy RiPPs, MetaMiner (i) creates a decoy database of RiPP precursor peptides by randomly shuffling each peptide in the target precursor database (Elias and Gygi, 2007), and (ii) creates decoy RiPPs from decoy precursors in the same way as the target RiPPs.

(c) Matching spectra against the target and decoy RiPPs. MetaMiner uses a modified version of Dereplicator (Mohimani et al., 2017) for searching spectra against the database of putative target/decoy RiPPs as follows (i) theoretical spectra for all the target/decoy RiPPs are constructed, (ii) peptide-spectrum matches are generated and scored, (iii) p-values of the peptide-spectrum matches are computed using MS-DPR (Mohimani et al., 2013), (iv) false discovery rates are computed using the decoy database, and (v) statistically significant peptide-spectrum matches are output as putative RiPP identifications.

While exhaustive generation of the 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 (Mohimani et al., 2014a; Pevzner et al., 2000; Pevzner et al., 2001; Tsur et al., 2005). 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 (Mohimani et al., 2014a) can handle modifications in linear peptides, it is not applicable to cyclic peptides. MetaMiner uses a brute-force approach to search all the RiPP modifications of each candidate cyclic peptide against all the spectra. To make this possible, MetaMiner uses a faster scoring strategy that utilized the sparsity of mass spectra and theoretical spectra. We do not currently perform blind modification searches for cyclic peptides due to the inherent computational complexity.

(d) Enlarging the set of identified RiPPs via spectral networking. The set of RiPP identifications is enlarged via spectral networks (Bandeira et al., 2007; Watrous et al., 2012).

Post-translational modifications of RiPPs.—While constructing target and decoy databases of post-translationally modified RiPPs (Figure 1b), MetaMiner considers various types of modifications based on the class of the RiPPs. Table S3 lists the modifications considered by MetaMiner.

Extraction and tandem mass spectrometry.—Below we describe the process of growth (for isolated samples), extraction, and analysis of each dataset by mass spectrometry.

ACTI dataset.: A total of 39 strains of *Streptomyces* were grown on A1, MS and R5 agar, extracted sequentially with ethyl acetate, butanol and methanol, and analyzed on Agilent 6530 Accurate-Mass Q-TOF spectrometer coupled to an Agilent 1260 LC system.

BACIL dataset.: *Bacillus* strains were grown on ISP2 agar, extracted with a solvent mixture of 65:35 acetonitrile:water with 0.05% formic acid, and analyzed on a NanoDESI LTQ-FT (Thermo Electron) mass spectrometer.

SPACE dataset.: Samples from International Space Station were extracted using 50% MeOH and analyzed on a maXis Impact mass spectrometer coupled to C18 RP-UHPLC.

CYANO dataset.: A total of 317 cyanobacterial collections were extracted repetitively with CH_2Cl_2:MeOH 2:1, dried in vacuo, and fractionated into nine fractions (A-I) by silica gel vacuum liquid chromatography (VLC) using a stepwise gradient of hexanes/EtOAc and EtOAc/MeOH, and analyzed on a Maxis Impact mass spectrometer coupled to C18 RP-UHPLC.

HUMAN-iso dataset.: Cultures of reference human microbiome isolates were extracted using 50% EtOH and analyzed on a maXis Impact mass spectrometer (Bruker Daltonics)

coupled to a UltiMate 3000 UPLC system (Thermo Scientific) as described here (Bouslimani et al. 2015).

HUMAN-CF dataset.: Microbial isolates from the sputum culture of cystic fibrosis patients were extracted using 50% MeOH and analyzed on a maXis qTof mass spectrometer coupled to UltiMate 3000 Dionex UPLC system.

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. (Nogle et al., 2003), $[\alpha]^{21}$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 × 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 × 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 MetaMiner identified wewakazole from sample PNG26APR06-4. 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 mg of wewakazole (t_R=13.0 min).

Practical guidelines for MetaMiner.—MetaMiner takes paired genomic and metabolomics data as input. For genomic data, MetaMiner accepts raw nucleotide sequences (.fasta file), antiSMASH output (.final.gbk file) or BOA output (.fasta file). For users who have raw DNA short reads data (.fastq file), we provide a brief guidance about how to

assemble the short read by SPAdes/metaSPAdes in the MetaMiner manual. For metabolomics data, MetaMiner accepts MGF or mzXML files. MetaMiner output is the report of the detected RiPPs in the plain text tab-separated files (.tsv). The spectral network step can be done either through the MetaMiner pipeline, or the GNPS infrastructure (https://ccms-ucsd.github.io/GNPSDocumentation/). For more details, please refer to the MetaMiner manual at https://github.com/mohimanilab/MetaMiner.

QUANTIFICATION AND STATISTICAL ANALYSIS

Estimating false discovery rate—For each dataset, we use 1% false discovery rate (FDR) threshold for RiPP identification. The FDR is estimated with a target-decoy approach. For each target RiPP, we create a corresponding decoy RiPP by first randomly shuffling its precursor peptide sequence and then applying all the modifications of the target RiPP to the shuffled precursor peptide. Given a p-value threshold, denote the number of peptide-spectrum matches in the decoy database and target database as N_{decoy} and N_{target}. The FDR can be estimated as follows:

$$FDR = \frac{N_{decoy}}{N_{target}}$$

DATA AND CODE AVAILABILITY

MetaMiner is available as a command line tool at https://github.com/mohimanilab/MetaMiner.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of H.M. and L.C. was supported by a start-up fund from the Computational Biology Department, Carnegie Mellon University. The work of P.D. and P.A.P. was supported by NIH 2-P41-GM103484. P.D. is supported by GM097509. A.G., A.M. and P.A.P. were supported by St. Petersburg State University, Russia (grant ID PURE 28396291). A.S. and A.K. were supported by Russian Science Foundation (grant 19-14-00172). 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 Oncogenesis (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.

Declaration of Interest

The work of W.H.G was supported by the University of California, San Diego, Scripps Institution of Oceanography, and a grant from the NIH (Award GM118815). W.H.G has an equity interest in Sirenas Marine Discovery, Inc., a company that may potentially benefit from the research results and also serves on the company's Scientific Advisory Board. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. The work of PAP was supported by the US National Institutes of Health grant 2-P41-GM103484. P.A.P. has an equity interest and receives income from Digital

Proteomics, LLC, a company that may potentially benefit from the research results. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

References

Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep 30, 108–160. [PubMed: 23165928]

Bandeira N, Tsur D, Frank A, and Pevzner PA (2007). Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A 104, 6140–6145. [PubMed: 17404225]

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19, 455–477. [PubMed: 22506599]

Blin K, Kazempour D, Wohlleben W, and Weber T (2014). Improved lanthipeptide detection and prediction for antiSMASH. PLoS One 9, e89420. [PubMed: 24586765]

Challis GL, and Ravel J (2000). Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS Microbiol Lett 187, 111–114. [PubMed: 10856642]

Charlop-Powers Z, Pregitzer CC, Lemetre C, Ternei MA, Maniko J, Hover BM, Calle PY, McGuire KL, Garbarino J, Forgione HM, et al. (2016). Urban park soil microbiomes are a rich reservoir of natural product biosynthetic diversity. Proc Natl Acad Sci U S A 113, 14811–14816. [PubMed: 27911822]

Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, Clardy J, Linington RG, and Fischbach MA (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158, 1402–1414. [PubMed: 25215495]

Donia MS, and Fischbach MA (2015). HUMAN MICROBIOTA. Small molecules from the human microbiota. Science 349, 1254766. [PubMed: 26206939]

Duncan KR, Crusemann M, Lechner A, Sarkar A, Li J, Ziemert N, Wang M, Bandeira N, Moore BS, Dorrestein PC, et al. (2015). Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. Chem Biol 22, 460–471. [PubMed: 25865308]

Eddy SR (2011). Accelerated Profile HMM Searches. PLoS Comput Biol 7, e1002195. [PubMed: 22039361]

Fischbach MA, and Walsh CT (2009). Antibiotics for emerging pathogens. Science 325, 1089–1093. [PubMed: 19713519]

Freeman MF, Gurgui C, Helf MJ, Morinaka BI, Uria AR, Oldham NJ, Sahl HG, Matsunaga S, and Piel J (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science 338, 387–390. [PubMed: 22983711]

Gilbert JA, Jansson JK, and Knight R (2014). The Earth Microbiome project: successes and aspirations. BMC Biol 12, 69. [PubMed: 25184604]

Gross H, Stockwell VO, Henkels MD, Nowak-Thompson B, Loper JE, and Gerwick WH (2007). The genomisotopic approach: a systematic method to isolate products of orphan biosynthetic gene clusters. Chem Biol 14, 53–63. [PubMed: 17254952]

Gurevich A, Mikheenko A, Shlemov A, Korobeynikov A, Mohimani H, and Pevzner PA (2018). Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. Nat Microbiol 3, 319–327. [PubMed: 29358742]

Hadjithomas M, Chen IM, Chu K, Ratner A, Palaniappan K, Szeto E, Huang J, Reddy TB, Cimermancic P, Fischbach MA, et al. (2015). IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. MBio 6, e00932. [PubMed: 26173699]

Harvey AL, Edrada-Ebel R, and Quinn RJ (2015). The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov 14, 111–129. [PubMed: 25614221]

Human Microbiome Project, C. (2012a). A framework for human microbiome research. Nature 486, 215–221. [PubMed: 22699610]

Human Microbiome Project, C. (2012b). Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214. [PubMed: 22699609]

Kersten RD, Yang YL, Xu Y, Cimermancic P, Nam SJ, Fenical W, Fischbach MA, Moore BS, and Dorrestein PC (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol 7, 794–802. [PubMed: 21983601]

Lautru S, Deeth RJ, Bailey LM, and Challis GL (2005). Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. Nat Chem Biol 1, 265–269. [PubMed: 16408055]

Li JW, and Vederas JC (2009). Drug discovery and natural products: end of an era or an endless frontier? Science 325, 161–165. [PubMed: 19589993]

Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A, Schaberle TF, Hughes DE, Epstein S, et al. (2015). A new antibiotic kills pathogens without detectable resistance. Nature 517, 455–459. [PubMed: 25561178]

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61–66. [PubMed: 28953883]

Lopez JA, Al-Lihaibi SS, Alarif WM, Abdel-Lateff A, Nogata Y, Washio K, Morikawa M, and Okino T (2016). Wewakazole B, a Cytotoxic Cyanobactin from the Cyanobacterium Moorea producens Collected in the Red Sea. J Nat Prod 79, 1213–1218. [PubMed: 26980238]

Luzzatto-Knaan T, Garg N, Wang M, Glukhov E, Peng Y, Ackermann G, Amir A, Duggan BM, Ryazanov S, Gerwick L, et al. (2017). Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. Elife 6.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, et al. (2018). American Gut: an Open Platform for Citizen Science Microbiome Research. mSystems 3.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, and Breitling R (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39, W339–346. [PubMed: 21672958]

Medema MH, and Fischbach MA (2015). Computational approaches to natural product discovery. Nat Chem Biol 11, 639–648. [PubMed: 26284671]

Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, et al. (2015). Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11, 625–631. [PubMed: 26284661]

Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, and Breitling R (2014). Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comput Biol 10, e1003822. [PubMed: 25188327]

Mohimani H, Gurevich A, Mikheenko A, Garg N, Nothias LF, Ninomiya A, Takada K, Dorrestein PC, and Pevzner PA (2017). Dereplication of peptidic natural products through database search of mass spectra. Nat Chem Biol 13, 30–37. [PubMed: 27820803]

Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, Shcherbin E, Nothias LF, Dorrestein PC, and Pevzner PA (2018). Dereplication of microbial metabolites through database search of mass spectra. Nat Commun 9, 4035. [PubMed: 30279420]

Mohimani H, Kersten RD, Liu WT, Wang M, Purvine SO, Wu S, Brewer HM, Pasa-Tolic L, Bandeira N, Moore BS, et al. (2014a). Automated genome mining of ribosomal peptide natural products. ACS Chem Biol 9, 1545–1551. [PubMed: 24802639]

Mohimani H, Liu WT, Kersten RD, Moore BS, Dorrestein PC, and Pevzner PA (2014b). NRPquest: Coupling Mass Spectrometry and Genome Mining for Nonribosomal Peptide Discovery. J Nat Prod 77, 1902–1909. [PubMed: 25116163]

Mohimani H, and Pevzner PA (2016). Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectral networks. Nat Prod Rep 33, 73–86. [PubMed: 26497201]

Morton JT, Freed SD, Lee SW, and Friedberg I (2015). A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. BMC Bioinformatics 16, 381. [PubMed: 26558535]

Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C, et al. (2013). MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci U S A 110, E2611–2620. [PubMed: 23798442]

Nogle LM, Marquez BL, and Gerwick WH (2003). Wewakazole, a novel cyclic dodecapeptide from a Papua New Guinea Lyngbya majuscula. Org Lett 5, 3–6. [PubMed: 12509876]

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol 20, 714–737. [PubMed: 24093227]

Nurk S, Meleshko D, Korobeynikov A, and Pevzner PA (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Res 27, 824–834. [PubMed: 28298430]

Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers GS, Mavrodi DV, DeBoy RT, Seshadri R, Ren Q, Madupu R, et al. (2005). Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. Nat Biotechnol 23, 873–878. [PubMed: 15980861]

Rince A, Dufour A, Uguen P, Le Pennec JP, and Haras D (1997). 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. [PubMed: 9361411]

Singh NK, Blachowicz A, Checinska A, Wang C, and Venkateswaran K (2016). Draft Genome Sequences of Two Aspergillus fumigatus Strains, Isolated from the International Space Station. Genome Announc 4.

Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, Wyatt MA, and Magarvey NA (2015). Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 43, 9645–9662. [PubMed: 26442528]

Tanizawa Y, Fujisawa T, and Nakamura Y (2018). DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34, 1037–1039. [PubMed: 29106469]

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. [PubMed: 29088705]

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

Venkateswaran K, Checinska Sielaff A, Ratnayake S, Pope RK, Blank TE, Stepanov VG, Fox GE, van Tongeren SP, Torres C, Allen J, et al. (2017). Draft Genome Sequences from a Novel Clade of Bacillus cereus Sensu Lato Strains, Isolated from the International Space Station. Genome Announc 5.

Walsh CT (2015). A chemocentric view of the natural product inventory. Nat Chem Biol 11, 620–624. [PubMed: 26284660]

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34, 828–837. [PubMed: 27504778]

Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, et al. (2012). Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109, E1743–1752. [PubMed: 22586093]

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, et al. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43, W237–243. [PubMed: 25948579]

Wilson MC, Mori T, Ruckert C, Uria AR, Helf MJ, Takada K, Gernert C, Steffens UA, Heycke N, Schmitt S, et al. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506, 58–62. [PubMed: 24476823]

Zhao X, and van der Donk WA (2016). Structural Characterization and Bioactivity Analysis of the Two-Component Lantibiotic Flv System from a Ruminant Bacterium. Cell Chem Biol 23, 246–256. [PubMed: 27028884]

Highlights:

- A tool for discovering and post-translationally modified peptides
- Handles unknown post-translational modifications
- Enables searching millions of spectra against thousands of genomes/ metagenomes
- Searches for lanthipeptides, lassopeptides, linaridins, glycocins, cyanobactins, etc.

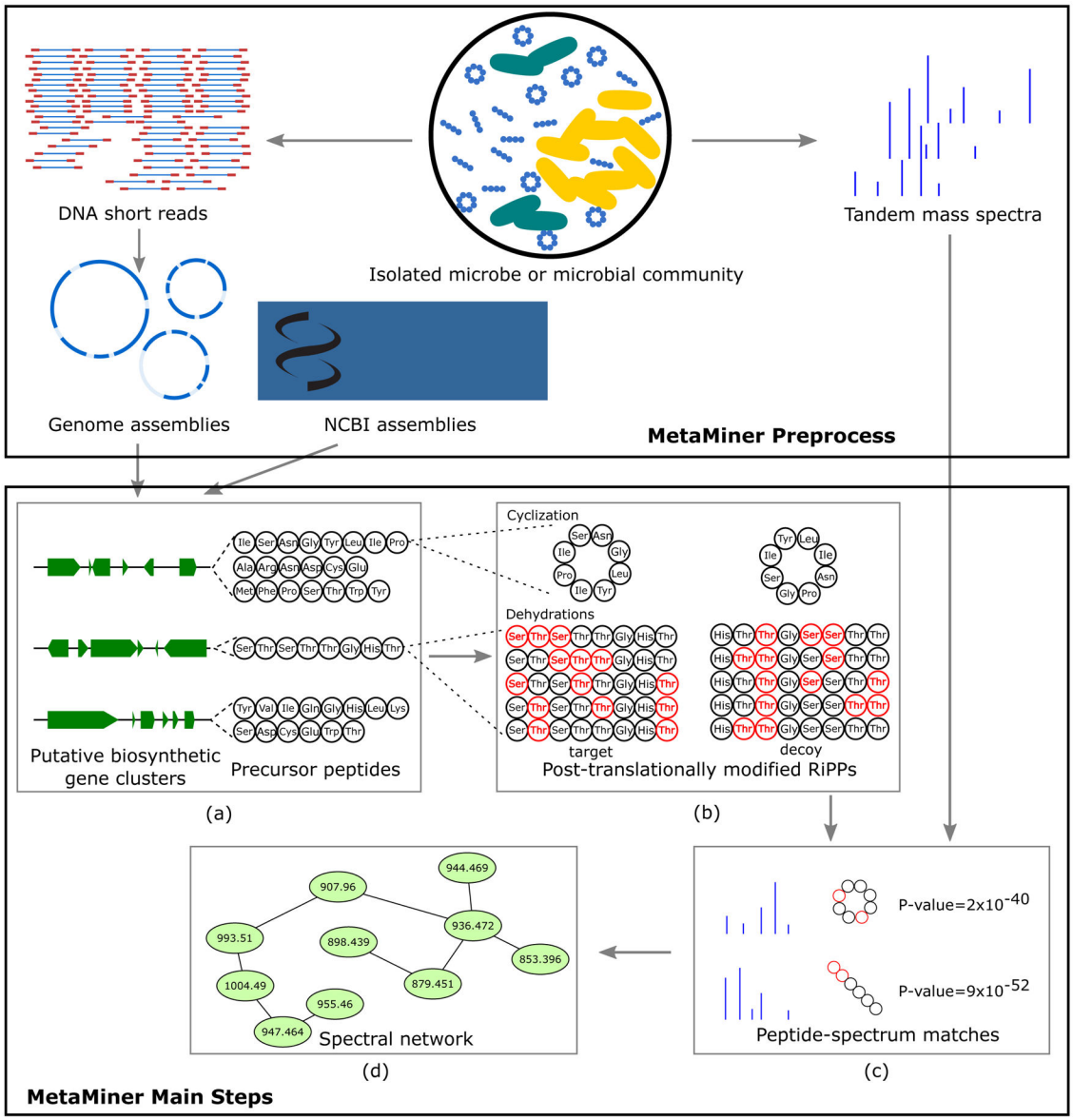


Figure 1.

MetaMiner pipeline analyzes the paired genome/metagenome assemblies and tandem mass spectra from isolated microbes or bacterial/fungal communities. Starting from the genome assemblies, MetaMiner (a) identifies putative Biosynthetic gene cluster and the corresponding precursor peptides, (b) constructs target and decoy putative RiPP structure databases (c) matches tandem mass spectra against the constructed post-translationally modified RiPP structures database using Dereplicator, and (d) enlarges the set of described RiPPs via spectral networking (Bandeira et al., 2007; Watrous et al., 2012). In addition to DNA assemblies, MetaMiner software can also accept various types of input data, including shotgun reads data, antiSMASH output and BOA output.

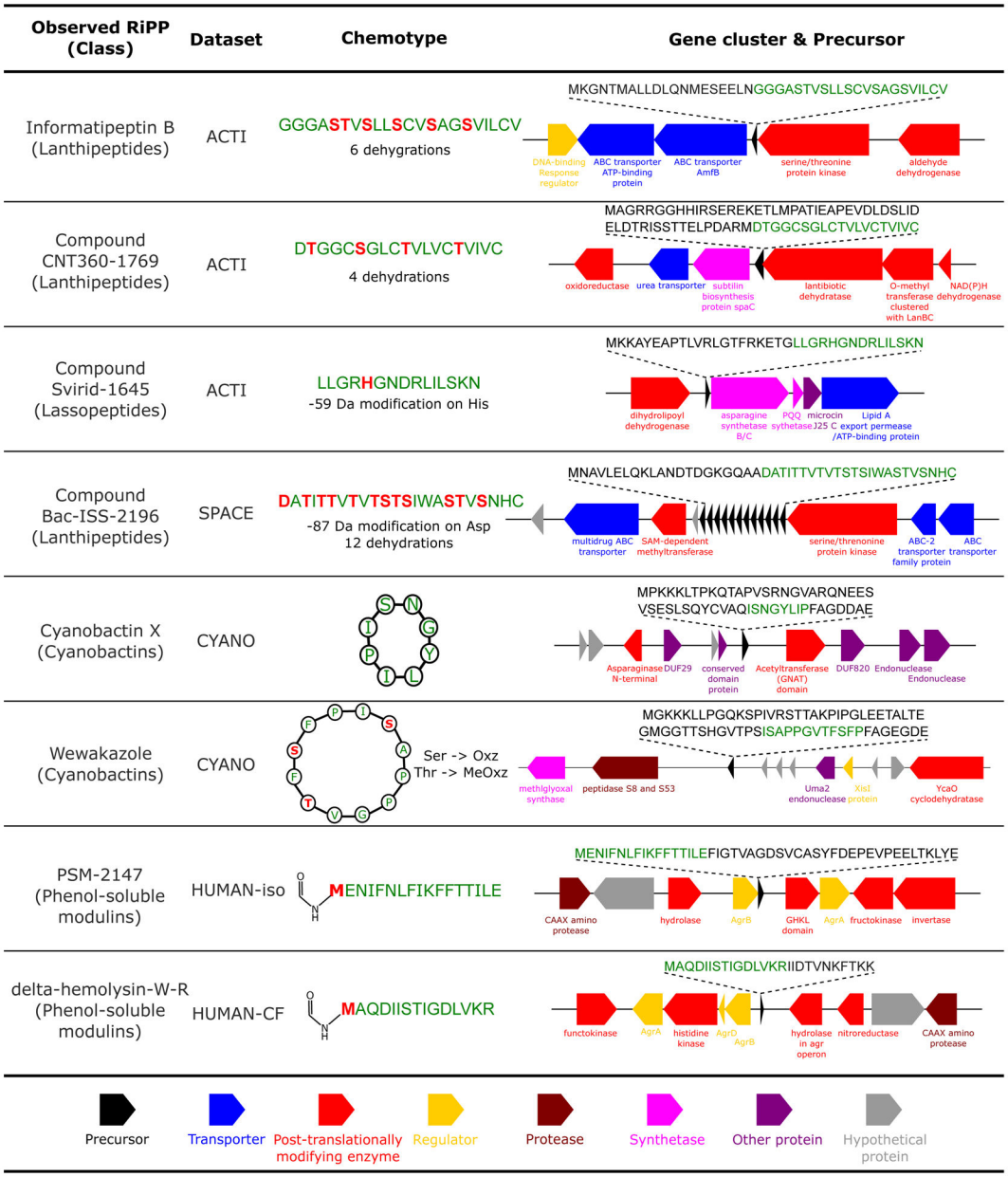


Figure 2.

Chemotypes and putative gene clusters of seven unknown RiPPs discovered by MetaMiner as well as wewakazole. In the column chemotype, post-translationally modified amino acids are shown in bold red. In the precursor column, the green part will be cut and modified, to produce the chemotype. In gene cluster, different colors indicate different types of proteins annotated by DFAST (Tanizawa et al., 2018) and HMMER (Eddy, 2011).

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Standard dataset (STANDARD) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000079506; ftp://massive.ucsd.edu/MSV000079622
Actinomycetes dataset (ACTI) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000078839; ftp://massive.ucsd.edu/MSV000078604
Bacillus dataset (BACIL) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000078552
Space station dataset (SPACE) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000080102
Sponge dataset (SPONGE) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000078670
Cyanobacteria dataset (CYANO) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000078568
Reference Human Microbiome isolates (HUMAN-iso) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000078556
Human Microbiome isolates from Cystic Fibrosis patients (HUMAN-CF) tandem mass spectra	GNPS	ftp://massive.ucsd.edu/MSV000080251
Software and Algorithms		
MetaMiner	This paper	https://github.com/mohimanilab/MetaMiner
antiSMASH	(Medema et al., 2011; Weber et al., 2015)	https://antismash.secondarymetabolites.org/#!/download
BOA	(Morton et al., 2015)	https://github.com/idoerg/BOA
SPAdes	(Bankevich et al., 2012)	https://github.com/ablab/spades
metaSPAdes	(Nurk et al., 2017)	https://github.com/ablab/spades
HMMER	(Eddy, 2011)	http://hmmer.org/
DFAST	(Tanizawa et al., 2018)	https://dfast.nig.ac.jp/
Dereplicator	(Mohimani et al., 2017)	https://github.com/ablab/npdtools
GNPS	(Wang et al., 2016)	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp
Spectral network	(Bandeira et al., 2007; Watrous et al., 2012)	https://ccms-ucsd.github.io/GNPSDocumentation/networking/