Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining

Michael A. Skinnider[a,b,1], Chad W. Johnston[a,b,1], Robyn E. Edgar[a,b], Chris A. Dejong[a,b], Nishanth J. Merwin[a,b], Philip N. Rees[a,b], and Nathan A. Magarvey[a,b,2]

[a]Department of Biochemistry and Biomedical Sciences, M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada L8S 4K1; and [b]Department of Chemistry and Chemical Biology, M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada, L8S 4K1

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved August 26, 2016 (received for review June 3, 2016)

Microbial natural products are an evolved resource of bioactive small molecules, which form the foundation of many modern therapeutic regimes. Ribosomally synthesized and posttranslationally modified peptides (RiPPs) represent a class of natural products which have attracted extensive interest for their diverse chemical structures and potent biological activities. Genome sequencing has revealed that the vast majority of genetically encoded natural products remain unknown. Many bioinformatic resources have therefore been developed to predict the chemical structures of natural products, particularly nonribosomal peptides and polyketides, from sequence data. However, the diversity and complexity of RiPPs have challenged systematic investigation of RiPP diversity, and consequently the vast majority of genetically encoded RiPPs remain chemical "dark matter." Here, we introduce an algorithm to catalog RiPP biosynthetic gene clusters and chart genetically encoded RiPP chemical space. A global analysis of 65,421 prokaryotic genomes revealed 30,261 RiPP clusters, encoding 2,231 unique products. We further leverage the structure predictions generated by our algorithm to facilitate the genome-guided discovery of a molecule from a rare family of RiPPs. Our results provide the systematic investigation of RiPP genetic and chemical space, revealing the widespread distribution of RiPP biosynthesis throughout the prokaryotic tree of life, and provide a platform for the targeted discovery of RiPPs based on genome sequencing.

natural product discovery | chemical space | genome mining | ribosomally synthesized natural product | cheminformatics

Natural products represent an important source of evolved bioactive small molecules, which form the basis for the majority of the small molecule drugs currently used in clinical practice (1). Despite declining discovery rates, genomic data now indicate the vast majority of natural products remain undiscovered (2). This observation has spurred interest in leveraging bacterial genome sequence data for natural product discovery (3–9). Several tools have been developed to integrate genomic data toward the genome-guided discovery of modular, assembly line-derived natural products, including polyketides and nonribosomal peptides, by applying the biosynthetic logic elucidated from the study of model pathways (8–19). However, few systematic strategies target other important classes of natural products, including ribosomally synthesized natural products.

Ribosomally synthesized and posttranslationally modified natural products (RiPPs) are a diverse class of natural products whose biosynthesis proceeds via a ribosomal pathway, followed by extensive posttranslational modification, rather than via modular enzymatic assembly lines. RiPPs are grouped into a number of distinct families based on shared biosynthetic or structural paradigms (20). In prokaryotes, the biosynthetic genes for RiPPs are typically clustered together at a single genetic locus, as with modular natural products. These biosynthetic gene clusters include genes coding for the precursor peptide and a set of biosynthetic enzymes responsible for the serial posttranslational modification of the precursor. The precursor peptide is composed of a core peptide, which is modified to form the final natural product, as well as sequences that flank the core peptide at either the N terminus (leader peptide), C terminus (follower peptide), or both. Leader and follower peptides are removed from the mature natural product by associated proteases. Posttranslational modifications that tailor the core peptide range in complexity from simple head-to-tail macrocyclization, as observed in cyclic bacteriocins, to intricate enzymatic cascades, such as those responsible for thiopeptide biosynthesis (21).

The genetic encoding of RiPPs within small precursor peptides has the potential to facilitate extremely accurate structure predictions, which could be used to guide natural product identification. However, the diversity of RiPPs has challenged the development of tools to predict and identify the numerous families of RiPPs directly from genetic information. To date, no computational framework exists to systematically investigate RiPPs within genetic information and drive the genome-guided discovery of new RiPPs. As a result, the chemical space represented by genetically encoded RiPPs has largely remained dark.

Previously, we developed prediction informatics for secondary metabolomes (PRISM), a software platform to identify NRPS and PKS gene clusters and predict the chemical structures of their associated products (18), and genomes to natural products (GNP), a platform for automated identification of genetically predicted modular natural products in liquid chromatography-tandem MS (LC-MS/MS) data (8). In this work, we provide a structure prediction algorithm that identifies biosynthetic gene

Significance

Natural products and their derivatives are essential to the treatment of many diseases. Ribosomally synthesized and posttranslationally modified peptides (RiPPs) are a class of natural products noted for their bioactivities. Genome sequencing has revealed that most natural products remain undiscovered, but the complexity and diversity of RiPPs has challenged the systematic identification of these molecules from genomic data. Here, we present an algorithm for RiPP structure prediction from prokaryotic genomes and systematically investigate the chemical space occupied by genetically encoded RiPPs. We reveal widespread biosynthesis of RiPPs by prokaryotes, identify candidates for targeted discovery, and isolate a RiPP from a rare family.

Author contributions: M.A.S., C.W.J., and N.A.M. designed research; M.A.S., C.W.J., R.E.E., C.A.D., N.J.M., and P.N.R. performed research; M.A.S. and C.W.J. analyzed data; and M.A.S., C.W.J., and N.A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]M.A.S. and C.W.J. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: magarv@mcmaster.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1609014113/-/DCSupplemental.

clusters and generates libraries of hypothetical structures for 21 families of RiPPs (Fig. 1). Libraries of 154 hidden Markov models and 58 motifs are leveraged to identify RiPP biosynthetic gene clusters, predict precursor peptide cleavage, and execute virtual tailoring reactions, resulting in accurate combinatorial structure prediction across a broad range of RiPP families. We validate leader peptide cleavage and chemical structure predictions, and we mine over 65,000 prokaryotic genomes to characterize the biosynthetic and structural landscape of RiPPs. We assemble known RiPPs to describe their chemical space and compare it to the chemical space occupied by genetically encoded RiPPs. Finally, we combine databases of predicted structures generated by PRISM with LC-MS/MS analysis to reveal a member of a rare RiPP family.

Results

Validating RiPP-PRISM Structure Prediction Accuracy. Having developed an algorithm for RiPP structure prediction from genome sequence data (*Materials and Methods*), we sought to validate its performance on known RiPPs. We first investigated the accuracy of precursor peptide identification and cleavage. A reference dataset of 161 RiPPs with both known biosynthetic gene clusters and experimentally elucidated structures was assembled from the minimum information about a biosynthetic gene cluster (MIBiG) repository (22) (Dataset S1). RiPP-PRISM predicted cleavage sites for 157 of 161 RiPPs (97.5%). Among the four RiPPs without predicted cleavage sites, three were lasso peptides from clusters with multiple precursors. The xanthomonin A2 precursor was identified, but precursor cleavage was not predicted, whereas the caulosegnin III and sphingonodin I precursors were not identified. In all three clusters, at least one other precursor peptide was identified and correctly cleaved. The final RiPP without a predicted cleavage site was the class II lantipeptide lactocin S, whose unique precursor peptide was identified but not cleaved. RiPP-PRISM therefore generated predicted structures for 136 of 137 RiPP clusters (99.3%).

Among precursor peptides with predicted N-terminal leader peptide cleavage, 124 of 157 (79.0%) were correctly predicted. A further 18, or 142 of 157 (90.4%), had predicted leader peptide cleavage sites within a single amino acid of the true cleavage site.

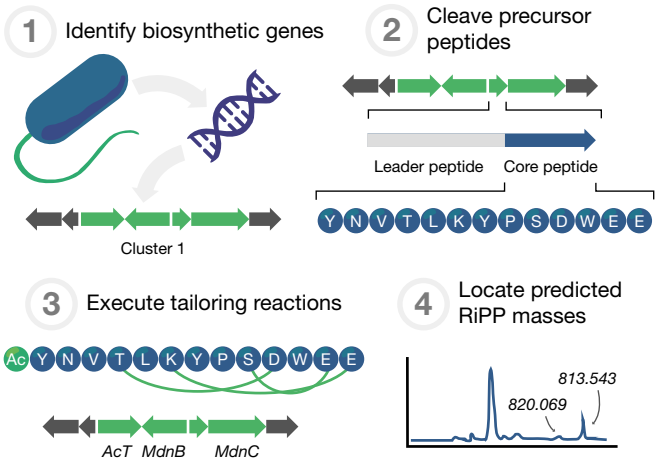

Fig. 1. Schematic overview of a genomic structure prediction algorithm for ribosomally synthesized and posttranslationally modified natural products. A library of 154 hidden Markov models and a set of heuristics for precursor peptides enable the identification and clustering of biosynthetic genes. A library of 53 motifs is used to predict precursor peptide N- and/or C-terminal cleavage. Finally, a set of 94 virtual tailoring reactions are executed based on identified biosynthetic information to generate a combinatorial library of predicted structures. The exact masses of predicted structures can subsequently be searched within a high-resolution LC/MS chromatogram.

Only 7 of 157 precursor peptides (4.5%) had predicted N-terminal leader peptide cleavage that differed by five or more amino acids from the true site. Among precursor peptides with C-terminal follower peptide cleavage, 22 of 24 were predicted correctly (91.7%), whereas predicted sites for the remaining two precursor peptides were within a single amino acid of the true site. The distributions of the differences between true and predicted N- and C-terminal precursor peptide cleavage are plotted in Fig. 2*A*, whereas the dataset of all predicted cleavage sites is included in Dataset S1.

We next validated the accuracy of RiPP-PRISM structure predictions, which leverage combinatorial library generation to elaborate posttranslational modifications to the cleaved precursor peptide. We generated libraries of hypothetical structures for all 136 RiPP clusters and compared true and predicted structures with the Tanimoto coefficient. The median Tanimoto coefficient between each hypothetical structure library and the corresponding true RiPP structure was used as a measure of predictive accuracy (Fig. 2*B*). We observed an average median Tanimoto coefficient of 0.69 ± 0.21, with a range of 0.43–1.0 for each RiPP family. Thiopeptides were the RiPP family with the lowest median Tanimoto coefficient, likely because of the extremely large combinatorial search space, as thiopeptide biosynthesis includes dehydration, heterocyclization, and heterocycle oxidation, and pyridine formation at a minimum, and may additionally include heterocycle methylation, pyridine hydroxylation, esterification, and glycosylation, among other posttranslational modifications. In fact, predicting thiopeptide structures within a reasonable computational time required extensive optimization of RiPP-PRISM to permit random sampling from a biosynthetically plausible combinatorial search space. However, it is notable that the sparse ECFP6 chemical fingerprint, which was chosen on the basis of its excellent performance in virtual screening benchmarks (23), generally produces low scores for any comparison of two structures that are not perfectly identical (24). Comparing the median Tanimoto coefficients for RiPP structure predictions to the median Tanimoto coefficients generated by RiPP-PRISM for thiotemplated structures, which at ∼0.25 currently represent the most accurate genomic structure predictions of NRPS and PKS products (18), provides a more subjective confirmation of the predictive accuracy of RiPP-PRISM for genetically encoded RiPPs.

We also determined the single best Tanimoto coefficient from each hypothetical structure library–true RiPP comparison and observed a significant increase in the average Tanimoto coefficient for each class ($P < 0.02$, Kolmogorov–Smirnov test; *SI Appendix*, Fig. S1). The average best Tanimoto coefficient was 0.84 ± 0.16 and ranged according to the RiPP family from 0.58 to 1.0. These data suggest that, even in clusters where structure prediction based on the identified biosynthetic information involves a large combinatorial search space, RiPP-PRISM is typically able to predict at least one structure with a high degree of chemical similarity to the true structure. The predicted structures with the median and top Tanimoto coefficient for each of the 136 clusters are presented in Dataset S1.

Finally, we tested against overfitting by systematically excluding all sequences from each cluster from the library of sequences within PRISM, then regenerating chemical structure predictions for that cluster (Dataset S1). Only a small decrease was observed in the overall accuracy of chemical structure prediction, with the average median Tanimoto coefficient decreasing from 0.70 to 0.60, and the average maximum Tanimoto coefficient decreasing from 0.84 to 0.73 (*SI Appendix*, Fig. S2 *A* and *C*). These results indicate that the utility of RiPP-PRISM for chemical structure prediction is generalizable beyond the data used to train the algorithm. However, larger decreases in predictive accuracy were observed for the smallest families of RiPPs: glycocins, proteusins, and YM-216931 were no longer predicted at all, and the median Tanimoto coefficient decreased by 25% or more for the linear azole-containing peptides, sactipeptides, and thioviridamide (*SI Appendix*, Fig. S2 *B*

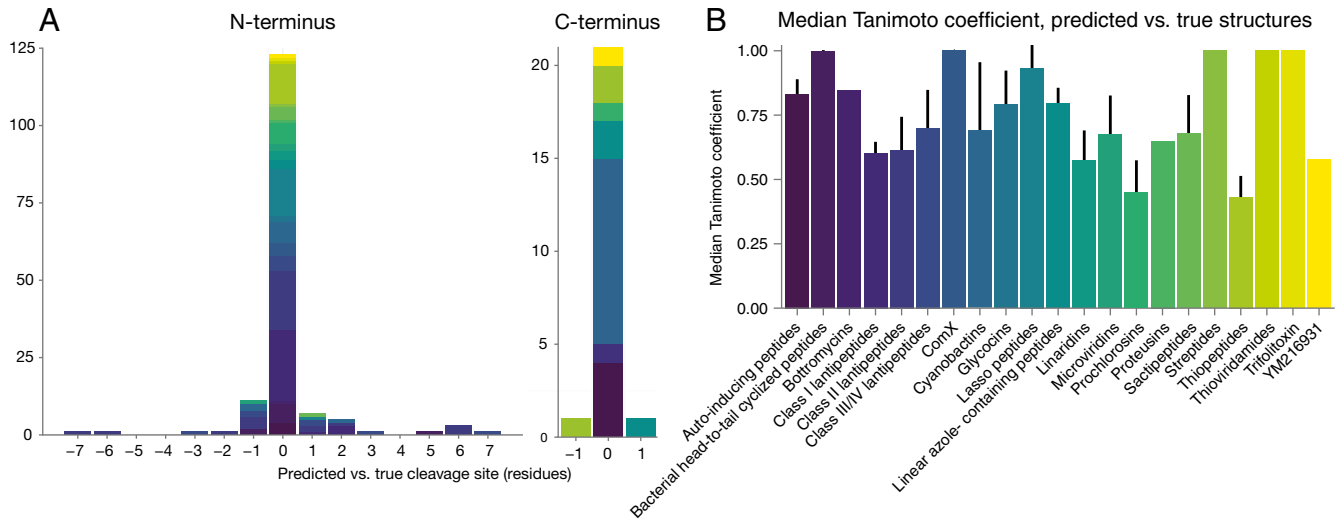

Fig. 2. Validation of RiPP-PRISM predictive accuracy. (*A*) Difference between true and predicted N- and C-terminal leader and follower peptide cleavage sites. (*B*) Average median Tanimoto coefficient between predicted structure libraries and true RiPP structures for 21 families of RiPPs. Error bars show SD.

and *D*), suggesting that RiPP-PRISM performance may be poorer for divergent new members of these small RiPP families.

Global Analysis of Genetically Encoded RiPP Chemical Space. We used RiPP-PRISM to chart the chemical space of genetically encoded RiPPs by analyzing the 65,421 prokaryotic genomes listed in the National Center for Biotechnology Information (NCBI). RiPP-PRISM identified 30,261 biosynthetic gene clusters encoding RiPPs (Fig. 3*A*). Among all prokaryotic genomes, 19,113 (20.0%) contained at least one cluster, corresponding to 2,118 of 12,439 unique species. The genome of the average RiPP-producing organism contained 1.58 ± 0.92 RiPP clusters, but significant variability was observed in biosynthetic potential. A small number of organisms were highly prolific, with 312 microbes (1.6%) producing five or more RiPPs. Three actinomycetes (*Streptomyces mobaraensis*, *Nonomuraea candida*, and *Streptacidiphilus albus*) produced 11 RiPPs, the maximum number of RiPP clusters observed in any single genome.

Although the greatest number of RiPP biosynthetic gene clusters were observed in Firmicutes and Actinobacteria (24,319 and 3,933 clusters, respectively), our results demonstrated that RiPPs are likely produced by nearly all bacterial phyla, including many that have never been known to produce natural products. Among bacterial phyla with at least one cultured representative (noncandidate phyla), we identified RiPP clusters in 17 of 33 phyla with sequenced genomes and failed to identify clusters only in sparsely sequenced phyla (i.e., with fewer than 50 genomes) and in the Tenericutes. Lantipeptides were the most widely distributed RiPPs, with clusters observed in 16 bacterial phyla and in Archaea. Moreover, whereas there exists to date only one example of a lantipeptide from a noncyanobacterial Gram-negative organism (pinensin) (25), we identified class I, II, and III/IV lantipeptide clusters throughout Gram-negative phyla, including Proteobacteria, Acidobacteria, and Gemmatimonadetes. We also identified lantipeptide clusters in unusual producers, such as the obligate intracellular pathogen *Coxiella burnetti*. Proteusins were previously known to be produced only by Cyanobacteria and species of the proposed Tectomicrobia phylum (26), but clusters were observed in Proteobacteria and in two sequenced isolates from the poorly described Verrucomicrobia phylum. Prochlorosins were surprisingly observed in a number of noncyanobacterial families, including α- and δ-proteobacteria, as well as in Verrucomicrobia. Microviridins, another cyanobacterial family of RiPPs, were also found to be more widely distributed than previously appreciated, with clusters in a number of Proteobacteria and Bacteroidetes families. Sactipeptides, which were known to be produced only by Firmicutes, were distributed across Archaea and Bacteria, appearing in six phyla, including several not previously associated with natural product biosynthesis, such as Thermotogae, Fusobacteria, and Dictyoglomi. Chlamydiae had likewise never been appreciated as natural product producers, but several species were observed to possess clusters for class I and II lantipeptides. Trifolitoxin, a narrow spectrum antibacterial agent with activity against Rhizobium spp. (27), was detected in multiple *Acinetobacter* isolates, indicating a potential ecological role for this family of RiPPs within a number of human pathogens. Thiopeptides are promising antibacterial drug candidates, which were previously only associated with Firmicutes and Actinobacteria. However, we observed thiopeptide clusters in a number of species from Proteobacteria, Chloroflexi, Bacteroidetes, and even Deinococcus-Thermus, a phylum previously considered devoid of natural product biosynthetic gene clusters. These results cumulatively demonstrate the surprisingly universal distribution of RiPP biosynthesis throughout the prokaryotic tree of life. We provide all detected RiPP clusters and the taxonomy of their producing organism in Dataset S2, as well as graphical representations of the clusters discussed in this section in *SI Appendix*, Fig. S3.

Analysis of RiPP clusters revealed that the rarest families were YM-216391 family peptides (n = 2), thioviridamide family RiPPs (n = 9), proteusins (n = 9), and bottromycins (n = 9) (Fig. 3*A*). Meanwhile, the most widespread families of RiPPs were autoinducing peptides (n = 8,741), lantipeptides (n = 4,420, 4,373, and 6,074 for classes I, II, and III/IV, respectively), bacterial head-to-tail cyclized peptides (n = 2,927), and lasso peptides (n = 1,466) (Fig. 3*A*). However, many RiPPs are ubiquitous signaling molecules, such as the autoinducing peptides, and consequently it is likely that many of the identified clusters produce identical products. We therefore used structure predictions to identify clusters that produced unique RiPPs. Each of the 24,756 clusters for which RiPP-PRISM predicted at least one structure were compared with one another to generate over 612 million Tanimoto coefficient matrices. A master similarity matrix was constructed by assigning the value of the median Tanimoto coefficient between two libraries of predicted structures to each cluster–cluster comparison, except when the clusters contained at least one identical predicted structure, in which case a value of 1.0 was assigned to the comparison. In this method, highly similar RiPPs, such as the structural isomers epidermin and gallidermin, would be considered distinct products,

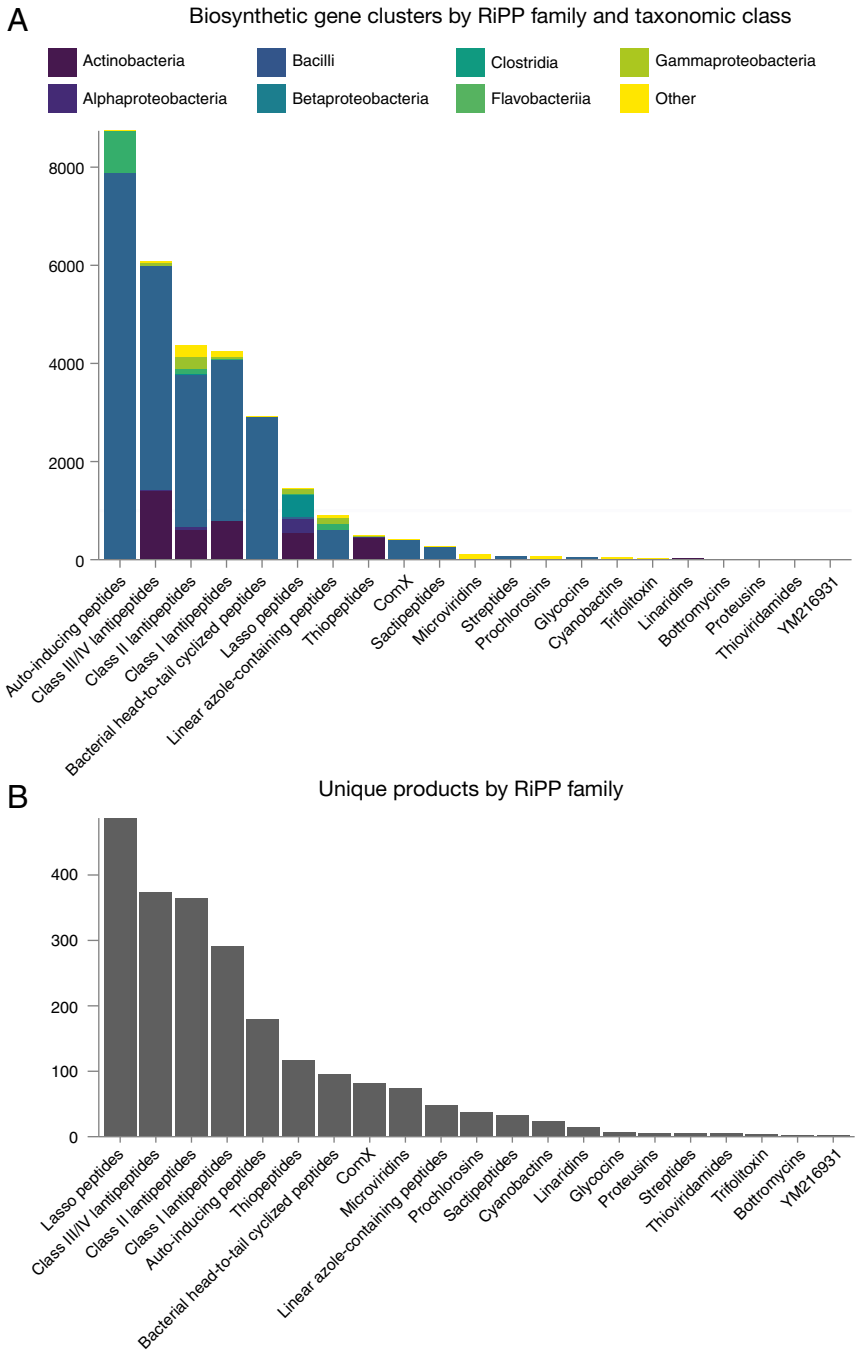

Fig. 3. Genome mining for RiPP biosynthetic gene clusters and their unique products. (*A*) Biosynthetic gene clusters identified in a sample of 65,421 prokaryotic genomes, organized by RiPP family (most abundant first) and taxonomic class of producer organism. "Other" includes all classes with fewer than 100 RiPP clusters. (*B*) Unique products identified by Tanimoto coefficient matrix analysis, organized by RiPP family (most abundant first).

whereas two clusters that encode the same RiPP with low sequence homology or a different enzyme order would not.

Tanimoto coefficient analysis of predicted structure libraries revealed 2,231 clusters producing unique RiPPs among the 24,756 clusters with at least one predicted structure (Fig. 3*B* and Dataset S4). Strikingly, comparing the most abundant unique cluster products to the most abundant clusters revealed a significant reordering of RiPP families: lasso peptides, not autoinducing peptides, are the most abundant when clusters producing the same product are dereplicated. Unique thiopeptides, microviridins, prochlorosins, and cyanobactins are considerably more abundant than the distribution of their clusters had suggested. In contrast, autoinducing peptides, bacterial head-to-tail cyclized peptides, ComX, and streptides are more homogeneous than the distribution of their clusters would suggest. One likely explanation for this discrepancy is the frequent and repetitive sequencing of the producers of these molecules, which are often pathogenic or otherwise human-associated Firmicutes, including *Bacillus*, *Clostridium*, *Staphylococcus*, *Streptococcus*, and *Enterococcus*. Indeed, RiPPs of these families are rarely, if ever, observed outside of the Firmicutes, with our analysis identifying only 25 cyclized bacteriocins and one ComX molecule in non-Firmicutes producers; streptides and

autoinducing peptides do not appear outside of this phylum. Of the 30 most common clusters producing the same product, observed between 110 and 3,697 times, all were from commonly sequenced human pathogens or laboratory strains, with 28 of 30 produced by Firmicutes.

We leveraged RiPP-PRISM structure predictions to chart the chemical space of genetically encoded RiPPs and compared it to the chemical space occupied by known RiPPs. A thorough review of the literature and both public and in-house databases revealed a set of 509 known RiPPs (Dataset S3), which was used to generate a Tanimoto coefficient similarity matrix for known RiPPs. We then used principal component analysis to plot the chemical space of known RiPPs (*Materials and Methods*) with the size of each node corresponding to the number of known RiPPs of each family, and its color corresponding to the within-family chemical diversity as measured by the average median Tanimoto coefficient (Fig. 4*A*). We subsequently used the Tanimoto coefficient matrix of structure predictions for 24,756 clusters to plot the chemical space of genetically encoded RiPPs (Fig. 4*B*).

Comparing the two reveals disparities between the number of known and genetically encoded natural products for many classes: in particular, the genetically encoded lantipeptides and lasso peptides vastly outnumber known products from these families. This analysis also highlights genetically encoded RiPP families with a low level of chemical diversity, such as bacterial head-to-tail cyclized peptides, trifolitoxins, thioviridamides, and streptides: these families of genetically encoded natural products are less likely to represent attractive pharmaceutical or industrial targets. Conversely, cyanobactins and thiopeptides demonstrated the highest within-family chemical diversity. This observation, combined with their broad distribution in organisms not previously known to produce these RiPPs, suggest these families are viable targets for discovery efforts.

These results also provide some insight into the number of RiPPs that remain undiscovered. Thorough investigation of the literature and both public and in-house databases revealed a set of 510 known RiPPs, but Tanimoto coefficient analysis revealed 2,231 unique cluster products. Removing congeners produced by the same cluster reduced the size of the set of known RiPPs to 398. Our analysis therefore suggests that at least 1,833 of 2,231, or 82% of genetically encoded RiPPs remain unknown if all known molecules are currently present in sequenced genomes. However, many known RiPPs were obtained from environmental isolates and other organisms without sequenced genomes, and therefore would not have been detected in this analysis. Thus, 82% is more likely to represent a conservative lower bounds for the percentage of undiscovered RiPPs than a truly accurate estimate, emphasizing our finding that the vast majority of genetically encoded RiPPs remain unknown.

Leveraging Accurate Structure Prediction for Genome-Guided RiPP Discovery. We finally sought to demonstrate the potential of the highly accurate structure predictions generated by RiPP-PRISM to facilitate the targeted, genome-guided discovery of novel RiPPs. Bioinformatic analysis revealed that natural products of the YM-216391 family were among the rarest RiPPs, with only two biosynthetic gene clusters identified in a sample of 65,421 genomes. Tanimoto coefficient matrix analysis suggested that neither cluster product was identical to any of the three known members of this family, which include urukthapelstatin and mechercharstatin in addition to YM-216391, the sole representative with a sequenced cluster. All three products are characterized by nanomolar cytotoxicity and a conserved, azole-rich macrocyclic structure. Putative YM-216931 family clusters were identified in the genomes of *Streptomyces aurantiacus* JA 4570 and *Streptomyces curacoi* DSM 40107. Because *S. aurantiacus* JA 4570 has been the subject of intense investigation over the past two decades, and is known to produce at least three distinct classes of natural products with diverse activities, we reasoned that it would be a useful target to demonstrate the utility of our approach.

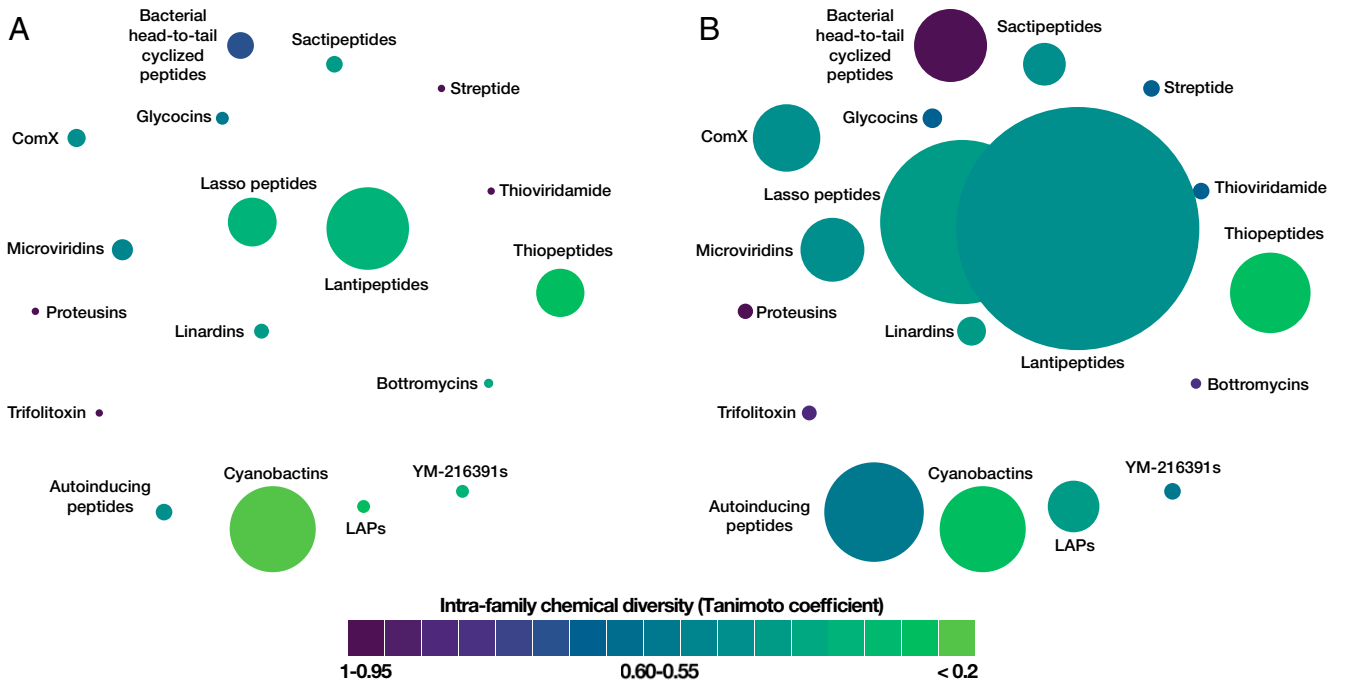

Fig. 4. Charting the chemical space of known and genetically encoded RiPPs. (*A*) Principal component analysis plot of 509 known ribosomal products, organized into 18 families, with node size corresponding to number of known RiPPs and node color corresponding to within-family chemical diversity (average median Tanimoto coefficient). (*B*) Principal component analysis plot for genetically encoded RiPPs, with node size corresponding to number of unique predicted RiPPs and node color corresponding to average median Tanimoto coefficient across all identified clusters.

We leveraged RiPP-PRISM structure predictions to facilitate targeted identification of natural products in LC-MS/MS data from bacterial extracts. Given a library of structures in SMILES format predicted by RiPP-PRISM, LC-MS/MS searches can be readily automated (8) for parent and daughter ion mass predictions. *S. aurantiacus* was cultured in a panel of 20 media for 3 d before cell pellets were collected and extracted with organic solvent, which was then processed using high-resolution LC-coupled MS (HR-LC/MS). Chromatograms from each of the media conditions were analyzed by searching for the library of 15 predicted structures (*SI Appendix*, Fig. S4) from the putative *S. aurantiacus* YM-216391 family cluster. A predicted structure was only identified in one media condition and at an extremely low concentration, with a mass consistent with the fully oxidized azole ring system of predicted scaffold 10, which was named aurantizolicin (Fig. 5). MS/MS fragmentation displayed the predicted tandem isoleucines, and incorporation of deuterium-labeled phenylalanine (d8-Phe) demonstrated incorporation of five deuterium atoms (with the other three being lost during hydroxylation, heterocyclization, and aromatization; *SI Appendix*, Fig. S5). Culture conditions were serially optimized for production, culminating in a large culture effort to isolate sufficient quantities of aurantizolicin to confirm the predicted structure with NMR spectroscopy. As expected, aurantizolicin possesses a number of conserved and highly predictable features unique to this small class of RiPPs, and is closely related to YM-216391, with nearly identical NMR chemical shifts (*SI Appendix*, *SI Text*). Like other RiPPs of this family, aurantizolicin is produced in extremely small quantities and consequently was not discovered despite previous bioactivity-guided studies. However, its identification by RiPP-PRISM indicates our platform can be leveraged toward the targeted discovery of novel, genetically encoded RiPPs.

Discussion

Microbial natural products and their derivatives are essential to the treatment of many diseases, and RiPPs are a class noted for their biological activities. However, the structural and biosynthetic diversity of RiPPs has to date challenged the development of a comprehensive platform for the analysis of genetically encoded RiPPs. The scope of current models of RiPP biosynthesis imposes inherent limits on the predictive accuracy of any platform designed for the systematic analysis of genetically encoded RiPPs, and necessarily precludes the development of software capable of perfectly classifying and predicting chemical structures for every RiPP cluster. However, the development of RiPP-PRISM has here enabled a systematic investigation of RiPP biosynthesis in over 60,000 prokaryotic genomes. Our analysis demonstrates the nearly universal distribution of RiPPs throughout the prokaryotic tree of life, with RiPP clusters absent from only noncandidate bacterial phyla with fewer than 50 sequenced genomes and the Tenericutes. We also observed the surprisingly widespread taxonomic distribution of many RiPP families thought to be produced by only one or two phyla, including prochlorosins, microviridins, and cyanobactins, and identified RiPP clusters in several phyla that had not previously been known to produce natural products. We leveraged accurate structure predictions to chart the chemical space of genetically encoded RiPPs, dereplicating clusters that produce the same products and ranking RiPP families according to their chemical diversity. Our analysis indicates that the vast majority of RiPPs remain undiscovered and provides a lower bounds for an estimate of the number of unknown RiPPs. Finally, we use the structure predictions of our bio- and cheminformatic platform to identify and isolate a member of the smallest family of RiPPs, those related to the cytotoxic peptide YM-216391.

This work expands significantly on previous approaches developed to facilitate genome-guided RiPP analysis or discovery. Three general properties distinguish RiPP-PRISM from previous efforts. Its ability to handle 21 families of RiPPs makes it a uniquely extensive platform. Moreover, its ability to generate predicted structures separates it from any existing tool. Finally, its application to global genomic analysis requires no manual annotation. Previously, Maksimov et al. (28) used a combination of heuristic precursor identification and motif-guided identification of biosynthetic enzymes to mine 3,000 genomes for lasso peptides and identify a novel lasso peptide, but their approach was limited to this family of RiPPs. Efforts capable of genome mining for RiPPs of multiple classes have often been limited by the amount of manual annotation required to identify clusters. For instance, Letzel et al. (29) used a set of three bioinformatics tools and BLAST searches to mine the genomes of 211 anaerobic bacteria for six families of RiPPs, but could not generate predicted

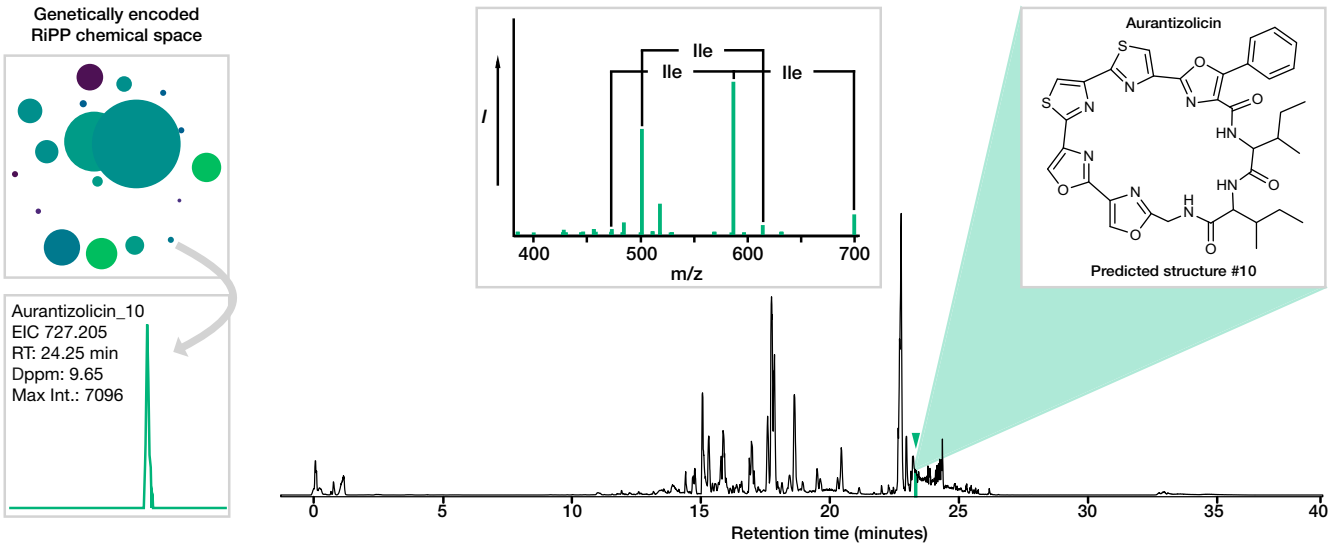

Fig. 5. Genome-guided isolation of a member of the rare YM-216391 family of RiPPs. Hypothetical structures for a *S. aurantiacus* RiPP were generated by RiPP-PRISM and were used to search LC-MS/MS chromatograms. A single peak was identified corresponding to predicted structure no. 10 of 15, including MS/MS data, which demonstrated that this candidate structure was the predicted molecule, aurantizolicin.

structures or match clusters to unique products except by manual annotation. Likewise, Cox et al. (30) used YcaO proteins to identify thiazole/oxazole-modified macrocycles (TOMMs), a superfamily encompassing members of linear azol(in)e-containing peptides, cyanobactins, thiopeptides, and bottromycins, and identified nearly 1,500 clusters; however, their method required extensive manual annotation. Whereas automated methods for RiPP genome mining exist, these are often limited to a subset of RiPP families. The widely used antiSMASH platform (19), for instance, is capable of identifying putative lantipeptide clusters and predicting their cleavage sites, but can neither generate predicted structures nor identify other classes of RiPPs. Cimermancic et al. (31) developed a machine-learning algorithm, ClusterFinder, to identify clusters of both known and unknown classes based on Pfam domain content. Their method identified several hundred RiPP clusters in a sample of 1,154 genomes, but was not capable of predicting the structures of their corresponding products. Moreover, manual annotation was required to identify the family of RiPPs that each cluster belonged to. BAGEL3 (32) is capable of identifying 12 families of RiPPs, but this platform cannot predict precursor cleavage or generate predicted structures. In contrast, RiPP-PRISM uses a library of 154 hidden Markov models to identify 21 families of RiPPs, an ensemble of hidden Markov models and heuristics to identify putative precursors, a set of 54 motifs to predict precursor cleavage, and a library of 94 virtual tailoring reactions to generate highly accurate predicted structures, making it a uniquely extensive resource for RiPP genome mining. The scale of the genome mining effort presented here, with the identification of over 30,000 clusters in over 60,000 prokaryotic genomes, as well as the ability of RiPP-PRISM to dereplicate clusters that produce the same natural product both attest to its advantages as a platform for RiPP discovery.

Although no existing platform is capable of predicting the structures of genetically encoded RiPPs, some existing methods are capable of automating, to varying degrees, the processes of RiPP cluster identification and classification. We compared the ability of RiPP-PRISM, antiSMASH 3.0, and ClusterFinder to identify RiPP clusters within a set of 5,049 complete microbial genomes from NCBI (Dataset S5 and *SI Appendix*, *SI Text*). Of the 2,258 unique clusters identified, 912 (40.4%) were detected by all three methods, 712 (31.5%) by two methods, and the remaining 634 (28.1%) by only one method (*SI Appendix*, Fig. S6 *A*, *D*, and *E*). A total of 586 clusters were identified by RiPP-PRISM only (30.0%), whereas 216 clusters were not identified by RiPP-PRISM (9.6%). However, manual inspection of clusters not identified by PRISM suggested a large fraction represented false positives: for instance, isolated homologs of the ATP-grasp enzyme RimK were identified by antiSMASH as microviridin clusters; a large number of putative lantipeptide clusters detected by antiSMASH based on a "LanC-like" hidden Markov model did not appear to have any of the enzymatic machinery associated with the biosynthesis of known lantipeptides; and isolated homologs of the PoyD radical SAM epimerase were identified by antiSMASH as proteusin clusters. Manual annotation of each of the 216 clusters not detected by RiPP-PRISM suggested that up to 87% corresponded to putative false positives. Although many such putative false positives were detected by both antiSMASH and ClusterFinder, we note that ClusterFinder predicted 198,302 clusters in a sample of 5,049 genomes, of which the vast majority are themselves likely to represent false positives; consequently, identification by ClusterFinder does not provide a high-confidence independent confirmation of putative clusters identified by antiSMASH. PRISM additionally identified at least 10% more clusters than antiSMASH and ClusterFinder combined, for 6 of 21 RiPP families (*SI Appendix*, Fig. S6 *F* and *H*), and exhibited less bias toward actinomycete clusters than either ClusterFinder or antiSMASH (*SI Appendix*, Fig. S6 *B*, *C*, *G*, and *I*). The results of a comparative analysis on a large set of genomes suggest that, in addition to introducing unique chemical structure prediction functionality, RiPP-PRISM expands the scope of RiPP genome mining by revealing a greater number of RiPP clusters, particularly for atypical producing organisms.

The approach presented here relies fundamentally on homology to known clusters and experimentally elucidated biosynthetic transformations to identify RiPP clusters from sequence data and predict the structures of their products. Although this approach enables the characterization of the biosynthetic and structural landscape of genetically encoded RiPPs with known chemotypes, it has at least two significant limitations with respect to RiPP discovery. In particular, our approach cannot identify novel families of RiPPs, and it cannot predict the presence or mechanisms of novel enzymatic tailoring reactions. Therefore, it is nearly certain that this analysis will have failed to identify RiPP clusters corresponding to novel chemotypes, and likewise will have failed to predict the action of enzymatic tailoring reactions with little or no homology to known enzymes in RiPP biosynthesis. However, these failings are not unique to the platform described here: to date, no computational strategy is capable of predicting the action of novel tailoring enzymes in natural product biosynthesis except by homology to known enzymes; and whereas some machine learning strategies are capable of identifying clusters from previously unknown families (31), they are not capable of predicting the structures of the genetically encoded products. Moreover, we emphasize that although our validation demonstrated PRISM is capable of highly accurate structure predictions, perfect accuracy is not necessary to dereplicate clusters that produce the same product, because PRISM will always generate the same predicted structures from the same set of identified biosynthetic information.

Several other limitations of PRISM should be highlighted. Some families of RiPPs are thought to be specific to eukaryotes, but because PRISM is designed for the analysis of prokaryotic genomes, these RiPP families were not included in the present analysis. Current models of RiPP biosynthesis are incomplete and therefore limit the accuracy of cluster detection and assignment to known RiPP families for divergent new clusters, particularly for families with a small number of available sequences. The analysis presented in *SI Appendix*, Fig. S2 suggests this is particularly likely to be the case for glycocins, proteusins, and YM-216931 family RiPPs, and that new members of these families were most likely to have remained undetected by our analysis. Finally, many microbial genomes exist as drafts or low-quality assemblies, a fact that may have prevented the identification of some fragmented clusters and furthermore has the potential to bias cluster identification and structure prediction for user-submitted sequences.

Although we identified 30,261 RiPP clusters in a sample of 65,421 prokaryotic genomes, PRISM generated predicted structures for only 24,756 (81.8%). This figure is considerably lower than the fraction of clusters for which at least one predicted structure was generated during validation on known products (99.3%). We plotted the remaining 5,505 clusters by RiPP family and producing organism taxonomy in *SI Appendix*, Fig. S7. Class I, II, and III/IV lantipeptides cumulatively accounted for over 86% of clusters without a structure predicted. This can likely be attributed to the diversity of lantipeptide precursors, which likely precluded precursor identification or cleavage for lantipeptides with little homology to known precursors. The only other families of RiPPs with more than 100 clusters without structure predictions were bacterial head-to-tail cyclized peptides and thiopeptides; these failures can presumably be attributed to the same factors as lantipeptides, as well as the absence of a heuristic for bacterial head-to-tail cyclized peptide precursor identification. Manual inspection of thiopeptide clusters without predicted structures revealed a large number (>50) from the frequently sequenced genus *Salinispora*, where a large transposon insertion had occurred between the precursor and biosynthetic genes, placing it outside of the cluster size

range considered by RiPP-PRISM: it is possible that transposon insertion produced similarly incomplete results for other clusters.

In this work, we have introduced a comprehensive platform for identifying RiPP gene clusters from prokaryotic genomes and predicting their chemical structures with a high degree of accuracy. We have leveraged this platform to conduct a global analysis of genetically encoded RiPPs, revealing these molecules are ubiquitous throughout prokaryotic phyla, with a substantial majority remaining unknown. Finally, by creating highly accurate predicted structures, RiPP-PRISM facilitates the targeted detection of new molecules from LC-MS/MS data based on genome sequence data, leading in this work to the discovery of a natural product from a rare family of RiPPs. Our results highlight the advantages of this platform for the genome-guided discovery of novel RiPPs.

Materials and Methods

General Computational Methods. Hidden Markov models were constructed by manual compilation of experimentally annotated sequences, which were supplemented with homologs identified by querying the Integrated Microbial Genomes database (IMG) (33) and NCBI BLAST (34) databases. Sequences were aligned with MUSCLE (version 3.8.31) (35) and trimmed using trimAl (version 1.2rev59) (36) to remove gaps. Hidden Markov models were compiled from the trimmed alignments using hmmbuild (version 3.1b1) (37) and bitscore cutoffs were determined by manual analysis of the results of searches of the UniProtKB and UniProt reference proteome databases (38). Motif discovery was performed using the MEME web server (39), allowing any number of motif occurrences with a minimum motif length of 6 aa and a maximum length of 25 aa. The Chemistry Development Kit (version 1.5.10) (40) implementation of the ECFP6 chemical fingerprint (41) was used to calculate Tanimoto coefficients.

Development of an Algorithm for Genome-Guided Chemical Structure Prediction of RiPPs. We developed an algorithm to identify biosynthetic gene clusters and predict chemical structures for 21 families of RiPPs by extending the open-source PRISM framework (18). PRISM is a Java 8 web application built for the Apache Tomcat 7 web server, which implements BLAST (version 2.2.25+) (42), HMMER (version 3.1b1) (43), BioJava (version 3.0.7) (44), BioPerl (version 1.006924) (45), RDKit (version 2014.03.1), the Chemistry Development Kit (version 1.4.19) (40), Prodigal (version 2.6.2) (46), and FIMO (version 4.11.0) (47). PRISM queries a user-input sequence with a library of several hundred hidden Markov models and curated BLAST databases to identify nonribosomal peptide, type I and II polyketide, deoxy sugar (8), and resistance (9) domains. The results of this search are used to identify natural product biosynthetic gene clusters. Linear scaffolds are generated and elaborated in a combinatorial manner based on predicted tailoring reactions, deoxy sugar moieties, and cyclizations to generate a combinatorial library of predicted structures.

We extended PRISM by developing 53 motifs, 150 hidden Markov models, and 94 virtual tailoring reactions specific to RiPP biosynthesis and developed rules for the identification of biosynthetic gene clusters for 21 families of RiPPs. Precursor peptides are identified with a combination of hidden Markov models and heuristic strategies and are cleaved at their N terminus and/or C terminus based on conserved motifs. Tailoring reaction domains are identified, and all potential reaction sites are determined. Virtual reactions are then executed combinatorially to produce a library of hypothetical structures corresponding to the identified biosynthetic gene cluster. A detailed description of precursor peptide cleavage and tailoring reaction execution is presented in *SI Appendix, SI Text*. We note that although tailoring reactions are described in the *SI Appendix, SI Text* according to the biosynthetic family, each tailoring reaction can occur within any class of RiPPs: thus, for example, the identification of a putative domain with homology to cyanobactin prenyltransferases within a lantipeptide cluster would result in the generation of *O*- or *N*-prenylated hypothetical lantipeptide structures. Furthermore, any precursor peptide associated with a cluster will result in the generation of predicted structures, regardless of cluster family.

All hidden Markov models developed in this study are presented in Dataset S6, table 1. All motifs developed in this study are presented in Dataset S6, table 2. All virtual tailoring reactions developed in this study are presented in Dataset S6, table 3. All rules for RiPP cluster detection are presented in *SI Appendix*, Table S1. We additionally provide graphical representations of the genes required to identify clusters of each RiPP family, and the chemical transformations associated with each gene, in Dataset S7. RiPP-PRISM is integrated into the publicly available PRISM web application at magarveylab.ca/prism, whereas source code is available at github.com/magarveylab/prism-releases.

Validation of Predictive Accuracy. PRISM was run on 136 known RiPP clusters with the following settings: tailoring, deoxy sugar, and RiPP domain HMM searches enabled; both Prodigal and all potential coding sequences used to identify ORFs; and cluster scaffold library limit of 100. Predicted and true structures were compared with the Tanimoto coefficient.

Global Analysis of Genetically Encoded RiPP Chemical Space. A total of 65,426 prokaryotic genomes were retrieved from NCBI Genome in March 2016. PRISM was run on each genome with the following settings: tailoring, deoxy sugar, and RiPP domain HMM searches enabled; Prodigal used to identify ORFs; cluster window of 5,000 bp; and a cluster scaffold library limit of 100. A total of 65,421 of 65,426 genomes were successfully run through PRISM. Taxonomic information for each genome was retrieved with the ETE module in Python (48). JSON output from PRISM was parsed to retrieve all predicted structures for each cluster and the RiPP family or families of the cluster. All predicted structures from each of the 24,756 clusters for which predicted structures were generated were compared with one another to generate Tanimoto coefficient matrices ranging in size from 1×1–100×100. The median value of all Tanimoto coefficients within the matrix was assigned to the cluster–cluster comparison unless the matrix contained one or more instances of 1.0, in which case the clusters were understood to produce the same product and a value of 1.0 was assigned to the cluster–cluster comparison.

Plotting RiPP Chemical Space. A Tanimoto coefficient similarity matrix was generated for a comprehensive collection of 509 known ribosomal natural products, including molecules from 18 distinct families. ECFP6 chemical fingerprints were used to calculate Tanimoto coefficients. Molecules with the highest median within-family Tanimoto coefficients were taken as representative structures, and the resulting 18-member Tanimoto similarity matrix was plotting with multidimensional scaling (MDS) using XLSTAT 2016. MDS was performed with default settings, presenting final results in two dimensions using an absolute model, measuring correspondence between the matrix input and final distances via Kruskal stress equation 1. Minimization was completed after maximum convergence (0.00001) was reached. Points for representative structures in the resulting plot were enlarged such that their area corresponded to the number of members of their associated RiPP class, with color corresponding to the median within-family Tanimoto coefficient. Lantipeptides of classes I, II, III, and IV were here classed as a single family, as this plot includes all known compounds, including structures whose biosynthetic origins are undetermined. To plot genetically encoded RiPP chemical space, the number of unique products was used to determine the size of each node, and the average median within-family Tanimoto coefficient was used to color nodes according to within-family chemical diversity.

General Experimental Procedures. NMR spectra 1D (1H and 13C) and 2D [1H-13C heteronuclear multiple bond correlation spectroscopy (HMBC), heteronuclear single quantum coherence spectroscopy (HSQC), nuclear Overhauser effect spectroscopy (NOESY), total correlation spectroscopy (TOCSY), and homonuclear correlation spectroscopy (COSY)] for aurantizolicin was recorded on a Bruker AVIII 700 MHz NMR spectrometer in d6-DMSO (Sigma-Aldrich). High-resolution LC-MS/MS spectra were collected on a SciEX 5600+ TripleTOF mass spectrometer (ABSciEX) with an electrospray ionization (ESI) source and using CID with helium for fragmentation, coupled with an Agilent 1100 series HPLC system using an Ascentis Express C8 column (150 mm $\times$ 2.1 mm, 2.7 mm; Sigma-Aldrich) for analytical separations, running acetonitrile with 0.1% formic acid, and double-distilled (dd)-H2O with 0.1% formic acid as the mobile phase. Preparative HPLC was performed using a Dionex UltiMate 3000 HPLC system, using a Luna C8 column (250 mm $\times$ 10 mm; Phenomenex), running acetonitrile with 0.1% formic acid and ddH2O with 0.1% formic acid as the mobile phase.

Microbial Strains and Culturing. *S. aurantiacus* JA 4570 was obtained from the Hans Knöll Institute (IMET 43917). It was maintained on Bennett's agar at 28 °C. Bennett's agar contains 15 g/L agar, 1 g/L beef extract, 1 g/L yeast extract, 2 g/L NZ-amine, and 10 g/L glucose, with pH adjusted to 7.3. KE media contains 1 g/L glucose, 10 g/L potato dextrin, 5 g/L yeast extract, 5 g/L NZ-amine, 3 g/L beef extract, 0.5 g/L calcium carbonate, and 0.05 g/L magnesium sulfate heptahydrate. After autoclaving, 2 mL/L sterile phosphate buffer (91 g/L potassium phosphate monobasic and 95 g/L potassium phosphate dibasic, pH 7) was added. Aurantizolicin was initially detected in a media containing 20 g/L sodium chloride, 10 g/L soluble starch, 0.3 g/L casein, 2 g/L potassium nitrate, 2 g/L potassium phosphate dibasic, 0.05 g/L magnesium sulfate heptahydrate, 0.02 g/L calcium carbonate, and 0.01 g/L iron (II) sulfate heptahydrate. Aurantizolicin production media contains 20 g/L sodium chloride, 10 g/L soluble starch, 3 g/L casein, 2 g/L potassium nitrate, 2 g/L potassium phosphate dibasic, 0.05 g/L

magnesium sulfate heptahydrate, 0.02 g/L calcium carbonate, 0.01 g/L iron (II) sulfate heptahydrate, and 30 g/L washed HP-20 resin (Diaion).

Production and Isolation of Aurantizolicin. *S. aurantiacus* was grown on Bennett's agar for 2 wk at 28 °C. Colonies were added to 250 mL Erlenmeyer flasks containing 50 mL of KE media and grown for 48 h at 28 °C and 225 rpm. From these, 10 mL of culture was added to a 2.8-L Fernbach flask containing 1 L of aurantizolicin production media, which was incubated at 28 °C and shaken in a Kuhner ISF4-X shaker incubator (Basel, Switzerland) at 225 rpm for 72 h. Mycelial mass and resins were collected via Buchner funnel vacuum filtration using nongauze milk filters (KenAG). The resultant cell pellet and resin cake was extracted multiple times with excess acetone that was evaporated to dryness using a rotary evaporator. The dried extract was resuspended and partitioned in 1:1 butanol:water, collecting the butanol phase and evaporating it to dryness. Dried extract was resuspended in methanol and loaded on a large open column containing LH20 resin (Sephadex) with methanol as the mobile phase. Fractions containing aurantizolicin were pooled, dried down, and resuspended in a small volume of DMSO. Aurantizolicin was purified by semipreparative scale HPLC, using a Luna C8 column (250 × 10 mm) with ddH2O and acetonitrile containing 0.1% formic acid as the mobile phase, pumping at 8 mL/min. Acetonitrile was 5% for the first 3 min, ramping to 50% by 10 min, then to 80% by 23 min, then proceeding to 100% at 24 min. Aurantizolicin eluted at 20.2 min. A total of 60 L of optimized production culture provided <1 mg of aurantizolicin. HPLC fractions containing aurantizolicin were dried down and resuspended in d6-DMSO for NMR. Structure elucidation of aurantizolicin, including HRMS measurement, NMR spectra, chemical shift table, and detailed description of chemical shift correlations are provided in *SI Appendix, SI Text*.

Incorporation of d8-Phenylalanine into Aurantizolicin. *S. aurantiacus* was grown on Bennett's agar for 2 wk at 28 °C. Colonies were added to 250 mL Erlenmeyer flasks containing 50 mL of KE media and grown for 48 h at 28 °C and shaken in a Kuhner ISF4-X shaker incubator (Basel, Switzerland) at 225 rpm. From these flasks, 1 mL of culture was added to a 250-mL Erlenmeyer flask containing 50 L of aurantizolicin production media without HP20 resin, which was incubated at 28 °C and 225 rpm for 72 h. After 24 h, d8-phenylalanine was added via sterile syringe filtration to a final concentration of 4 mM. Cultures were extracted and analyzed by LC-MS/MS.

ACKNOWLEDGMENTS. This work was supported by the National Science and Engineering Research Council (RGPIN 371576-2009, 101997-2006, and 388681-2011). Support for N.A.M. was obtained through a Canadian Institute for Health Research (CIHR) New Investigator Award, Ontario Early Investigator Award, the Canada Research Chairs Program, a CIHR Operating Grant, and a CIHR–Joint Programming Initiative on Antimicrobial Resistance Research Grant. C.W.J. is supported by a CIHR Doctoral Fellowship.

1. Newman DJ, Cragg GM (2016) Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79(3):629–661.
2. Nett M, Ikeda H, Moore BS (2009) Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat Prod Rep* 26(11):1362–1384.
3. Lautru S, Deeth RJ, Bailey LM, Challis GL (2005) Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat Chem Biol* 1(5):265–269.
4. Nguyen T, et al. (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol* 26(2):225–233.
5. Kersten RD, et al. (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat Chem Biol* 7(11):794–802.
6. Doroghazi JR, Metcalf WW (2013) Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 14:611.
7. Doroghazi JR, et al. (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 10(11):963–968.
8. Johnston CW, et al. (2015) An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat Commun* 6:8421.
9. Johnston CW, et al. (2016) Assembly and clustering of natural antibiotics guides target identification. *Nat Chem Biol* 12(4):233–239.
10. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* 7(3):211–224.
11. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: A knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res* 32(Web Server issue):W405–W413.
12. Bachmann BO, Ravel J (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol* 458:181–217.
13. Kim J, Yi GS (2012) PKMiner: A database for exploring type II polyketide synthases. *BMC Microbiol* 12:169.
14. Ziemert N, et al. (2012) The natural product domain seeker NaPDoS: A phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* 7(3):e34064.
15. Conway KR, Boddy CN (2013) ClusterMine360: A database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res* 41(Database issue):D402–D407.
16. Ichikawa N, et al. (2013) DoBISCUIT: A database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 41(Database issue):D408–D414.
17. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH (2009) Automated genome mining for natural products. *BMC Bioinformatics* 10:185.
18. Skinnider MA, et al. (2015) Genomes to natural products prediction informatics for secondary metabolomes (PRISM). *Nucleic Acids Res* 43(20):9645–9662.
19. Weber T, et al. (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43(W1):W237–243.
20. Arnison PG, et al. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* 30(1):108–160.
21. Wieland Brown LC, Acker MG, Clardy J, Walsh CT, Fischbach MA (2009) Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. *Proc Natl Acad Sci USA* 106(8):2549–2553.
22. Medema MH, et al. (2015) Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* 11(9):625–631.
23. Cereto-Massagué A, et al. (2015) Molecular fingerprint similarity search in virtual screening. *Methods* 71:58–63.
24. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* 7:20.
25. Mohr KI, et al. (2015) Pinensins: The first antifungal lantibiotics. *Angew Chem Int Ed Engl* 54(38):11254–11258.
26. Morinaka BI, et al. (2014) Radical S-adenosyl methionine epimerases: Regioselective introduction of diverse D-amino acid patterns into peptide natural products. *Angew Chem Int Ed Engl* 53(32):8503–8507.
27. Breil BT, Ludden PW, Triplett EW (1993) DNA sequence and mutational analysis of genes involved in the production and resistance of the antibiotic peptide trifolitoxin. *J Bacteriol* 175(12):3693–3702.
28. Maksimov MO, Pelczer I, Link AJ (2012) Precursor-centric genome-mining approach for lasso peptide discovery. *Proc Natl Acad Sci USA* 109(38):15223–15228.
29. Letzel AC, Pidot SJ, Hertweck C (2014) Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC Genomics* 15:983.
30. Cox CL, Doroghazi JR, Mitchell DA (2015) The genomic landscape of ribosomal peptides containing thiazole and oxazole heterocycles. *BMC Genomics* 16:778.
31. Cimermancic P, et al. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158(2):412–421.
32. van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP (2013) BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* 41(Web Server issue): W448–W453.
33. Markowitz VM, et al. (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* 42(Database issue):D560–D567.
34. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res* 40(Database issue):D130–D135.
35. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
36. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
37. Eddy SR (2011) Accelerated profile HMM searches. *PLOS Comput Biol* 7(10):e1002195.
38. Magrane M; UniProt Consortium (2011) UniProt Knowledgebase: A hub of integrated protein data. *Database (Oxford)* 2011:bar009.
39. Bailey TL, et al. (2009) MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res* 37(Web Server issue):W202–W208.
40. Steinbeck C, et al. (2003) The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J Chem Inf Comput Sci* 43(2):493–500.
41. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50(5): 742–754.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
43. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37.
44. Prlić A, et al. (2012) BioJava: An open-source framework for bioinformatics in 2012. *Bioinformatics* 28(20):2693–2695.
45. Stajich JE, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10):1611–1618.
46. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
47. Grant CE, Bailey TL, Noble WS (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27(7):1017–1018.
48. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: A python environment for tree exploration. *BMC Bioinformatics* 11:24.