HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2017 August 28.

Published in final edited form as:

Nat Chem Biol. 2017 May ; 13(5): 470–478. doi:10.1038/nchembio.2319.

A new genome-mining tool redefines the lasso peptide biosynthetic landscape

Jonathan I. Tietz[1,†], **Christopher J. Schwalen**[1,†], **Parth S. Patel**[1], **Tucker Maxson**[1], **Patricia M. Blair**[1], **Hua-Chia Tai**[1], **Uzma I. Zakai**[1], and **Douglas A. Mitchell**[1,2,3,*]

[1]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[2]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[3]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

Abstract

Ribosomally synthesized and post-translationally modified peptide (RiPP) natural products are attractive for genome-driven discovery and re-engineering, but limitations in bioinformatic methods and exponentially increasing genomic data make large-scale mining difficult. We report RODEO (Rapid ORF Description and Evaluation Online), which combines hidden Markov model-based analysis, heuristic scoring, and machine learning to identify biosynthetic gene clusters and predict RiPP precursor peptides. We initially focused on lasso peptides, which display intriguing physiochemical properties and bioactivities, but their hypervariability renders them challenging prospects for automated mining. Our approach yielded the most comprehensive mapping of lasso peptide space, revealing >1,300 compounds. We characterized the structures and bioactivities of six lasso peptides, prioritized based on predicted structural novelty, including an unprecedented handcuff-like topology and another with a citrulline modification exceptionally rare among bacteria. These combined insights significantly expand the knowledge of lasso peptides, and more broadly, provide a framework for future genome-mining efforts.

*Correspondence: douglasm@illinois.edu, phone: 1-217-333-1345, fax: 1-217-333-0508.
†contributed equally to this work

URLs

The RODEO source code for local installation, along with the SVM optimization and classification scripts, tutorials for using RODEO, and web tool access are available at: **http://www.ripprodeo.org**.

ACCESSION CODES

Solution-state NMR structure of LP2006 deposited in the RCSB Protein Data Bank (PDB code 5JPL).
The genome of *S. albulus* was deposited in GenBank with the accession number LWBU00000000.

AUTHOR CONTRIBUTIONS

J.I.T. and C.J.S. contributed equally to this work. Experiments were designed by J.I.T., C.J.S. and D.A.M and were performed by J.I.T., C.J.S., T.M., P.M.B., H-C.T. and U.I.Z. J.I.T., P.S.P. and C.J.S. wrote code. The manuscript was written by J.I.T., C.J.S., and D.A.M.

COMPETING FINANCIAL INTEREST

The authors declare no competing financial interest.

Natural products have transformed the practice of medicine, where a major percentage of clinically relevant molecules are NPs, mimics, or derivatives[1]. Despite this, the discovery of novel NPs remains an endeavor largely based on trial and error; the bioassay-guided isolation method historically was fruitful, but abundant, long-known compounds dominate such screens, resulting in the costly and time-consuming "rediscovery problem"[2]. Reconciling the challenges of traditional NP discovery with the opportunities of modern genomics, computational methods represent a promising bridge between genes and molecules[3, 4]. In this way, genomic information allows for NP structural prediction, facilitates prioritization based on predicted novelty, and guides structure elucidation[5, 6].

Biosynthetic gene cluster (BGC) analysis tools that predict structure based on functional domain content and architecture have been developed[3, 4, 6, 7]. These tools work especially well for modular assembly-line complexes, such as polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS), and has enabled their broad-scale profiling[8, 9]. This type of analysis is not as straightforward for other NP classes, such as the ribosomally synthesized and post-translationally modified peptides (RiPPs)[10]. As gene-encoded peptides, RiPP precursors are enzymatically tailored to produce diverse compounds (Fig. 1a). RiPPs also present unique characteristics for genomic identification: (i) their BGCs are relatively small, (ii) RiPP precursors are typically encoded near the modification enzymes, (iii) their sequence can be used to predict novelty and aid structure elucidation[5], and (iv) RiPP precursors have separate sites for enzyme-binding and modification (referred to as the leader and core regions, respectively). The unmodified leader region is proteolytically removed during maturation. This provides a facile route to evolving new NPs since RiPP enzymes can be highly selective for a particular leader sequence but promiscuously process many core sequences[11].

Methods for mining RiPPs include single-input whole-genome analysis (*e.g.* antiSMASH and BAGEL3)[4, 12] and RiPPquest, a mass spectrometry-based method for connecting BGCs to molecules. However, a customizable platform to predict RiPP BGCs across all taxa in a single query has remained an unmet challenge. To develop such a platform, we have initially focused on lasso peptides, a RiPP subfamily. Lasso peptides present additional challenges, for instance, there can be as few as two essential biosynthetic enzymes, both closely related to primary metabolism. The BGCs are thus very small and easily overlooked. Furthermore, nearly all lasso peptide precursors are not recognized as coding sequences and their hypervariability makes their identification nontrivial.

Lasso peptides biosynthesis requires two enzymes: a lasso cyclase (C-protein) homologous to asparagine synthase and a leader peptidase (B-protein) homologous to transglutaminase. Lasso peptides contain a macrocyclic linkage between an Asp or Glu side chain to the N-terminus of the core peptide. The C-terminal tail is threaded through the macrocycle, giving a lariat topology for which the lasso peptides are named (Fig. 1b). Threading is maintained either by disulfide bonds between the macrocycle and tail or by bulky tail residues[13]. These features impart exceptional stability towards proteases and heat as well as a diversity of bioactivities. Lasso peptide BGCs frequently feature transporters and a RiPP recognition element (RRE, E-protein), which binds the leader peptide and directs enzymatic

modification[11]. Interest in these scaffolds is evident from the numerous studies geared toward generating unnatural lasso peptides[14–16].

Herein we report the development of RODEO (Rapid ORF Description and Evaluation Online), an algorithm developed to analyze RiPP BGCs via a combination of profile hidden Markov model (pHMM)-based local genomic analysis and precursor peptide/structure prediction using a combination of heuristic scoring, motif analysis, and machine learning. We have demonstrated the utility of RODEO by surveying and annotating the observable lasso peptide genomic space, revealing >1,400 BGCs, which expands the family by an order of magnitude. Our data revealed unforeseen trends in both precursor sequences and BGC architecture that yielded the most complete picture of lasso peptides to date. In addition to connecting all previously isolated lasso peptides to their respective BGCs, our analysis revealed a large number of conserved lasso peptide families. Furthermore, trends in core peptide sequence and BGC architecture, including co-occurrence of previously-unrecognized tailoring enzymes are described. Lastly, we leveraged this new data set to isolate and characterize several new lasso peptides, which revealed a modification heretofore unseen in any RiPP, as well as a novel lasso topology. Hence, RODEO aids in the generation of high-quality, extensive data sets that here revealed the extent of chemical/genetic diversity in a RiPP class while facilitating the discovery of novel NPs through genomic prioritization.

RESULTS

Identification of lasso peptide biosynthetic gene clusters

To address the challenges associated with lasso peptides, RODEO can rapidly analyze a large number of BGCs by predicting the function for genes flanking query proteins. This is accomplished by retrieving sequences from GenBank followed by analysis with HMMER3. The results are compared against the Pfam database with the data being returned to the user in spreadsheet and HTML format[17–19] (Fig. 1c). For analysis of BGCs encoding proteins not covered by Pfam, RODEO allows usage of additional pHMMs (either curated databases or user-generated). Although RODEO was designed for lasso peptides, it provides a logically expandable framework to target other RiPP and non-RiPP BGCs.

Taking advantage of RODEO's ability to rapidly analyze genes neighboring a query, we compiled a list of all observable lasso peptide BGCs in GenBank (**Online Methods**). We reasoned that a comprehensive evaluation of this data set would provide unprecedented insight into a widespread, but understudied, NP family. Lasso peptide BGCs were defined by the local presence of genes encoding proteins matching the Pfams for the lasso cyclase, leader peptidase, and RRE (Fig. 1b). This procedure identified 1,419 prospective lasso peptide BGCs.

Identification of lasso peptide precursor peptides

The previously mentioned challenges regarding the confident identification of RiPP precursors render traditional methods insufficient. For example, in our final set of lasso peptide precursors described below, ~85% were not recognized as coding sequences in GenBank. Moreover, nucleotide and protein BLAST searches using all known lasso peptides

as the query sequence cumulatively returned <300 hits, of which validation without laborious manual inspection of each BGC was unreliable. Low alignment scores between even closely-related peptides made distinguishing false positives from true lasso precursors impossible in most cases. To confidently predict lasso precursors, RODEO next performed a six-frame translation of the intergenic regions within each of the potential 1,419 BGCs (**Supplementary Results**, Supplementary Fig. 1). The resulting peptides were assessed based on length and essential sequence features and split into predicted leader and core regions. A series of heuristics based on known lasso peptide characteristics[13, 20] were defined to predict precursors from a large pool of false positives (Supplementary Table 1). After optimization of heuristic scoring, we obtained good prediction accuracy, but only for BGCs closely related to known lasso peptides (Fig. 1d–e and Supplementary Fig. 2).

We reasoned that machine learning, particularly support vector machine (SVM) classification, would be more effective in locating precursor peptides from predicted BGCs more distant to known lasso peptides[21]. Although SVM has been previously used to predict NRPS substrate specificity[22], to our knowledge, it has not yet been used to predict NP structures themselves. SVM is well-suited for RiPP discovery due to availability of SVM libraries that perform well with large data sets with numerous variables and the ability of SVM to minimize unimportant features[23]. After calculation of >100 features (Supplementary Table 2), we optimized the SVM classifier using a randomly selected and manually curated training set from the unrefined whole data. Of these, a random sub-population was withheld as a test set to avoid over-fitting (Supplementary Fig. 1). We found that by combining SVM classification with motif (MEME) analysis,[24] along with our original heuristic scoring, prediction accuracy was greatly enhanced as evaluated by recall and precision metrics (Fig. 1d–e, Supplementary Figs. 2–3). This tripartite procedure yielded a high-scoring, well-separated population of 1,315 lasso precursor peptides from >24,000 candidate peptides. The training set was found to display nearly identical scoring distributions upon comparison to the full data set.

The identification of 1,315 high-scoring lasso precursors within 1,419 BGCs further bolstered our confidence in the predictive value of RODEO. Most BGCs contain one lasso precursor peptide, although two contain as many as six (Supplementary Fig. 4). For the cases where a precursor was not identified, it is possible the peptide deviated too far from prediction guidelines, is encoded elsewhere in the genome, or the BGC is defunct.

Comparison of RODEO to existing bioinformatic tools

To compare predictive abilities with existing tools, the 1,419 RODEO-identified lasso peptide-containing genomes were submitted to antiSMASH. Contrary to RODEO, each genome had to be individually queried given that antiSMASH is a single-input whole-genome analysis tool[4]. AntiSMASH identified 87% of the lasso peptide BGCs found by RODEO (Supplementary Table 3). This good agreement was expected given the dual reliance on pHMM analysis in identifying BGCs. AntiSMASH, however, does not predict RiPP precursor peptides but the program does attempt to predict local coding sequences. For 16% of cases, the RODEO-identified lasso precursor gene was predicted as a gene by antiSMASH, albeit along with many others (Supplementary Table 3). A companion

repository, the antiSMASH Database, was recently developed to analyze antiSMASH data compiled from many genomes[25]. This program currently surveys only complete genomes listed in GenBank and thus predicted only 351 lasso peptide BGCs with no predictions for candidate precursor peptides.

Recently, PRISM's functionality has been expanded to analyze prokaryotic genomes for RiPP BGCs, including lasso peptides[26]. RiPP-PRISM, which combines heuristics with pHMM validation, returned a similar estimate to the total number of lasso BGCs as RODEO, although the composition varied significantly (Supplementary Table 3). Lastly, we compared the predictive capabilities with another popular tool, BAGEL3[12]. This program was capable of identifying the majority of a randomly selected subset of lasso peptide BGCs; however, BAGEL3 either misclassified the RRE protein as the precursor peptide or missed the precursor entirely.

Analysis of lasso peptide core sequences

A sequence similarity network (SSN) of the 1,315 lasso precursors was constructed to visualize diversity, distribution, and discovery status (Fig. 2 and Supplementary Data Set 2)[27]. It was immediately apparent that the majority of lasso peptide structural diversity remains unexplored, particularly in the Firmicutes, Cyanobacteria, Euryarchaeota, and Bacteroidetes. Intriguingly, many large lasso peptide families do not have any isolated members. Also, two of the best-known lasso peptides, microcin J25 and lariatin, are in fact unique examples.

As sequence biases have important discovery and engineering implications, we next analyzed the residue composition of the 1,315 precursors. We surmised this would reveal a more universal perspective on sequence trends compared to rules inferred from the limited number of previously-isolated lasso peptides. For instance, it was long believed that lasso peptide biosynthesis required Gly to be the first core peptide residue[20, 28]. Recently, lasso peptides with N-terminal Cys, Ala, and Ser have been reported[29, 30], demonstrating that sequence diversity at this position has been markedly underestimated. Upon inspection of the full set of lasso peptides, the first core position does indeed prefer small residues, with the majority being Gly (62%) or Ala (16%). However, Leu is unexpectedly common (7%) and fittingly, we isolated members of the lagmysin and citrulassin families (Fig. 3) that contain the first examples of N-terminal Leu. Pro is the only residue not represented in the first position, possibly due to restrictions on macrocyclization.

The hallmark lasso peptide macrocycle poses many questions that only an extensive data set is poised to answer. For instance, the conformation-restricting Pro residues of caulosegnin aids in maintaining stability[31], but the extent to which Pro is globally represented in the core region of lasso peptides has remained unknown. We found that Pro is only slightly overrepresented in the macrocycle region of the core, accounting for 6.1% (expected 4.7%, Supplementary Table 4)[32]. In contrast, Gly is significantly overrepresented in the macrocycle (22.8% versus 7.1% expected), suggesting that flexibility is biosynthetically crucial. Aromatics are overrepresented in the tail region, likely due to their role in maintaining a threaded topology (Fig. 3a)[13]. The unexpected prevalence of C-terminal Ser strongly correlates with the co-occurrence of a Ser kinase gene in several lasso peptide

families, including the paeninodins[31] and yet-unexplored mesonodins and aneurinodins (Supplementary Tables 5–6), suggesting an *O*-phosphorylation site. These sequence biases point to roles for particular residues to interact with enzymes, maintain threading, and elicit bioactivity. Knowledge of these trends will facilitate future biosynthetic, engineering, and functional studies on these peptides that remain inaccessible to synthetic chemistry.

Evolutionary insights into lasso peptide biosynthesis

SSNs of the lasso peptide biosynthetic enzymes show strong correlation to those with closely-related precursor peptides (Fig. 4 and Supplementary Figs. 5–7). The similarity of these proteins also strongly correlates with phylum, which in conjunction with %GC comparison of the BGC to the whole genome (Supplementary Fig. 4), suggests that lasso peptide BGCs are primarily transferred vertically.

Akin to other RiPPs, lasso peptide biosynthesis is orchestrated by the binding of the leader peptide by a ~90 residue RRE domain[11]. This facilitates interaction with modification enzymes and largely accounts for the substrate specificity of RiPP biosynthesis. With rare exception (discussed later), the lasso peptide RRE is present as a discrete protein (E; previously, B1) or fused to the leader peptidase (B; previously, B2). We noted from our survey that E-B fusions occur only in a subset of Proteobacteria (and a few scattered additional examples), distinct E and B proteins are present in all phyla and vastly outnumber fused cases (Supplementary Fig. 8). Therefore, we recommend adoption of an E/B nomenclature, as it is more logical evolutionarily and less ambiguous than the older B1/B2 nomenclature.

In addition to inferring evolutionary lineages, analysis of the data set enabled us to evaluate interactions between the precursor and biosynthetic proteins. We reasoned that physically interacting residues may be discernable by analyzing conserved motifs that correlate between the two binding partners. Therefore, paired alignments of lasso precursors and RREs were analyzed for the presence of gapless sequence motifs using MEME (Supplementary Fig. 9). Whereas most RRE motifs did not correlate to any core motifs (and vice versa), this analysis revealed a widely-distributed leader sequence motif (motif L1: $YxxPxLx_3Gx_5Tx$) that showed excellent correlation to three motifs in the RRE. Motif L1 is present in 800 (63%) lasso precursors with the first portion of the motif, YxxP, represented in all lasso peptide-containing phyla (Supplementary Fig. 10). Notably, the strong YxxP correlation was only observed for BGCs harboring distinct E- and B-proteins. Upon surveying RREs from fused E-B proteins, only 1 motif remained significantly correlated to L1.

To augment motif correlations, we further reasoned that interacting residues could be identified via co-evolutionary analysis, given that if one binding partner were to mutate at a critical location, the other would need a compensatory mutation to maintain binding[33, 34]. We thus employed GREMLIN[35] to identify such residues and the results agreed well with previously-reported binding constants for StmE, the RRE for streptomonomicin (Supplementary Fig. 9)[11, 30]. For instance, StmE-D69A bound StmA with 75-fold lower affinity compared to wild-type. Several locations of the leader peptide are clearly co-evolving with the RRE, especially towards regions of the RRE involved in peptide binding

(α3/β3). Upon mapping the co-evolving sites onto a StmE homology model, the residues localized primarily to one face of the protein[36]. No evidence of co-evolution was found between the core region and the RRE, underscoring the functionally bipartite nature of RiPP precursors[10]. These results provide an evolutionary model for leader peptide recognition by RRE-containing proteins, and more broadly, support the hypothesis that variation in RiPP core peptides seeds NP diversification.

Identification of BGCs with unusual architectures

Our survey revealed a large number of lasso peptide BGCs containing unusual domain architectures. Fusions of the leader peptidase and transporter, for instance, are found in 38 BGCs, including various *Ruminococcus* species (Supplementary Fig. 11) as well as *Streptococcus*, *Enterococcus*, *etc*. Other notable findings include fusions of a lasso cyclase and RRE, as well as a leader peptidase-lasso cyclase fusion in *Bifidobacterium reuteri* and *Acidobacteriaceae* bacterium TAA166, respectively. Also noteworthy from *Acidobacteriaceae* is the precursor, which replaces an "invariant" Thr at the penultimate position of the leader with Ile. This first identification of a natural lasso precursor lacking the invariant Thr reconciles previous reports where conservative substitutions still produced microcin J25 and capistruin[15].

Other genes preferentially co-occur with lasso peptide BGCs

In addition to essential biosynthetic genes, other proteins families were found to frequently co-occur with lasso peptide BGCs (Supplementary Table 6). As expected, ABC transporters were common, appearing in 62% of BGCs (Supplementary Fig. 8). Many transferases also frequently co-occur, including kinases (32%), nucleotidyltransferases (26%), glycosyltransferases (14%), and acetyltransferases (6.6%), *etc*. Taken together, we propose there exist undiscovered lasso peptides bearing additional tailoring. Whereas the transferases mentioned above are found in many phyla, there is a clear subset of proteobacteria devoid of these genes. Instead, proteobacteria encode an isopeptidase reported to linearize the lasso macrocycle[37].

Assignment of BGCs to known lasso peptides

Although the first reported lasso peptide, anantin, was discovered 25 years ago[38], the BGC has not yet been reported, despite commercial availability and usage as an atrial natriuretic factor antagonist. In fact, BGCs for 10 known lasso peptides from 6 families remain elusive due to the unavailability of the producer's genome. We sought to connect all known lasso peptides to BGCs by querying our data set with the core sequences of the known lasso peptides. Since ~50% of lasso peptides belong to families (Fig. 3b), we expected those with unknown BGCs could be traced to RODEO-predicted precursor peptides with identical or nearly identical sequences. Indeed, queries using 4–5 contiguous core residues located precursors for every lasso peptide of previously unknown origin. (Supplementary Fig. 12). All lasso peptides with newly assigned BGCs were from Actinobacteria. Consistent with others from this phylum, each encode a single lasso precursor immediately upstream of discrete C, E, and B proteins. By assigning BGCs to all known lasso peptide families, this study underscores the utility of comparative genomics in providing insight into NP biosynthesis even in the absence of a genome for the original producing organism.

Structure-prioritized discovery of lasso peptides

We next applied RODEO's predictive capabilities to discover lasso peptides expected to harbor intriguing structural features. We examined our data set for lasso peptides predicted in the USDA-ARS actinomycete collection (http://nrrl.ncaur.usda.gov/). Instances of widely-distributed families, as well as rare ones, with unique variations in sequence and topology, were selected for further investigation. Thirty-seven strains were chosen, grown on three media, and screened by MALDI-TOF-MS for the presence of peaks consistent with the RODEO predictions (Supplementary Fig. 13). Of these, metabolites from five cultures were isolated for further study and their identity was confirmed by high-resolution, tandem MS (Fig. 5, Supplementary Fig. 14). For citrulassin A (**1**), lagmysin (**2**), and LP2006 (**3**), which all contained unprecedented structural features, extensive NMR analysis was conducted. Chemical shifts of individual residues were assigned by ^{1}H-^{1}H COSY, ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC, and ^{1}H-^{1}H NOESY (Supplementary Table 7, Supplementary Fig. 15). Residue order was assigned by analysis of sequential NH-CαH NOEs in conjunction with MS/MS and the sequence of the precursor's coding gene. The site of the isopeptide linkage was confirmed by the presence of strong NOE signals from the N-terminal residue to the Asp/Glu side chain. Lagmysin, LP2006, and citrulassin A each showed a wide range of NH chemical shift values (~7.25–9.50); unthreaded lasso peptides show a narrower distribution of NH chemical shift values[39]. We also observed NOE correlations between the ring and tail, indicating their proximity (Supplementary Fig. 15). Further, TOCSY data for anantin B_1 (**4**), citrulassin A, lagmysin, and LP2006 indicated that multiple backbone amides were resistant to deuterium exchange, implying solvent occlusion (Supplementary Fig. 15). All lasso peptides were thermally stable, showing no unthreading by HPLC after extended heating, and were either resistant or impervious to carboxypeptidase Y digestion (Supplementary Fig. 14)[46, 47]. These data strongly support a threaded lasso topology. Marfey-type analysis also confirmed all amino acids to be in the L-configuration (Supplementary Fig. 14).

The precursor from *Nocardiopsis alba* was particularly interesting because it was predicted to bear a novel topology (Fig. 5). This peptide contains two Cys, that presumably would form a disulfide; however, both Cys were predicted to reside on the tail whereas in all other known cases, disulfides link the tail to the macrocycle. This resulting "handcuff" topology of LP2006 was confirmed by solution-state NMR. We propose this rare lasso peptide topology be referred to as class IV.

Moomysin (**5**), isolated from *Streptomyces cattleya*, has no known structural analogues but topologically belongs to class II (Fig. 5). Lagmysin, another class II lasso peptide from *Streptomyes* sp. NRRL S-118, belongs to one of the largest families (Fig. 3) and contains an unprecedented N-terminal Leu. Currently, this is the largest residue with demonstrated biosynthetic compatibility. Additionally, we characterized by HR-MS/MS anantin B_1 and a congener lacking the C-terminal Phe (anantin B_2, **6**) from a previously unknown producer, *Streptomyces* sp. NRRL S-146 (Fig. 5 and Supplementary Fig. 14). The anantins comprise a lasso peptide family with 8 currently identifiable members.

The citrulassins also contain N-terminal Leu and form a family with 55 members (Fig. 3). During screening, the masses of seven of nine citrulassins were 1 Da heavier than predicted (Supplementary Tables 8–9). This caught our attention, as additional tailoring of lasso peptides is rare[31, 40, 41]. One previously unsequenced organism, *Streptomyces albulus* NRRL B-3066, also produced citrulassin, as gleaned from C-terminal tail MS/MS fragmentation. This compound purified most readily of all citrulassins; therefore, we sequenced the producer's genome and deemed the compound citrulassin A, as mentioned above. NMR, HR-MS/MS, and quantitative amino acid analysis of citrulassin A confirmed that its genetically encoded Arg9 was modified to citrulline (Fig. 5 and Supplementary Fig. 14), accounting for the mass difference. Intriguingly, Arg9 is invariant among the 55 citrulassins, although codon usage varies. To our knowledge, this is the first known instance of citrulline in any RiPP, but also, only the second known instance of peptidylarginine deiminase (PAD) activity in the entire bacterial domain[42]. In an attempt to locate the responsible PAD, we performed conservation analysis of the genes flanking the citrulassin BGCs. Not only were these flanking genes highly variable, no obvious candidate for a PAD was found. To determine if the responsible PAD was locally encoded, we generated a fosmid library of *S. albulus* to heterologously express the citrulassin BGC. Two fosmids were expressed in *Streptomyces lividans*, together covering ~20 kb upstream and downstream of the BGC (Fig. 6). MS analysis revealed that expression of either fosmid led to the production of Arg-containing *des*-citrulassin (**7**). These data suggest the responsible PAD is distally encoded.

Antibacterial activity of lasso peptides

We tested all of the lasso peptides isolated herein for growth-suppressive activity against a panel of bacteria (Supplementary Table 10). Citrulassin A, moomysin, and lagmysin lacked growth suppressive activity towards all organisms tested. In contrast, LP2006 was active against *Enterococcus faecium*, *Bacillus subtilis*, *Bacillus anthracis*, and *Mycobacterium smegmatis* while anantin B_1/B_2 showed weak activity against *E. coli*. Interestingly, the anantin B_2 was also moderately active against *B. subtilis*.

DISCUSSION

Genome-mining tools have been closing the gap between the large number of predicted BGCs and the significantly smaller number of known NPs. Despite the development of methods for identifying and predicting structures for NRPs and PKs, similar methods for RiPPs have been hampered by the homology of some RiPP biosynthetic proteins to those in primary metabolism and the difficulty in locating precursor peptides[3]. However, the gene-encoded nature of RiPP precursors provides an opportunity: a confident precursor prediction near a reasonable BGC is strong evidence for a valid RiPP. Thus, a robust platform for BGC identification and precursor scoring provides predictive ability not possible in other NP classes. Although manual identification of RiPP precursors is common, it is performed on a case-by-case basis which is tedious and inefficient. Methods that facilitate predictions for large numbers of BGCs will be increasingly necessary given that genome sequencing significantly outpaces the rate of functional annotation. These issues are amplified for lasso peptides, which are sequence hypervariable. Machine learning is an ideal tool to address this

challenge: it is scalable for large data sets, and in an unbiased fashion, can determine which scoring parameters to prioritize. This results in a powerful, adaptive scoring metric that only requires a comparatively small but diverse training set. Despite these strengths, machine learning remains underutilized in NP discovery.

We herein describe the development and utility of RODEO, an algorithm for BGC identification. RODEO provides a graphical and tabular output of BGCs and peptides that can be later analyzed to reveal statistical correlations and examine phylogenetic relationships. The customizability of the program in taking user-supplied configuration files and pHMMs renders RODEO extensible to any gene neighborhood analysis, although currently, the precursor-scoring portion of RODEO is optimized exclusively for lasso peptides. With customization to other RiPP classes, or by turning off the precursor scoring function, RODEO complements existing tools focused on individual whole genomes, such as BAGEL3[12] and antiSMASH[4], as well as general gene context tools such as the Gene Ortholog Neighborhood analysis option within JGI-IMG and the Prokaryotic Sequence homology Analysis Tool, PSAT[43, 44]. Beyond RiPPs, we have already used RODEO to analyze polyketide BGCs by constructing pHMMs for enzymes responsible for plecomacrolide-specific backbone tailoring[45].

RODEO here was used to identify 1,315 lasso precursors within 1,419 BGCs from a basis set of only 28 characterized BGCs. This is notable in light of the sequence hypervariability seen in lasso peptides relative to other RiPPs. This study revealed that ~50% of lasso peptides exist in structural families. In addition to locating BGCs for all genetically unassigned lasso peptides, we isolated six new examples, four of which were chosen on the basis of predicted structural novelty, and subsequently characterized their structures and antibacterial activities. Of these, LP2006 was found to adopt an unprecedented topology and displayed activity against several pathogens. Lasso peptides are reported to display diverse bioactivities, such as receptor antagonists and enzyme inhibitors; however, it is common for lasso peptides to be reported without an associated activity. This suggests some will display more esoteric activities or influence signal transduction. This last possibility is consistent with sviceucin's inhibition of *Enterococcus faecalis* quorum sensing[46]. One such lasso peptide described herein that lacks bacterial growth-suppressive activity, citrulassin A, belongs to a 55-membered family. Citrulassin A is exemplified by the conversion of an Arg to citrulline by a yet-unidentified PAD. MS-based screening suggests that additional citrulassins contain citrulline. This simple modification is quite unique: not only was citrulline previously unknown in RiPPs, but only one example of PAD activity—a virulence factor—is known in the entire bacterial domain[42, 47, 48].

The RODEO-enabled genome-mining methods described here has allowed for broad-scale investigations of a NP family only possible with a large, high-confidence dataset. Our analysis indicates that lasso peptide BGCs are widely distributed among bacteria, with *Streptomyces* harboring the plurality. We identified rare cases of lasso peptide BGCs in unprecedented phyla, although most appear to be inherited vertically. Despite their widespread nature, fewer than 50 lasso peptides have been isolated. This work provides a roadmap for the future discovery of additional lasso peptides, novel RiPP classes, and allows for a rapid means to curate and interpret emerging BGCs.

ONLINE METHODS

General methods

Unless otherwise specified, chemicals were purchased from Sigma, Fisher Scientific, or GoldBio. DNA sequencing was performed by the Roy J. Carver Biotechnology Center (Univ. of Illinois at Urbana-Champaign). Restriction enzymes were purchased from New England Biolabs (NEB), Platinum Taq HiFi was purchased from Thermo Scientific, and dNTPs were purchased from NEB. Oligonucleotide primers were synthesized by Integrated DNA Technologies (IDT).

Lasso peptide cluster dataset assembly

GI numbers for essential biosynthetic genes (lasso cyclase, lasso protease, and RRE) of all known lasso peptides were collected. The set of lasso cyclases was then used as a BLAST-P query against the nr database in GenBank, returning the top 1,000 hits for each with default settings. Hits were ranked by maximum similarity to any query sequence, and the top local genomic region (6 genes in either direction of cyclase) of top hits were sequentially analyzed with RODEO until no new lasso peptide-like BGCs were returned. Because our initial query set contained only Actinobacteria and Proteobacteria, we repeated the above using as queries representative lasso peptide cyclases, obtained during the first search, from the Bacteroidetes, Cyanobacteria, Firmicutes, Fusobacteria, and Verucomicrobia, and Archaea. The RODEO data sets were then merged for analysis, giving a set of 4,672 potential BGCs with 25,506 potential precursor peptides. Refinement and scoring (below) of this data yielded a final dataset of 1,419 lasso peptide BGCs and 1,315 lasso peptide precursors.

Precursor peptide scoring optimization

From the initial set of prospective lasso peptide BGCs, the RODEO output of a randomly-chosen sample (~25%; whole training set) were inspected, and gene clusters were classified qualitatively as lasso, non-lasso, or indeterminate. Peptides in each cluster were also classified as valid likely precursors, non-precursors, or indeterminate. This formed an unbiased training set for scoring optimization. In addition to the scoring features output by RODEO directly, an additional set of features were calculated in Excel for each peptide (Supplementary Tables 1 and 2). The linear combination score of each peptide was calculated as the dot product S of weight vector A and feature vector X,

$$S = \boldsymbol{A} \cdot \boldsymbol{X} = \sum_{i=1}^{n} A_i X_i$$

where for each of i features n, A_i is the feature point weight and X_i is the absence, 0, or presence, 1, of the feature.

For analysis and optimization, 80% of the manually classified set was randomly selected (“training set”), with the other 20% retained as a validation set to avoid over-fitting. Feature weighting optimization was done by plotting comparative score histograms of classified precursor peptides and stochastically adjusting individual weights to minimize overlap. For final analysis, additional points were given to any peptide classified as valid by the SVM

classifier (see below) or showing MEME/FIMO-identified sequence motifs (see below), resulting in two minimally overlapping populations (Supplementary Fig. 2).

Sequence motif-based unsupervised learning of precursor peptides

Precursor peptide sequences from the training set (see above) were analyzed for the top 16 sequence motifs with zero or one occurrences per sequence using MEME (parameters: -nmotif 16 -mod zoops)[24]. All output peptides from RODEO (n = 25,506) were scanned for these motifs using FIMO, and both the presence of motifs and the resulting scores were passed to final linear-combination scoring and used as features for SVM. Motifs were only considered for linear-combination scoring if their score from FIMO was greater than 1, which we found to increase accuracy. The sequence motifs used by FIMO are included with the Supplementary Data.

Support vector machine (SVM) supervised learning classifier

Machine learning model training and classification was performed using the scikit-learn library for Python. The manually classified training set (see above) was used as model data. Five-fold cross-validation (*i.e.* randomly witholding one-fifth of the training data as a testing data set) was used to optimize the prediction method and avoid over-fitting. Kernel modes (linear, polynomial, rbf) and parameters (C, gamma, polynomial degree, class weight) were systematically varied to optimize prediction accuracy, assessed as the product of *precision* [precision = true positives/(true positives + false positives)] and *recall* [recall = true positives/(true positives + false negatives)]. Optimal parameters used an rbf kernel with C = 25, gamma = 2.75E-06, and 'auto' class weight. This gave 97.9% precision and 92.8% recall, which we estimated would likely result in ~30 false positives and ~100 false negatives when applied to the entire data set. Thus, we subjected the entire 25,506-member RODEO-derived peptide set to classification. **Code availability:** The Python scripts used for SVM features, optimization, and classification are available with the source code (http://www.ripprodeo.org/) and are also listed in Supplementary Table 2.

Precursor peptide set determination

We utilized a combination of SVM classification, motif analysis, linear-combination scoring, and SSN analysis to determine valid lasso precursor peptides. To minimize the false positive and negative rate, we compared SVM classifications to linear-combination scores and inspected the highest-scoring SVM negatives and lowest-scoring SVM positives, resulting in reclassification of several peptides. We also manually inspected peptides that clustered abnormally on an SSN (alignment score cutoff = 1E-6) of the precursor peptides (*e.g.* singleton SVM negatives clustering with SVM). SVM positives and negatives with high or low scores, respectively, that clustered with valid or invalid precursor peptides were assumed to be most likely correct. *Leader/core alignment verification*: RODEO predicts the leader/core cut site on the basis of a conserved Thr found in the penultimate leader position and a putative cyclization site (Asp/Glu) in the core. We manually refined the RODEO-generated leader/core split sites based on conserved leader residues (YxxP and Gx_5T motifs); although correct in 94% of cases, 75 required manual adjustment due to the presence of multiple possible Thr and Asp/Glu residues in the precursor.

Gene cluster classification

Gene clusters were classified on the basis of cluster architecture (identifiable E, B, or C proteins) as well as the presence of valid precursor peptides. Any cluster with a valid lasso precursor was classified as a lasso peptide BGC. Those containing only one identifiable E, B, or C protein and no precursor peptide were discarded. The remaining clusters were classified by analysis of an SSN of the lasso cyclases. Those that contained two or more identifiable lasso biosynthetic genes and showed high similarity to valid BGCs, as judged by network connectivity, were retained. In this way, the initial data set of 4,672 potential gene clusters was refined to a set of 1,419 predicted lasso peptide BGCs, 1,165 of which (82%) contained a RODEO-identified lasso peptide precursor.

RRE/precursor peptide co-evolution analysis

A paired alignment of precursor peptide sequences and RREs (PF05402 in lasso BGCs) (n = 846) was generated; RRE sequences were aligned using MUSCLE on default settings in MEGA7[49]. The aligned precursor peptides and RREs were concatenated. *(1) Residue/residue co-evolution analysis*. The paired alignment was analyzed using the GREMLIN web server (http://gremlin.bakerlab.org/)[34] on default settings, filtered at 75% gap removal and 25% coverage thresholds. No multiple sequence alignment (MSA) generation was performed. The precursor peptide and RRE (GenBank: WP_040271204.1) from the streptomonomicin BGC were used as a reference. *(2) Sequence motif analysis with MEME/FIMO*. Sets of paired precursor peptides/RREs were analyzed using MEME 4.11.1, which detects ungapped sequence motifs[24]. Precursor sequences were trimmed to include only the leader region and analyzed to find the 8 most significant motifs with zero or one occurrences per sequence (parameters: -nmotif 8 -mod zoops). All precursors and RREs were then analyzed for the presence of each discovered motif using FIMO on default settings[24]. Motif correlation was calculated as the number of co-occurrences of each of two motifs divided by the number of total occurrences of each. Sequence motifs were mapped onto the StmE homology model generated as previously described.[11] (3) *Homology modeling*. HHpred (http://toolkit.tuebingen.mpg.de)[50] was used to generate sequence alignments for the RRE from the gene cluster of streptomonomicin (GenBank: WP_040271204.1). LynD (PDB ID: 4V1T) was used to generate a model using the HHpred modeling toolkit (aligned residues 8–84). Inter-protein GREMLIN-output probability scores of residue-residue co-evolution were used to map regions of predicted evolutionary significance

Analysis of GC content

The %GC of nucleotide sequences for a representative set of lasso cyclases (n = 1,085) were calculated and compared to NCBI Taxonomy records (organism %GC content). Lasso cyclase %GC was plotted against average organism genome %GC. Linear regression was calculated to draw a best-fit line and calculate absolute residual regression for each point. The top 5% of outliers (n = 54) were highlighted.

Phylogenetic tree generation

For visual clarity, we chose to represent only a representative subset of lasso cyclases (n = 185 of 1,419). From the phyla with the most lasso BGCs (Actinobacteria, Proteobacteria,

Firmicutes), ~10% of sequences were randomly selected. From all others, which had many fewer lasso BGCs, approximately half were selected in order ensure qualitative representation of all phyla. As diverse outgroups, non-lasso asparagine synthases (PF00733) were chosen from *E. coli* (GenBank: KJW30448.1; tree root), *B. subtilis* (GenBank: AID00505.1), and *Streptomyces rimosus* (GenBank: KEF03673.1). Protein sequences were first aligned in MEGA7[49] using MUSCLE on default settings. A tree was generated from the alignment using the Maximum Likelihood method based on the JTT matrix-based model in MEGA7 on otherwise default settings[49] and visualized with iTOL (http://itol.embl.de/)[51] and Adobe Illustrator. For suggestions on how to incorporate RODEO-generated data into tree-visualization programs like iTOL, see: http://www.ripprodeo.org/advanced.html

Sequence similarity network (SSN) generation

SSNs were generated using EFI-EST (http://efi.igb.illinois.edu/efi-est/)[27]. Initial networks were generated using alignment score cutoffs of 1E-1 (precursor peptides), 1E-5 (RREs), 1E-7 (leader proteases), or 1E-25 (lasso cyclases) as clustering thresholds. Networks were visualized using the Organic layout within Cytoscape v. 3.2.0[52] and Adobe Illustrator (Supplementary Data Set 2). Nodes were annotated using RODEO output (via Microsoft Excel) and the NCBI Taxonomy database[53]. For suggestions on how to incorporate RODEO-generated data into Cytoscape, see: http://www.ripprodeo.org/advanced.html

Lasso data set Circos visualization

Alignment scores for precursor peptides were obtained during SSN generation. Individual precursor peptides were annotated from a RODEO-generated and curated spreadsheet for the presence of co-occurring Pfams, structural novelty, phylum, and structural family. Lasso peptide, BGCs, and and their genomic contexts were visualized using Circos (Supplementary Data Set 3)[54]. For suggestions on how to import RODEO data for Circos visualization, see: http://www.ripprodeo.org/advanced.html

Sequence logo generation

Sequence logos were generated using WebLogo 2.8.2 (http://weblogo.berkeley.edu/)[55]. Core sequences of some families (*i.e.* sphingonodin, caulonodin, sphingopyxin and paeninodin) were first aligned using MUSCLE v.3.8.31[56] prior to sequence logo generation.

Microbiological methods

All actinomyces strains were grown in either ATCC medium 172 [ATCC172] (per L: glucose, 10 g; soluble starch, 20 g; yeast extract, 5 g; N-Z-Amine type A, 5 g; $CaCO_3$, 1 g; pH 7.3), ISP Medium no. 4 [ISP4] (per L: soluble starch, 10 g; K_2HPO_4, 1 g; $MgSO_4{\cdot}7H_2O$, 1 g; NaCl, 1 g; $(NH_4)_2SO_4$, 2 g; $CaCO_3$, 2 g; $FeSO_4$, 1 mg; $MnCl_2$, 1 mg; $ZnSO_4$, 1 mg), or AltMS (per L: mannitol, 10 g; soy flour, 10 g; malt extract, 10 g). All solid media included 1.5% (w/v) agar. *Neisseria meningitidis* was grown on gonococcal broth [GCB] (per L: proteose peptone #3, 15 g; K_2HPO_4, 4 g; KH_2PO_4, 1 g; NaCl, 1 g) with sterile filtered bicarbonate (100x: $NaHCO_3$, 0.42 g to 10 mL H_2O) and sterile filtered Kellogg's Supplements I (100x to 500 mL: D-glucose, 200 g; L-glutamine, 5 g; thiamine pyrophosphate, 0.01 g) and II (1000x to 50 mL: $Fe(NO_3)_3{\cdot}9H_2O$, 0.025 g).

Escherichia coli strains for DNA manipulation were DH5α, grown in LB medium (10 g tryptone, 10 g NaCl, 5 g yeast extract to 1 L of deionized water; Fisher Scientific) with 50 μg/mL kanamycin. A Gibson assembly master mix (NEB) was used to clone the anantin gene cluster from *Streptomyces* sp. S-146 into a pET28a backbone. Qiagen QIAprep Spin Miniprep Kit was used to purify plasmids. Primer walk sequencing was used to generate overlapping sequencing reads connecting the biosynthetic machinery for anantin.

Screening of Actinomycetes for lasso peptide production

Prioritized strains were grown on each ATCC Media 172, ISP Media 4, and AltMS on 6-well plates with 5 mL of agar media per well. Cultures were incubated at 30 °C for 7 d prior to whole-cell mass spectrometry. A portion of cell mass (pinhead-sized) including aerial hyphae was picked from the plate and placed in 2 μL of sat. 50% aq. MeCN solution of α-cyano-4-hydroxycinnamic acid (CHCA) with 0.1% trifluoroacetic acid (v/v) on a ground steel MALDI target. Additional aliquots of matrix (2 μL) were spotted and allowed to dry on the cell mass three times to extract metabolites onto the plate. Samples were analyzed using a Bruker UltrafleXtreme MALDI-TOF MS using manufacturer methods for reflector positive mode.

Isolation of high molecular weight DNA

Streptomyces albulus str. NRRL B-3066 was grown in liquid ATCC172 media supplemented with 0.8% glycine at 30 °C with agitation until a dense culture had developed (~5 d). A 30 mL portion of the culture was pelleted (4,000 × *g*, 10 min), and the supernatant was removed by aspiration. The cells were resuspended in 2 mL of TE25S buffer (25 mM Tris, 25 mM EDTA disodium salt, 30% sucrose, pH 8) and large clumps of cells were broken up by physical disruption with a glass rod. Lysozyme was added to a final concentration of 10 mg/mL and the suspension was incubated at 37 °C with agitation. After 1 h, 6 mL of modified lysis buffer (569 mM 4-aminosalicylic acid sodium salt, 69 mM sodium dodecyl sulfate, 100 mM Tris, pH 8) and 20 μL of proteinase K (NEB, 30 U/mg at 20 mg/mL) were added. The suspension was then incubated at 75 °C for 30 min, with the suspension being mixed once by gentle inversion several times after 15 min. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1, v/v) was then added and mixed by gentle inversion several times. The mixture was centrifuged at 10,000 × *g* at 4 °C for 10 min, and the aqueous phase was transferred to a new container using a wide bore pipet (or pipet tips with the ends cut off to provide a wider opening). To the aqueous phase, an equal volume of chloroform:isoamyl alcohol (24:1, v/v) was added and mixed by gentle inversion several times. The mixture was again centrifuged at 10,000 × *g* at 4 °C for 10 min, and the aqueous phase was transferred to a new container using a wide bore pipet. Another round of chloroform:isoamyl alcohol extraction was performed if solid material was still present in the aqueous phase. The aqueous phase was then transferred to a 50 mL conical tube, to which 0.1 volume equivalents of 5 M NaCl was added. The tube was then filled to the top with isopropyl alcohol and the mixture was gently inverted several times to mix. A small glass hook fashioned from a melted Pasteur pipet was used to collect the precipitated DNA. The DNA was washed twice with 75% aq. EtOH followed by a single wash with 100% EtOH. The DNA was then allowed to air dry on the glass hook overnight.

Fosmid library generation

High molecular weight DNA was resuspended in 600 μL TE buffer (10 mM Tris, 1 mM EDTA, pH 8), 330 μL 10x CutSmart buffer (NEB), and 2310 μL DNase-free water. Gentle pipetting with a wide bore pipet and heating at 65 °C were used to help dissolve the DNA. After the majority of the DNA was dissolved, 30 μL of RNase A was added and the solution was mixed by gentle inversion. The DNA solution was split into 14 fractions and Sau3AI (NEB) was added to each in a 2-fold serial dilution, starting with 2.5 μL of Sau3AI into 400 μL of DNA solution. Solutions were mixed by gentle pipetting with a wide bore pipet. Digest reactions were incubated at 37 °C for 1 h and were then inactivated by incubation at 75 °C for 30 min. Fractions were analyzed by field-inversion gel electrophoresis, and fractions containing approximately 40 kb pieces of DNA were utilized in cosmid construction. These fractions were dephosphorylated by the addition of 20 μL of shrimp alkaline phosphatase and incubation at 37 °C for 1 h. DNA fragments were isolated by phenol:chloroform:isoamyl alcohol extraction and air dried for 3 h.

The previously reported vector pJK050 was used for fosmid construction[57]. pJK050 was doubly digested with NheI-HF and BamHI-HF (NEB) and dephosphorylated with shrimp alkaline phosphatase. Cut vector was purified with a GeneJET PCR Purification Kit (Thermo Scientific). Approximately 1 μg of cut vector, 2 μL of 10x T4 ligase buffer (NEB), and water (to give 20 μL total volume) was used to resuspend the digested high molecular weight DNA. A portion (5 μL) was removed as a control and 1 μL of high concentration T4 ligase (NEB, 2,000,000 U/mL) was added to the remainder. The ligation reaction was incubated at 4 °C overnight and used without further modification.

A culture of *E. coli* WM4489 was started in LB and grown at 37 °C overnight. The culture was then used to inoculate a new 10 mL culture of LB supplemented with 10 mM maltose. This culture was grown at 37 °C with shaking until it reached an optical density of 0.8 to 1.0. The culture was pelleted at 4,000 × *g* for 10 min, and the supernatant was removed by aspiration. The pellet was resuspended in 2 mL of sterile 10 mM $MgCl_2$ and chilled on ice until needed. Ligated DNA was then packaged into phage and transfected into *E. coli* WM4489 with MaxPlax Lambda Packaging Extracts (Epicentre) following the manufacturer's instructions. The transfection reaction (2 mL) was pelleted at 17,000 × *g* for 5 min and 1.4 mL of the supernatant was discarded. The remaining supernatant was used to resuspend the pellet and the suspension was plated onto 5 LB agar plates (100 μL/plate) containing 12.5 μg/mL chloramphenicol. The plates were incubated at 37 °C overnight, resulting in thousands of colonies per plate. Colonies were scraped from the plates utilizing 1 mL of LB per plate, pooled together, and mixed with 3 mL of sterile 50% glycerol in water to give a library stock solution. The solution was plated on fresh LB agar plates containing 12.5 μg/mL chloramphenicol at an appropriate dilution to provide individual colonies. Colonies were screened by PCR for the presence of approximately 650 base sequences before and after the citrulassin A gene cluster. Positive hits were verified by an additional round of PCR screening and were then grown in LB containing 12.5 μg/mL chloramphenicol and 20 mM rhamnose at 37 °C with agitation overnight. Two hits were found to contain the citrulassin A gene cluster and were named p1F3 and p3H4. Cosmid DNA was isolated with a GeneJET Plasmid Miniprep Kit (Thermo Scientific). The start and end of the genomic

insert in each cosmid was determined via sequencing with primers designed to anneal up- and downstream of the BamHI or NheI cut sites.

Heterologous expression of citrulassin A gene cluster

Cosmid DNA was recombined in vitro with pAE4 using Gateway BP Clonase II (Thermo Scientific) following the manufacturer's instructions to insert the functions necessary for transfer and integration into *Streptomyces lividans*. The resulting fosmids were transformed into the conjugation strain *E. coli* WM6026 by electroporation, with 20 μg/mL of sterile diaminopimelic acid (DAP) added during the recovery step (*E. coli* WM6026 is a DAP auxotroph). Transformants were plated on LB agar containing 12.5 μg/mL chloramphenicol, 50 μg/mL apramycin, and 20 μg/mL DAP. Cultures of *E. coli* WM6026 containing the fosmids were started in LB supplemented with 12.5 μg/mL chloramphenicol, 50 μg/mL apramycin, and 20 μg/mL DAP and grown overnight at 37 °C with agitation. The overnight cultures were then used to inoculate fresh 5 mL cultures with the same supplements, which were grown to an optical density of 0.4 to 0.6. The cultures were pelleted by centrifugation at 4,000 × *g* for 10 min, and the supernatants were discarded. The pellets were washed twice with sterile 2xYT media (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl), resuspended in 1 mL of 2xYT containing 20 μg/mL DAP, and chilled on ice until use. A frozen suspension of *S. lividans* 66 spores was heat shocked at 56 °C for 15 min straight from the freezer. After 15 min, 500 μL of 2xYT was added and mixed by pipetting. Conjugation reactions were set up with 200 μL of the spore suspension and either 4 μL (50:1), 40 μL (5:1), or 200 μL (1:1) of a fosmid containing *E. coli* WM6026 solution. The suspensions were briefly centrifuged to settle the cells together at the bottom of the tubes and were incubated at room temperature for 10 min. The cells were then resuspended by pipet and the entire solutions were plated on individual mannitol-soy plates (20 g/L mannitol, 20 g/L soy flour, 16 g/L agar) supplemented with 10 mM $MgCl_2$. The plates were incubated at 30 °C for 16 h, and then the plates were flooded with 2 mL of a 1 mg/mL aqueous apramycin solution. The plates were dried while open in a biosafety cabinet and then incubated at 30 °C until colonies appeared. The strains were named *S. lividans* 1F3 and *S. lividans* 3H4, corresponding to the respective cosmids used to construct them.

Colonies of *S. lividans* 1F3 and *S. lividans* 3H4 were used to inoculate ATCC172 liquid media containing 12.5 μg/mL chloramphenicol and 50 μg/mL apramycin. The cultures were grown at 30 °C with agitation until a dense culture was obtained. These cultures were plated on 5 ATCC172 agar plates each (antibiotic free) to generate a lawn and these plates were grown at 30 °C for 10 d. The plates were frozen, thawed, and squeezed to collect the liquid from the agar. The liquid was filtered through Whatman paper by vacuum filtration and then applied to a 0.5 g Hypersep C18 solid phase extraction column (Thermo Scientific) that had been equilibrated with 20 mL of MeCN, 20 mL of 50% aq. MeCN, and 30 mL of H_2O. The column was washed with 20 mL of water and eluted with 5 mL each of 10%, 20%, 40%, 60%, and 100% aq. MeCN. Fractions were analyzed by MALDI MS with CHCA as a matrix. *Des*-citrulassin A was detected in the 40%, 60%, and 100% fractions. The 60% fraction was subsequently utilized for HRMS and MS/MS analysis.

The primers used for heterologous expression are given in Supplementary Table 11.

Isolation of anantin B_1 and B_2

Streptomyces sp. S-146 was grown on ATCC 172 with 1.5% agar on 15 cm sterilized petri dishes ($n = 48$) and incubated at 30 °C for 7 d. Cells were harvested by scraping cell mass from the plates with razors and collectively extracted with 600 mL MeOH with shaking at room temperature overnight. The methanolic extract was vacuum-filtered and dried under reduced pressure onto 5 g of Celite 545 adsorbent.

Spent agar was frozen at −20 °C overnight and then allowed to thaw. The agar was squeezed and filtered through cheesecloth to harvest the aqueous portion, which was vacuum-filtered with Whatman filter paper and loaded onto a HyperSep C18 10 g SPE column (Thermo Scientific) pre-equilibrated with H_2O. The loaded column was washed with 200 mL of H_2O, and bound material was eluted with successive 50 mL portions of 5%, 10%, 25%, 50% and 75% aq. MeCN, followed by 100 mL of MeCN. Fractions were analyzed via MALDI-TOF MS. Fractions containing anantin B_1 and B_2 were pooled with the methanolic extract and dried under reduced pressure onto 5 g of Celite 545 adsorbent.

Anantin adsorbed onto Celite was purified by medium pressure liquid chromatography (MPLC) using a Teledyne Isco Combiflash EZprep equipped with a RediSep Rf C18 cartridge (130 g media, 60 Å pore size, 40–63 μm particle size, 230–400 mesh) using a mobile phase of 10 mM aq. NH_4HCO_3/MeCN at 85 mL min^{-1} over a gradient from 20–60% MeCN (10 column volumes), then 60–95% MeCN (5 column volumes). Fractions were analyzed via MALDI-TOF and dried under reduced pressure.

Anantin B_1 and B_2 were dissolved in 7 mL of water each and filtered with a 0.22 μm syringe filter (BD Technologies) prior to injection. HPLC was performed using a Teledyne Isco Combiflash EZprep equipped with a RediSep Prep C18Aq column (100 Å pore size, 150 × 20 mm, 5 μM particle size) with a mobile phase of 10 mM aq. NH_4HCO_3/MeCN (gradient from 10 to 95% MeCN) at 19 mL min^{-1}. Fractions containing either anantin B_1 or B_2 were dried under reduced pressure, resuspended in H_2O, flash-frozen, and lyophilized overnight to dryness.

Preparative HPLC was performed with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 10 mm, 5 μm particle size). Anantin B_1 and B_2 were dissolved separately in 1.5 mL of water, subjected to centrifugation at 17,000 × *g* to remove precipitate, and injected with a mobile phase of water/MeCN + 0.1% formic acid (v/v) with a gradient from 10–95% over 30 min. Elution was monitored at 220 nm and appropriate fractions were dried under reduced pressure, dissolved in water (2 mL), flash-frozen, and lyophilized overnight to dryness. Anantin B_1 and B_2 were collected as colorless powders (1.8 mg and 2.1 mg, respectively). Yields for anantin B_1 and B_2 were 75 μg/15 cm dish and 88 μg/15 cm dish respectively.

Analytical HPLC to assess purity was performed using a Betasil C18 column (100 Å pore size, 250 × 4.6 mm, 5 μM particle size) (Thermo Scientific) with a mobile phase of H_2O/MeCN + 0.1% (v/v) formic acid (FA) gradient from 10% to 95% aq. MeCN over 30 minutes at 1 mL min^{-1}. For anantin B_2, a mobile phase of 10 mM aq. NH_4HCO_3/MeCN gradient from 5 to 25% MeCN over 30 minutes, then ramp to 95% MeCN over 5 minutes and held at

95% MeCN for 10 minutes at 1 mL min^{-1}. Both anantin B_1 and anantin B_2 were detected eluting at 220 nm.

Isolation of LP2006

Nocardiopsis alba NRRL B-24146 was grown on ATCC 172 and 1.5% agar in sterilized aluminum cake pans (12-¼ inches by 8-¼ inches; $n = 12$; ~300 mL media ea.) and incubated at 30 °C for 21 d. Cells were harvested by scraping cell mass from the plates with a razor blade and collectively extracted with 100 mL MeOH with shaking at rt overnight. Following filtration, the remaining cell mass was re-extracted with 100 mL of *n*-BuOH shaking at rt overnight. The MeOH and *n*-BuOH extract was vacuum-filtered and dried under reduced pressure onto 10 g of Celite 545 adsorbent.

LP2006 adsorbed on Celite was purified by MPLC using a Teledyne Isco Combiflash EZprep equipped with a RediSep Rf C18 cartridge (86 g media, 60 Å pore size, 40–63 μm particle size, 230–400 mesh) using a mobile phase of 10 mM aq. NH_4HCO_3/MeCN at 60 mL min^{-1} over a gradient from 10–95% MeCN (12 column volumes), and held at 95% MeCN (5 column volumes). Fractions were analyzed via MALDI-TOF MS. Fractions containing LP2006 were combined and the solvent was evaporated.

HPLC was performed using a Teledyne Isco Combiflash EZprep equipped with a RediSep Prep C18Aq column (100 Å pore size, 150 × 20 mm, 5 μM particle size) with a mobile phase of 10 mM aq. NH_4HCO_3/MeCN (gradient from 25 to 95% MeCN) at 19 mL min^{-1}. Fractions found by MALDI-TOF to contain LP2006 were combined and evaporated to dryness.

Preparative chromatography on a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 10 mm, 5 μm particle size) was performed with a mobile phase of 10 mM aq. NH_4HCO_3/MeCN (gradient from 20–40% MeCN for 20 min, then 40–95% MeCN for 5 min) at 4 mL min^{-1}. This yielded 1.9 mg of LP2006, which was dissolved in water, flash-frozen, and lyophilized to dryness with a final yield of 158 μg/pan.

The purity of LP2006 was checked by HPLC with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 4.6 mm, 5 μm particle size, 100 Å pore size). A mobile phase of 10 mM aq. NH_4HCO_3/MeCN gradient from 5–25% MeCN over 15 minutes, then ramped to 95% MeCN over 5 min and held at 95% MeCN for 10 min at a flow rate of 1 mL min^{-1}. Elution of LP2006 was detected at 220 nm.

Isolation of moomysin

Streptomyces cattleya NRRL 8057 was grown on AltMS medium with 1.5% agar on sterilized aluminum cakepans (12-¼ inches by 8-¼ inches; $n = 12$; ~300 mL media ea.) and incubated at 30 °C for 14 d. Cells were harvested by scraping cell mass from the plates with razors and cell material was extracted with 200 mL of 50% aq. MeCN overnight with shaking at room temperature. The cell mass was filtered with Whatman filter paper and the filter cake was re-extracted 3 additional times. Extracts were combined and dried under

reduced pressure onto ~5 g Celite 545 adsorbent. Agar extracts did not contain detectable quantities of moomysin by MALDI-TOF MS, and were not purified or analyzed further.

Adsorbed moomysin was purified by MPLC using a Teledyne Isco Combiflash EZprep equipped with a RediSep Rf C18 cartridge (130 g media, 60 Å pore size, 40–63 μm particle size, 230–400 mesh) using a mobile phase of 10 mM NH_4HCO_3/MeCN at 85 mL min^{-1} over a gradient from 10–20% MeCN (5 column volumes), 20–50% MeCN (5 column volumes), then 50–100% MeCN (2.5 column volumes) then holding at 100% MeCN (5 column volumes). Fractions were analyzed via MALDI-TOF MS, and fractions containing moomysin were pooled and dried under reduced pressure and stored at −20 °C.

Samples for HPLC were resuspended in 7 mL of water, subjected to centrifugation at 17,000 × *g* to remove precipitate and filtered with a 0.22 μm syringe filter (BD Technologies) prior to injection. Moomysin was injected onto a Teledyne Isco Combiflash EZprep equipped with a RediSep Prep C18Aq column (100 Å pore size, 150 × 20 mm, 5 μM particle size) at 19 mL min^{-1} with a mobile phase of 10 mM NH_4HCO_3/MeCN gradient starting at 10% MeCN (10 column volumes), 10 to 45% MeCN (7 column volumes) then 45 to 100% MeCN (5 column volumes) and held at 100% MeCN for 5 column volumes. Elution of moomysin was monitored at 214 nm, and the corresponding fractions were dried under reduced pressure and stored at −20 °C.

Preparative chromatography was performed with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 10 mm, 5 μm particle size). Moomysin was dissolved in 1.5 mL of H_2O, subjected to centrifugation at 17,000 × *g* to remove precipitate and injected with a mobile phase of water/MeCN + 0.1% formic acid (v/v) gradient from 10–95% over 30 min with a flow rate of 4 mL min^{-1}. Elution was monitored at 220 nm and fractions containing moomysin were dried under reduced pressure, resuspended in 2 mL water, flash-frozen, and lyophilized overnight to dryness. Moomysin was collected as a colorless powder (0.8 mg) after lyophilization, with an overall yield of 133 μg/pan.

The purity of moomysin was verified by HPLC with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 4.6 mm, 5 μm particle size, 100 Å pore size). A mobile phase of 10 mM aq. NH_4HCO_3/MeCN gradient from 5–15% MeCN over 25 minutes, then ramped to 95% MeCN for 5 minutes and held at 95% for 10 minutes at a flow rate of 1 mL min^{-1}. Elution of moomysin was monitored at 220 nm.

Isolation of citrulassin A

Streptomyces albulus NRRL B-3066 was grown on ATCC172 medium with 1.5% agar on sterilized, aluminum half sheet cake pans ($n = 4$; 750 mL per pan) and incubated at 30 °C for 10 d. The agar was frozen, thawed, and squeezed (filtering through cheesecloth) to collect the bulk aqueous extract. The solid agar was then extracted with 600 mL of 50% methanol in water for 2 h. The methanol was removed under vacuum and the remaining aqueous extraction was combined with the liquid from the agar squeeze. The combined aqueous extract was vacuum filtered through Whatman filter paper to remove residual solid material. The extract was loaded onto a HyperSep C18 10 g SPE column (Thermo Scientific) pre-

equilibrated with 100 mL MeCN, 100 mL 50% aq. MeCN, and 200 mL water prior to loading. The column was washed with an additional 200 mL of water. Product was eluted with successive 50 mL portions of 10%, 20%, 30%, 40%, and 50% aq. MeCN, followed by 100 mL of MeCN. Fractions were analyzed via MALDI-TOF MS (Supplementary Methods: screening Actinomycetes for lasso peptide production). Fractions containing citrulassin A were pooled and dried under reduced pressure onto ~15 g Celite 545 adsorbent.

Adsorbed citrulassin A was purified by MPLC using a Teledyne Isco Combiflash EZprep equipped with a RediSep Rf C18 cartridge (130 g media, 60 Å pore size, 40–63 μm particle size, 230–400 mesh) using a mobile phase of 10 mM aq. NH_4HCO_3/MeCN at 85 mL min^{-1} over a gradient from 10–100% MeCN (25 column volumes). Fractions were analyzed via MALDI-TOF MS, pooled, and dried under reduced pressure. Partially purified material was stored at −20 °C between purification steps.

The partially purified material was resuspended in 5 mL water and was injected onto a Teledyne Isco Combiflash EZprep equipped with a RediSep Prep C18Aq column (100 Å pore size, 150 × 20 mm, 5 μM particle size). The product was purified with a mobile phase of 10 mM aq. NH_4HCO_3/MeCN at a flow rate of 19 mL min^{-1} utilizing a method of: 5% MeCN (isocratic, 5 min), 5–40% MeCN (gradient, 20 min), 40–100% MeCN (gradient, 5 min), 100% MeCN (isocratic, 5 min). Elution of citrulassin A was monitored at 214 nm. Fractions containing the product were dried under reduced pressure and stored at −20 °C.

Final preparative chromatography was performed with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 4.6 mm, 5 μm particle size, 100 Å pore size). The partially purified product was dissolved in 6 mL of water and injected over 6 consecutive runs. A mobile phase of 10 mM aq. NH_4HCO_3/MeCN was utilized at a flow rate of 1 mL min^{-1} with a method of: 20% MeCN (isocratic, 5 min), 20%–30% MeCN (gradient, 35 min), 30%–95% MeCN (gradient, 5 min), 100% MeCN (isocratic, 3 min). Elution was monitored at 220 nm and fractions containing citrulassin A were dried under reduced pressure, resuspended in 5 mL water, flash frozen, and lyophilized overnight to yield a white, cotton-like solid (8 mg, ~2 mg/half sheet cake pan).

The purity of isolated citrulassin A was checked by HPLC with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 4.6 mm, 5 μm particle size, 100 Å pore size). A mobile phase of 10 mM aq. NH_4HCO_3/MeCN was utilized at a flow rate of 1 mL min^{-1}. Citrulassin A was injected as 10 μL of a 1 mM solution with a method of: 15% MeCN (isocratic, 5 min), 15–27% MeCN (gradient, 20 min). Elution was monitored at 220 nm.

Isolation of lagmysin

Streptomyces sp. S-118 was grown on ISP Medium 4 (**Online Methods**: microbiological methods) with 1.5% agar on 15 cm sterilized petri dishes ($n = 24$) and incubated at 30 °C for 10 d. Cells were harvested by scraping cell mass from the plates with razors and collectively extracted with 600 mL MeOH with shaking at room temperature overnight. The cell mass was filtered with Whatman filter paper and re-extracted with 600 mL BuOH with shaking at

room temperature overnight before filtering again. Cell extracts were combined and dried under reduced pressure onto ~5 g Celite 545 adsorbent.

Spent agar was frozen at −20 °C overnight and then allowed to thaw. The agar was squeezed and filtered through cheesecloth to harvest the aqueous portion, then vacuum filtered with Whatman filter paper and loaded onto a HyperSep C18 10 g SPE column (Thermo Scientific) pre-equilibrated with 100 mL MeCN, 100 mL 50% aq. MeCN, and 200 mL water prior to loading. The column was then washed with an additional 200 mL of water, then eluted with successive 50 mL portions of 5%, 10%, 25%, 50%, 75% aq. MeCN, followed by 100 mL MeCN. Fractions were analyzed via MALDI-TOF MS. Fractions containing lagmysin were pooled and dried under reduced pressure onto ~5 g Celite 545 adsorbent and combined with the adsorbed cell extract.

Adsorbed lagmysin was chromatographed by MPLC using a Teledyne Isco Combiflash EZprep equipped with a RediSep Rf C18 cartridge (130 g media, 60 Å pore size, 40–63 μm particle size, 230–400 mesh) using a mobile phase of 10 mM NH_4HCO_3/MeCN at 85 mL min^{-1} over a gradient from 20–60% MeCN (10 column volumes), then 60–95% MeCN (5 column volumes) then holding at 95% MeCN (5 column volumes). Fractions were analyzed via MALDI-TOF MS, pooled and dried under reduced pressure and stored at −20 °C.

To prepare samples for preparatory HPLC, dried lagmysin was resuspended in 7 mL of water, subjected to centrifugation at 17,000 × *g* to remove precipitate, and filtered with a 0.22 μm syringe filter (BD Technologies) prior to injection. Lagmysin was injected onto a Teledyne Isco Combiflash EZprep equipped with a RediSep Prep C18Aq column (100 Å pore size, 150 × 20 mm, 5 μM particle size) utilizing a flow rate of 19 mL min^{-1} with a mobile phase of 10 mM NH_4HCO_3/MeCN gradient from 10 to 50% MeCN (15 column volumes) then 50 to 100% MeCN (5 column volumes) and held for an additional 5 column volumes. Elution of lagmysin was monitored at 214 nm, and fractions containing the desired material were dried under reduced pressure and stored at −20 °C.

Final preparative chromatography was performed with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 (Thermo Scientific) reverse phase column (250 × 10 mm, 5 μm particle size). Lagmysin was dissolved in 1.5 mL of water, subjected to centrifugation at 17,000 × *g* to remove precipitate and injected with a mobile phase of water/MeCN + 0.1% formic acid (v/v) gradient from 20–95% over 30 min with a flow rate of 4 mL min^{-1}. Elution was monitored at 220 nm and fractions containing lagmysin were dried under reduced pressure, resuspended in 2 mL water, flash frozen and lyophilized overnight to dryness. Lagmysin was collected as a colorless powder (6.5 mg) after lyophilization, with an overall yield of 270 μg/15 cm plate.

Analytical HPLC to assess purity of lagmysin was performed with a Perkin Elmer Flexar HPLC equipped with a Betasil C18 column (100 Å pore size, 250 × 4.6 mm, 5 μM particle size) (Thermo Scientific) with a mobile phase of H_2O/MeCN + 0.1% (v/v) formic acid (FA) gradient from 10% to 95% aq. MeCN over 30 minutes and held at 95% MeCN for 10 minutes at 1 mL min^{-1}. Elution was monitored at 220 nm.

Broth microdilution antimicrobial assay

A broth microdilution assay was used to determine the minimum inhibitory concentration (MIC) against a panel of bacteria. *E. coli* MC4100, *Pseudomonas aeruginosa* PA01, *Klebsiella pneumoniae* ATCC 27736, *Acinetobacter baumannii* ATCC 19606, *Enterococcus faecium* U503 (VRE), *Staphylococcus aureus* USA300 (MRSA), *Bacillus subtilis* 168, *Bacillus anthracis* Sterne, *Listeria monocytogenes* serotype 4b str. F2365, and *Mycobacterium smegmatis* NRRL B-14616 were grown in Mueller-Hinton broth (BD). *Neisseria meningitidis* 8013 was grown in gonococcal broth (GBC) with Kellogg's supplements. Bacteria were grown overnight in 5 mL of medium at 37 °C to stationary phase. Cultures were diluted 50-fold into 5 mL of fresh media, allowed to grow back to mid-exponential phase (OD_{600} = 0.500), and diluted to OD_{600} = 0.01. Lasso peptide stocks (10 mM in DMSO; or 1 mM in 0.2 μm filtered H_2O for citrulassin A) were serially diluted (2-fold) in broth in a 96-well microplate. An equal volume of bacterial culture in broth was added to each well (final vol. 100 μL/well) to give a final range of 100 μM to 49 nM compound. Plates were sealed with Parafilm M and incubated at 37 °C with shaking. The MIC is the concentration that resulted in no visible growth after 18 h, or 24 h for *M. smegmatis*. All measurements were made with at least 2 replicates. These data are shown in Supplementary Table 10.

Sequencing of *Streptomyces albulus* str. NRRL B-3066 genome

The genome of *S. albulus* was sequenced in a manner identical to previous publication.[30] The draft genome was deposited in GenBank with the accession number LWBU00000000.

Quantitative amino acid analysis of citrulassin A

Purified citrulassin A was submitted to the Texas A&M Protein Chemistry Laboratory (http://tamupcl.com/) for quantitative amino acid analysis in triplicate using a total of 75 μg.

Chiral amino acid analysis

Microwave assisted vapor phase acid hydrolysis of peptides was performed in a CEM Discover 3000 Microwave System (CEM). Peptides were hydrolyzed in a sealed chamber with 6 N deuterium chloride (DCl) at 150 °C for 30 min under anaerobic conditions. Hydrolysates were derivatized with Marfey's reagent (1-fluoro-2,4-dinitrophenyl-5-L-alanine amide, FDAA) (Thermo-Fisher Scientific) to enhance separation of L-/D-amino acids. Hydrolysates were reconstituted in 25 μL of 0.5 M NaHCO3, combined with 20 μL of 1 mg/mL FDAA in MeCN, and incubated at 60 °C for 3 h. Derivatized amino acids from each peptide sample were analyzed using a LC-MS/MS-MRM system consisting of an Advance Ultra-High Pressure Liquid Chromatography (UHPLC) (Bruker) system coupled to an EVOQ Elite Triple-Quadrupole MS (Bruker) with a Kinetex 2.6 μm Phenyl-Hexyl 100 Å LC column (100 × 2.1 mm) (Phenomenex). Gradient elution was performed with binary solvents (Solvent A: 25 mM ammonium formate, Solvent B: MeOH) at 300 μL/min. The gradient was 5% B for 2 min, 5–15% B for 5 min, 15–60% B for 5 min, 60% B for 3 min, 60–100% B for 3 min, 100% B for 3 min, 100–5% B for 1 min, 5% B for 2 min. Data Review (Bruker) was used to analyze the MRM data.

High resolution mass spectrometry

Lasso peptides were dissolved in 80% aq. MeCN with 1% AcOH and subjected to centrifugation (17,000 × *g*, 5 min). Samples were infused onto a ThermoFisher Orbitrap Fusion ESI-MS using an Advion TriVersa NanoMate. The MS was calibrated weekly using calibration mixture, following manufacturer instructions, and tuned daily with Pierce LTQ Velos ESI Positive Ion Calibration Solution (ThermoFisher). Spectra were collected in profile mode with a resolution of 100,000. Ions were selected for fragmentation in the Ultra-High-Field Orbitrap Mass Analyzer using an isolation width of 5 *m/z*, a normalized collision energy of 35, an activation q value of 0.4, and an activation time of 30 ms. Data analysis was performed using Thermo Xcalibur software. All MS data are shown in Supplementary Figure 14.

NMR spectroscopy

HPLC-purified and lyophilized samples (ca. 1–5 mg) were dissolved in 600 μL of CD_3OD (99.96 atom % D; Sigma-Aldrich) or CD_3OH (99.5 atom % D; Cambridge Isotope Labs). Spectra were obtained with either a Bruker AVANCE 900 MHz narrow bore spectrometer equipped with an inverse 5 mm TCI cryogenic probe with z-axis pulsed field gradient (pfg) capability or on an Agilent VNMRS 750 MHz narrow bore magnet spectrometer equipped with a 5 mm triple resonance (^1H-^{13}C-^{15}N) triaxial gradient probe and pulse-shaping capabilities. Samples were held at 298 K during acquisition. Standard Bruker or Varian pulse sequences were used for each of the following experiments: ^1H, ^1H with solvent suppression using excitation sculpting (Bruker: zgesgp), ^1H-^1H TOCSY (80 ms mixing time; Bruker: dipsi2esgpph), ^1H-^{13}C HSQC (multiplicity-edited; Bruker: hsqcedetgpsisp2), and ^1H-^1H NOESY (400 ms mixing time; Bruker: noesyesgpph; Varian: dpfgse_noesy). 2D spectra in CD_3OH employed solvent suppression as per the listed pulse sequences. Spectra were recorded with Bruker TopSpin 1.3 or VNMRJ 3.2A software, and data was processed using MestReNova 8.1.1. Chemical shifts (δ, ppm) were referenced internally to the solvent peak. All NMR data are shown in Supplementary Figure 15.

Three-dimensional solution structure calculation

Solution structures were calculated by simulated annealing using distance and angle restraints within XPLOR-NIH v 2.36[58]. Standard XPLOR-NIH potentials for bond angles, improper angles, van der Waals, and favored/allowed Ramachandran regions were used. The lasso-forming isopeptide linkage was specified using a manual patch of the "protein-1.0.top" XPLOR-NIH file (Supplementary Note 2); the disulfide linkage was created using a pre-installed patch (DISL). Distance restraints were derived from peak area integrations in the NOESY data. The program nmrPipe was used for processing of NOESY data and conversion from Bruker to UCSF format (Sparky), and Sparky was used for peak picking, volume integration, and creation of the distance restraint table for use in XPLOR. Two hundred structures were calculated, of which the 20 of lowest energy were retained.

Thermal unthreading assay

Lasso peptide stock solutions were diluted to 40 μM in a final volume of 50 μL, heated at 95 °C for 4 h, and cooled to rt. Samples were then analyzed by liquid chromatography

electrospray ionization tandem mass spectrometry (LC/ESI-Q/TOF MS) using a Synapt ESI quadrupole TOF Mass Spectrometry System (Waters) equipped with an Acquity Ultra Performance Liquid Chromatography (UPLC) system (Waters). Lasso peptides were analyzed from 2 to 98% aq. MeCN (with 0.1% formic acid) over 24 min at a flow rate of 130 μL min^{-1}. Total ion chromatograms were used to monitor retention times via Waters MassLynx MS Software.

Enzymatic digestion assay

Lasso peptide stock solutions were diluted to 20 μM in 50 mM MES (pH 6.7) and 1 mM $CaCl_2$ with 1 U carboxypeptidase Y. Lasso digestions were carried out overnight at 22 °C before being desalted and dissolved in 75% MeCN (with 1% acetic acid) for analysis by high resolution mass spectrometry as described above.

Databases

The solution-state NMR structure of LP2006 was deposited in the RCSB Protein Data Bank (PDB code 5JPL). The genome of *S. albulus* was deposited in GenBank with the accession number LWBU00000000.

Data availability

All data generated or analyzed during this study are included in this published article (and its supplementary information) or are available from the corresponding author on reasonable request. The source code for the RODEO program and extended documentation of the algorithm are available at: http://www.ripprodeo.org/.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Lingyang Zhu (Univ. Illinois Urbana-Champaign) and Ben Ramirez (Univ. Illinois Chicago) for NMR assistance. We acknowledge Courtney Cox, Kisurb Choe (Univ. Illinois Urbana-Champaign), and Nicholas Tietz for valuable computational and programming input. This work was supported in part by a NIH Director's New Innovator Award Program (DP2 OD008463 to D.A.M.), the David and Lucile Packard Fellowship for Science and Engineering (to D.A.M.), the Robert C. and Carolyn J. Springborn Endowment (to J.I.T.), and ACS Division of Medicinal Chemistry Predoctoral Fellowships (to P.M.B. and J.I.T.). The Bruker UltrafleXtreme MALDI TOF/TOF mass spectrometer was purchased in part with a grant from the National Center for Research Resources, National Institutes of Health (S10 RR027109 A). The 900 MHz NMR spectrometer was purchased with funds provided by GM068944.

REFERENCES

1. Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. 2016; 79:629–661. [PubMed: 26852623]
2. Winter JM, Behnken S, Hertweck C. Genomics-inspired discovery of natural products. Curr. Opin. Chem. Biol. 2011; 15:22–31. [PubMed: 21111667]
3. Medema MH, et al. Minimum Information about a Biosynthetic Gene cluster. Nat. Chem. Biol. 2015; 11:625–31. [PubMed: 26284661]
4. Weber T, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015; 43:W237–W243. [PubMed: 25948579]

5. Tietz JI, Mitchell DA. Using genomics for natural product structure elucidation. Curr. Top. Med. Chem. 2015; 16:1645–1694.
6. Stachelhaus T, Mootz HD, Marahiel MA. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem. Biol. 1999; 6:493–505. [PubMed: 10421756]
7. Skinnider MA, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 2015; 43:9645–9662. [PubMed: 26442528]
8. Cimermancic P, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014; 158:412–421. [PubMed: 25036635]
9. Doroghazi JR, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat. Chem. Biol. 2014; 10:963–968. [PubMed: 25262415]
10. Arnison PG, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat. Prod. Rep. 2013; 30:108–160. [PubMed: 23165928]
11. Burkhart BJ, Hudson GA, Dunbar KL, Mitchell DA. A prevalent peptide-binding domain guides ribosomal natural product biosynthesis. Nat. Chem. Biol. 2015; 11:564–570. [PubMed: 26167873]
12. van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. 2013; 41:W448–W453. [PubMed: 23677608]
13. Hegemann JD, Zimmermann M, Xie X, Marahiel MA. Lasso peptides: an intriguing class of bacterial natural products. Acc. Chem. Res. 2015; 48:1909–19. [PubMed: 26079760]
14. Al Toma RS, et al. Site-directed and global incorporation of orthogonal and isostructural noncanonical amino acids into the ribosomal lasso peptide capistruin. Chembiochem. 2015; 16:503–509. [PubMed: 25504932]
15. Pan SJ, Rajniak J, Maksimov MO, Link AJ. The role of a conserved threonine residue in the leader peptide of lasso peptide precursors. Chem. Commun. 2012; 48:1880–1882.
16. Zong C, Maksimov MO, Link AJ. Construction of lasso peptide fusion proteins. ACS Chem. Biol. 2016; 11:61–68. [PubMed: 26492187]
17. Kans, J. Entrez Direct: E-utilities on the UNIX Command Line. National Center for Biotechnology Information; 2010.
18. Finn RD, et al. HMMER web server: 2015 update. Nucleic Acids Res. 2015; 43:W30–W38. [PubMed: 25943547]
19. Finn RD, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–D230. [PubMed: 24288371]
20. Maksimov MO, Pelczer I, Link AJ. Precursor-centric genome-mining approach for lasso peptide discovery. Proc. Natl. Acad. Sci. U.S.A. 2012; 109:15223–15228. [PubMed: 22949633]
21. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 2015; 16:321–332. [PubMed: 25948244]
22. Rottig M, et al. NRPSpredictor2-a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 2011; 39:W362–W367. [PubMed: 21558170]
23. Pedregosa F, et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011; 12:2825–2830.
24. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43:W39–W49. [PubMed: 25953851]
25. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Research. 2016
26. Skinnider MA, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. Proceedings of the National Academy of Sciences. 2016; 113:E6343–E6351.
27. Gerlt JA, et al. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): a web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta. 2015; 1854:1019–1037. [PubMed: 25900361]

28. Maksimov MO, Pan SJ, Link AJ. Lasso peptides: structure, function, biosynthesis, and engineering. Nat. Prod. Rep. 2012; 29:996–1006. [PubMed: 22833149]
29. Zimmermann M, Hegemann JD, Xie XL, Marahiel MA. Characterization of caulonodin lasso peptides revealed unprecedented N-terminal residues and a precursor motif essential for peptide maturation. Chem. Sci. 2014; 5:4032–4043.
30. Metelev M, et al. Structure, bioactivity, and resistance mechanism of streptomonomicin, an unusual lasso peptide from an understudied halophilic actinomycete. Chem. Biol. 2015; 22:241–250. [PubMed: 25601074]
31. Hegemann JD, et al. The ring residue proline 8 is crucial for the thermal stability of the lasso peptide caulosegnin II. Mol. Biosyst. 2016; 12:1106–1109. [PubMed: 26863937]
32. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204–D212. [PubMed: 25348405]
33. Hopf TA, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife. 2014; 3
34. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3:e02030. [PubMed: 24842992]
35. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins. 2011; 79:1061–1078. [PubMed: 21268112]
36. Koehnke J, et al. Structural analysis of leader peptide binding enables leader-free cyanobactin processing. Nat. Chem. Biol. 2015; 11:558–563. [PubMed: 26098679]
37. Maksimov MO, Link AJ. Discovery and characterization of an isopeptidase that linearizes lasso peptides. J. Am. Chem. Soc. 2013; 135:12038–12047. [PubMed: 23862624]
38. Weber W, Fischli W, Hochuli E, Kupfer E, Weibel EK. Anantin--a peptide antagonist of the atrial natriuretic factor (ANF). I. Producing organism, fermentation, isolation and biological activity. J. Antibiot. 1991; 44:164–171. [PubMed: 1849131]
39. Xie X, Marahiel MA. NMR as an effective tool for the structure determination of lasso peptides. Chembiochem. 2012; 13:621–5. [PubMed: 22278977]
40. Ogawa T, et al. RES-701-2, −3 and −4, novel and selective endothelin type B receptor antagonists produced by *Streptomyces* sp. I. Taxonomy of producing strains, fermentation, isolation, and biochemical properties. J. Antibiot. 1995; 48:1213–1220. [PubMed: 8557559]
41. Gavrish E, et al. Lassomycin, a ribosomally synthesized cyclic peptide, kills *Mycobacterium tuberculosis* by targeting the ATP-dependent protease ClpC1P1P2. Chem. Biol. 2014; 21:509–518. [PubMed: 24684906]
42. Goulas T, et al. Structure and mechanism of a bacterial host-protein citrullinating virulence factor, *Porphyromonas gingivalis* peptidylarginine deiminase. Sci. Rep. 2015; 5:11969. [PubMed: 26132828]
43. Markowitz VM, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014; 42:D560–D567. [PubMed: 24165883]
44. Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ. PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC Bioinformatics. 2008; 9:170. [PubMed: 18366802]
45. Molloy EM, Tietz JI, Blair PM, Mitchell DA. Biological characterization of the hygrobafilomycin antibiotic JBIR-100 and bioinformatic insights into the hygrolide family of natural products. Bioorg. Med. Chem. 2016; 24:6276–6290. [PubMed: 27234886]
46. Li Y, et al. Characterization of Sviceucin from Streptomyces Provides Insight into Enzyme Exchangeability and Disulfide Bond Formation in Lasso Peptides. ACS Chem. Biol. 2015; 10:2641–2649. [PubMed: 26343290]
47. McGraw WT, Potempa J, Farley D, Travis J. Purification, characterization, and sequence analysis of a potential virulence factor from *Porphyromonas gingivalis*, peptidylarginine deiminase. Infect. Immun. 1999; 67:3248–56. [PubMed: 10377098]
48. Gabarrini G, et al. The peptidylarginine deiminase gene is a conserved feature of *Porphyromonas gingivalis*. Sci. Rep. 2015; 5:13936. [PubMed: 26403779]

ONLINE METHODS REFERENCES

49. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016
50. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009; (77 Suppl 9):128–32. [PubMed: 19626712]
51. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 2011; 39:W475–W478. [PubMed: 21470960]
52. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. Curr Protoc Bioinformatics. 2014; 47(813):1–24. [PubMed: 25199789]
53. Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. Methods Mol. Biol. 2011; 696:291–303. [PubMed: 21063955]
54. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–45. [PubMed: 19541911]
55. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004; 14:1188–90. [PubMed: 15173120]
56. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]
57. Eliot AC, et al. Cloning, expression, and biochemical characterization of Streptomyces rubellomurinus genes required for biosynthesis of antimalarial compound FR900098. Chem Biol. 2008; 15:765–70. [PubMed: 18721747]
58. Schwieters CD, Kuszewski JJ, Clore GM. Using Xplor-NIH for NMR molecular structure determination. Prog. Nucl. Magn. Reson. Spectrosc. 2006; 48:47–62.

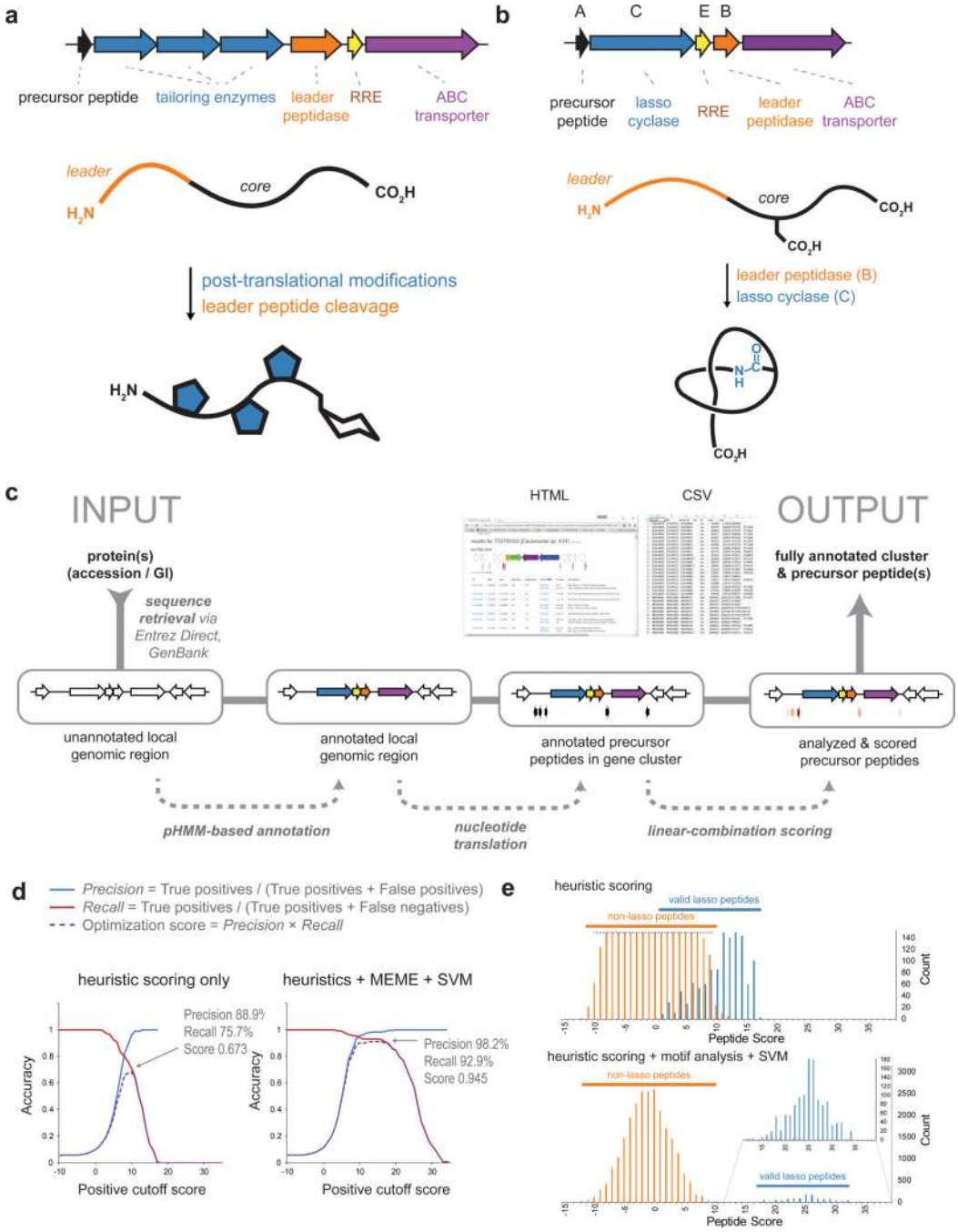

Figure 1. Ribosomal natural product (RiPP) biosynthesis and overview of RODEO

(**a**) General overview of RiPP biosynthesis. (**b**) Overview of lasso peptide biosynthesis by leader peptidase, lasso cyclase, and RRE. (**c**) RODEO workflow and output. (**d**) Comparison of scoring accuracy on a randomly selected training set using heuristic scoring only or scoring with combined motif analysis (MEME) and machine learning (SVM). Sensitivity is represented by recall; specificity is represented by precision. (**e**) Comparative scoring distribution on the final peptide dataset using either heuristics only or scoring with MEME and SVM integrated. Scoring distributions were practically indistinguishable between

training and full data sets (Supplementary Fig. 2). Currently, RODEO is optimized to score potential lasso peptides.

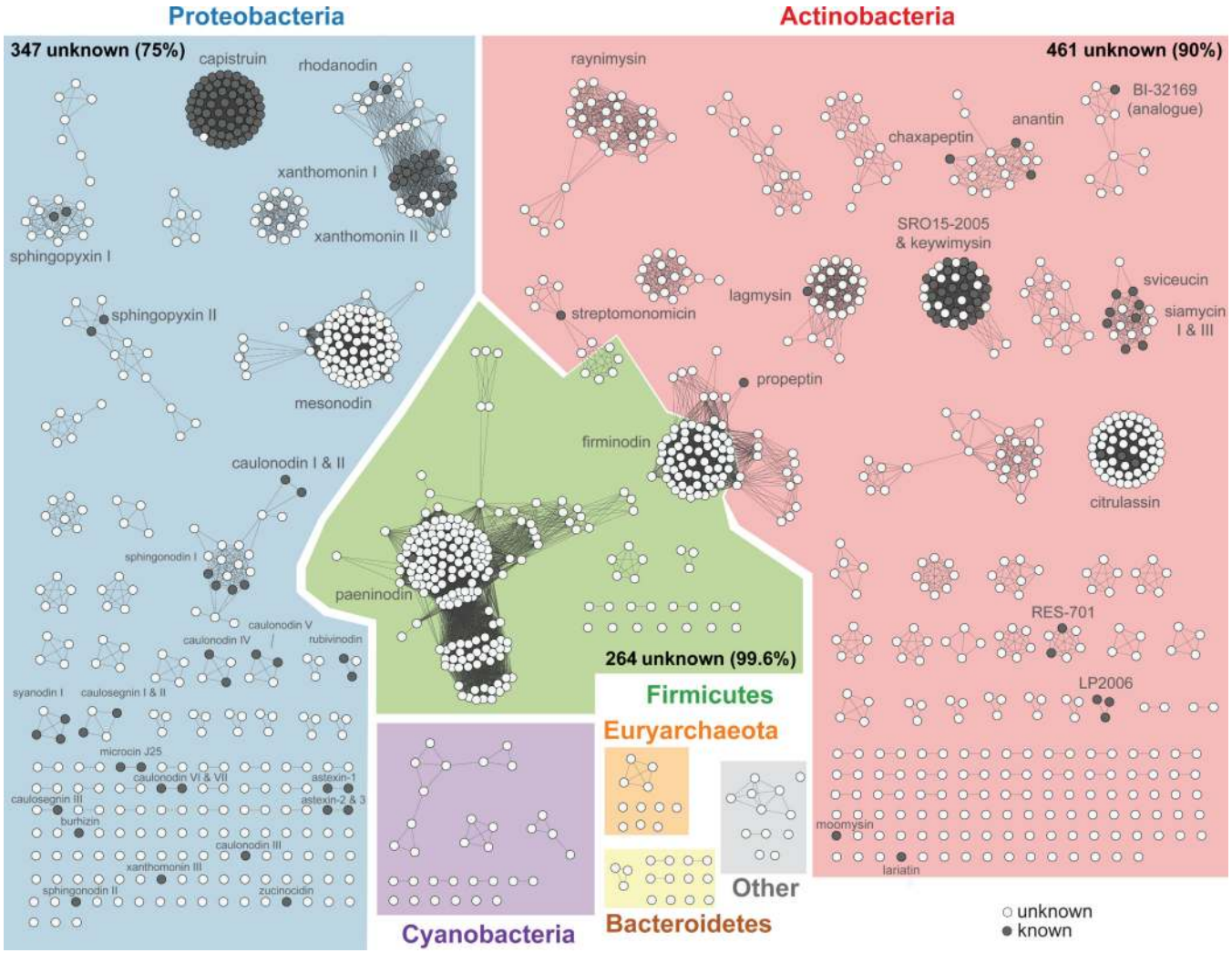

Figure 2. Phylogenetic map of all identified lasso peptides

A SSN of precursor peptides is shown with an E-value cutoff of 10^{-8}. Background shading indicates phylum. Node shading indicates if the NP has been isolated or detected in culture (including this study). Co-occurrence of conserved genes in the local genomic region for the peptides above are indicated in Supplementary Data Sets 1, 3.

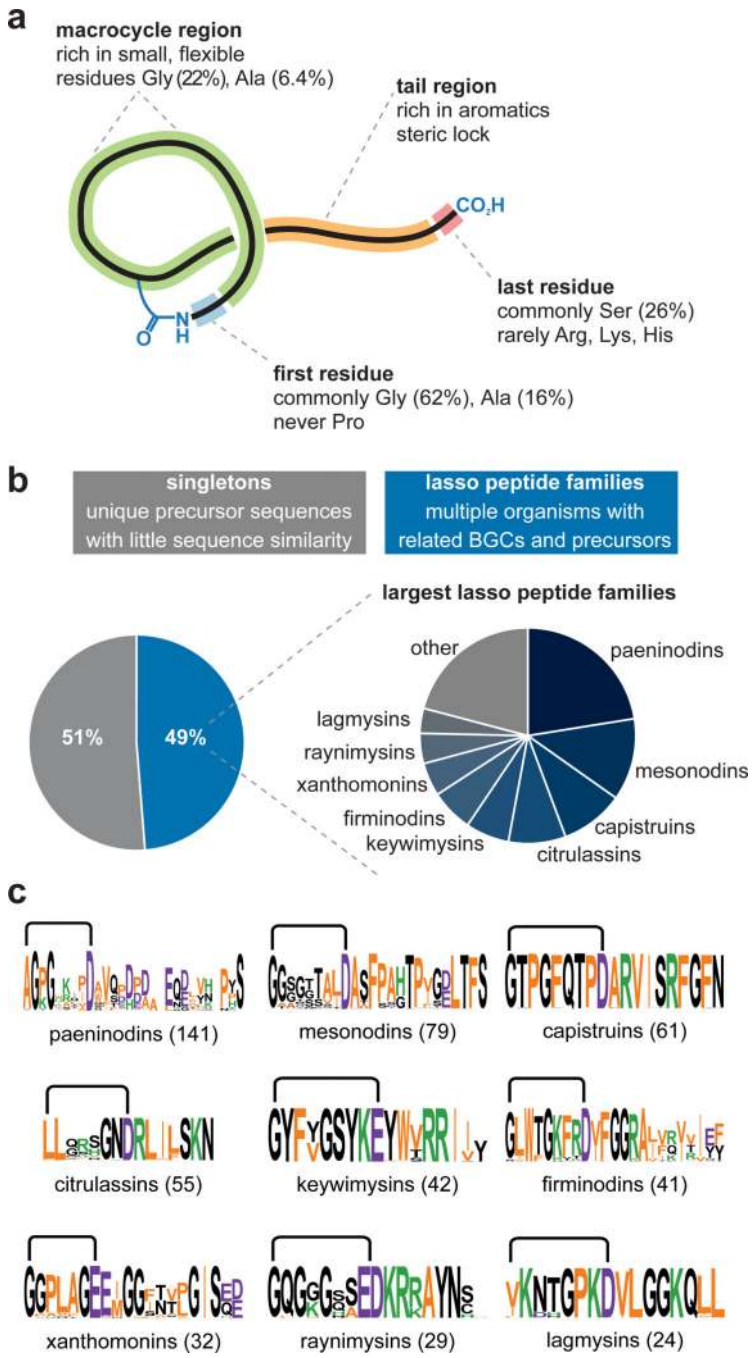

Figure 3. Lasso peptide sequence analysis

(**a**) Lasso peptide structural features and common residues mined from RODEO analysis. (**b**) Organization of predicted lasso peptides into families. Roughly half of the identified BGCs had identical or closely related clusters in other organisms. (**c**) Sequence logos for lasso peptide families with >20 members show wide variance in sequence composition and degree of conservation. Numbers in parentheses refer to number of instances found per family. Residues are color-coded according to basic (green), acidic (purple), or hydrophobic (orange) character.

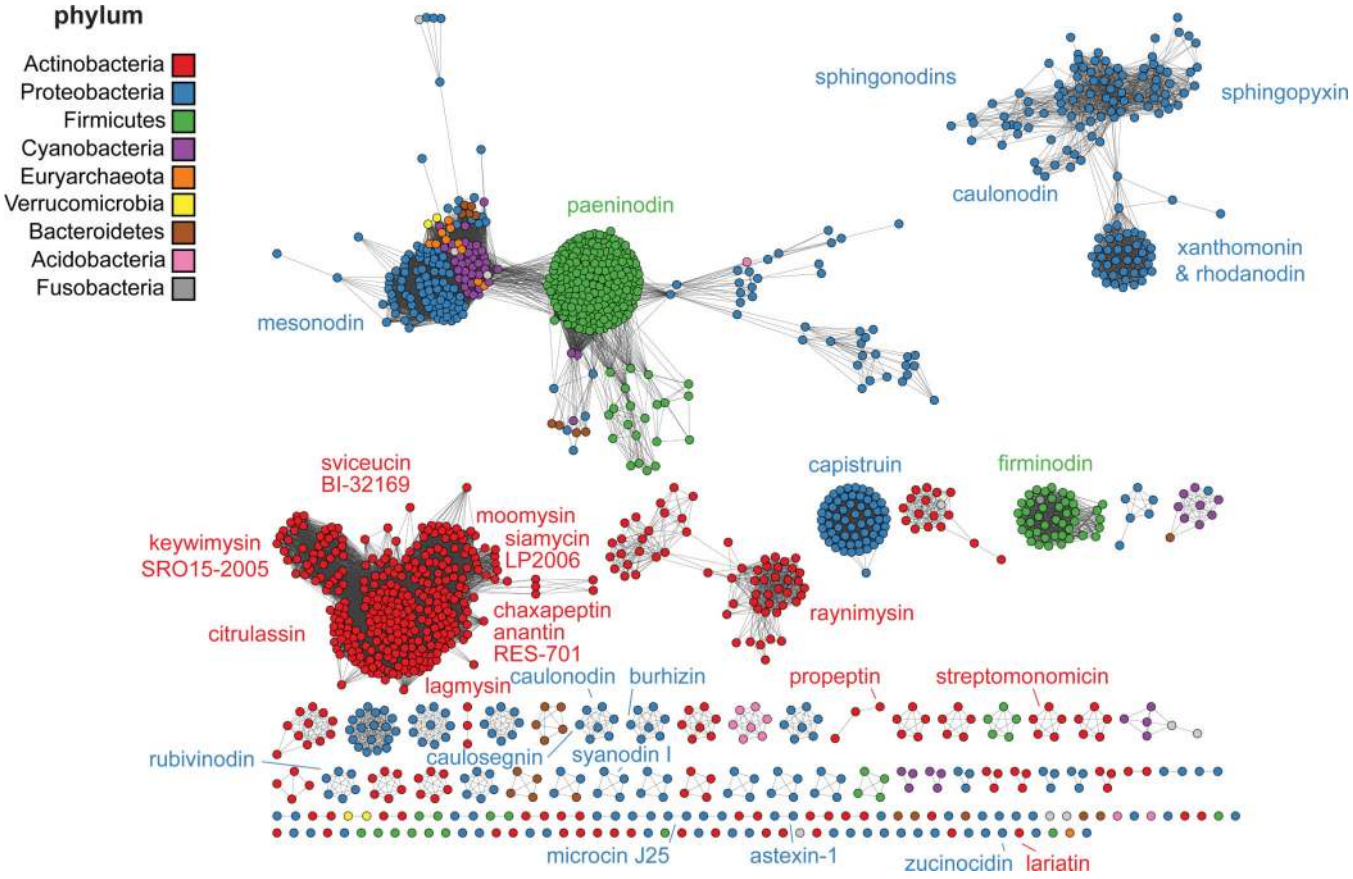

Figure 4. SSNs of lasso cyclase protein

Network was visualized with an edge cutoff E-value threshold of 10^{-80} with nodes colored by phylum. Location within the network of known lasso peptides (including this study) are indicated.

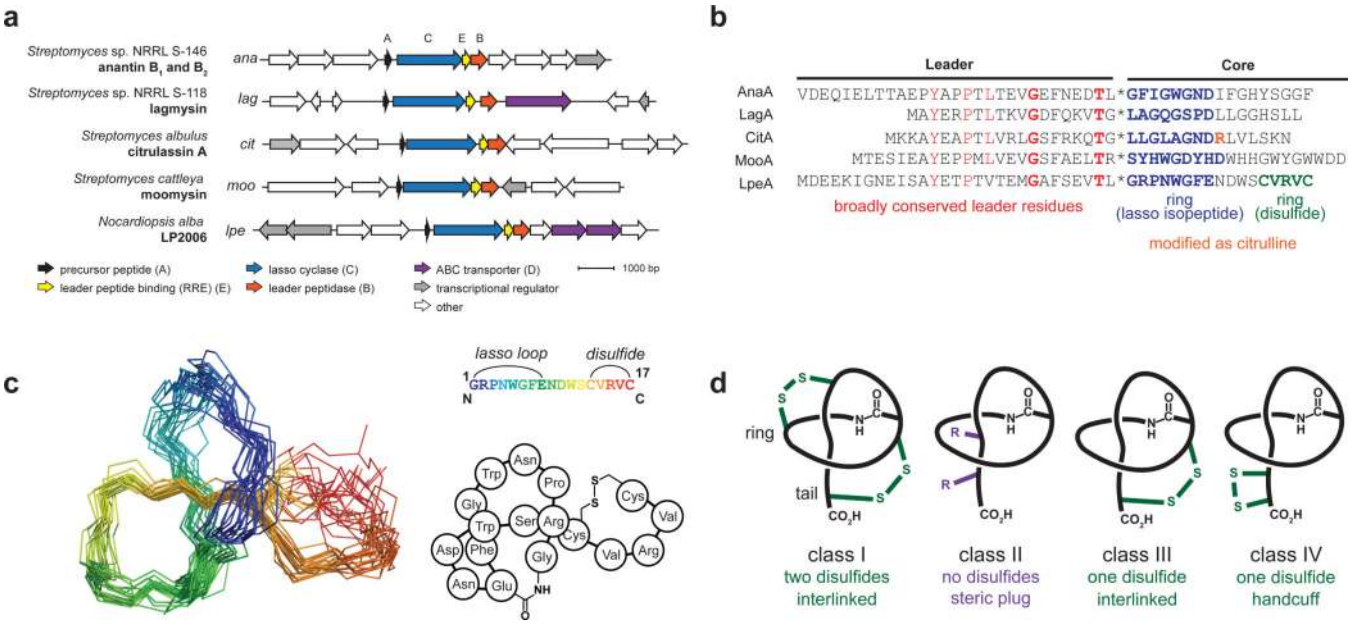

Figure 5. Lasso peptides discovered via RODEO-based prioritization

(**a**) Five gene cluster diagrams for the six investigated lasso peptides. (**b**) Precursor sequences and predicted modified sites from BGCs in (**a**). (**c**) NOE-based NMR ensemble structure and schematic diagram of LP2006 showing the looped-handcuff topology. (**d**) Comparison of LP2006 (class IV) to previous lasso topologies.

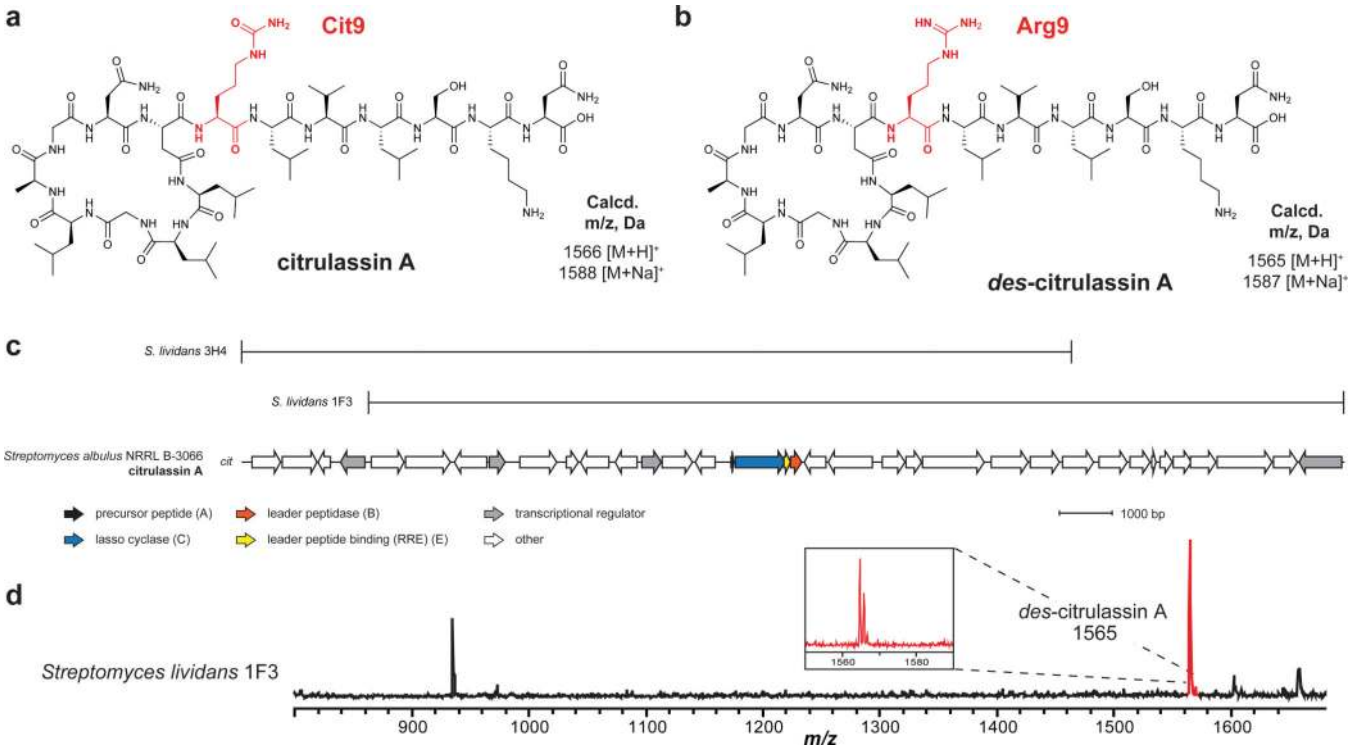

Figure 6. Citrulassin, a rare example of bacterial PAD activity

(**a**) The two-dimensional structure of citrulassin A with Cit9 highlighted. (**b**) Two-dimensional structure of heterologously expressed *des*-citrulassin A with Arg9 highlighted. (**c**) Gene cluster diagrams are shown for two fosmids (3H4 and 1F3) from *S. albulus* NRRL B-3066 expressed in *S. lividans*. (**d**) Production of heterologous *des*-citrulassin A as indicated by MALDI-TOF MS.