ARTICLE
PUBLISHED ONLINE: 9 OCTOBER 2011 | DOI: 10.1038/NCHEMBIO.684

nature chemical biology

A mass spectrometry–guided genome mining approach for natural product peptidogenomics

Roland D Kersten[1], Yu-Liang Yang[2], Yuquan Xu[2], Peter Cimermancic[3], Sang-Jip Nam[1], William Fenical[1,2], Michael A Fischbach[3], Bradley S Moore[1,2]* & Pieter C Dorrestein[1,2,4]*

Peptide natural products show broad biological properties and are commonly produced by orthogonal ribosomal and nonribosomal pathways in prokaryotes and eukaryotes. To harvest this large and diverse resource of bioactive molecules, we introduce here natural product peptidogenomics (NPP), a new MS-guided genome-mining method that connects the chemotypes of peptide natural products to their biosynthetic gene clusters by iteratively matching *de novo* tandem MS (MS^n) structures to genomics-based structures following biosynthetic logic. In this study, we show that NPP enabled the rapid characterization of over ten chemically diverse ribosomal and nonribosomal peptide natural products of previously unidentified composition from *Streptomycete* bacteria as a proof of concept to begin automating the genome-mining process. We show the identification of lantipeptides, lasso peptides, linardins, formylated peptides and lipopeptides, many of which are from well-characterized model *Streptomycetes*, highlighting the power of NPP in the discovery of new peptide natural products from even intensely studied organisms.

Peptide natural products (PNPs) are ubiquitous chemicals found in all life forms, where they have diverse biological functions in development, protection and communication[1]. Nature has evolved two orthogonal biosynthetic pathways to these highly modified peptides involving ribosomal and nonribosomal processes[2]. Although nonribosomal peptides have limited distribution, being restricted mainly to microorganisms with large genomes[3], ribosomally synthesized and post-translationally modified peptides seem to have a much broader distribution throughout nature, including being present in humans[4,5]. The enormous diversity and distribution of PNPs and their associated biological functions, however, are only now being fully realized because of time-consuming discovery options. We report here a new MS–guided genome mining method that quickly connects the chemotypes of expressed PNPs to their biosynthetic pathways, thereby enabling the rapid identification of transcriptionally active PNP biosynthetic gene clusters and the classification of their associated products in a streamlined discovery platform.

Among PNPs, ribosomally synthesized peptides encompass a rapidly expanding group of natural products[6]. Multiple classes of ribosomal peptide natural products (RNPs) of prokaryotic origin have been characterized through their biosynthetic pathways, which entail diverse post-translational modification strategies to yield lantipeptides[7], thiopeptides[8], cyanobactins[9], lasso peptides[10] and other microcins[11]. Consequently, traditional RNP classification systems based on bioactivity, producer and structure[11,12] have shifted toward a new classification based largely on biosynthesis (**Supplementary Results** and **Supplementary Table 1**). In RNP biosynthesis, the peptide sequence is encoded by a precursor gene directly translated by the ribosome to consist of leader peptide and core peptide regions[13]. The leader peptide serves as a scaffold and contains recognition sites for processing enzymes that introduce post-translational modifications of the RNP biosynthetic machinery, whereas the core peptide constitutes the primary sequence of the produced peptide natural product that is modified. Post-translational modification of the core peptide by biosynthetic enzymes can often be extensive and can provide a wealth of structural diversity rendering these peptides, at first glance, unrecognizable as ribosomally synthesized molecular entities[6] (**Fig. 1**). Nonribosomal peptides are conversely synthesized by multifunctional assembly line proteins that instead code for their amino acid precursors through an adenylating enzyme that selects and transfers its substrates to carrier proteins to facilitate peptide synthesis by the nonribosomal peptide synthetase (NRPS) machinery[14]. This process can capture a much wider array of substrates beyond the 20 proteinogenic amino acid building blocks, which limit input into RNPs, to produce notable examples such as the clinical agents vancomycin, daptomycin and cyclosporin[2] (**Fig. 1**).

To estimate the extent of PNP chemical diversity in bacteria, we systematically queried the Joint Genome Institute (JGI) database of 1,035 completed genomes as of September 2010 for RNP and NRPS pathways. Searching for gene clusters harboring characteristic RNP biosynthetic Pfam (protein family) domains[15], we estimate that at least 71% of the deposited bacterial genomes contain biosynthetic features that support common RNP classes (**Supplementary Table 3**). We identified 1,966 candidate RNP gene clusters, 637 of which have two or more of the nine Pfam domains found most frequently in RNP gene clusters (**Supplementary Table 2**). In comparison, 69% of the genomes we searched contained NRPS Pfam domains and 53% had hybrid NRPS-PKS biosynthetic features (**Supplementary Table 3**). Because the training set for our algorithm contained only 24 known RNP gene clusters, our estimate of RNPs is not comprehensive. Nonetheless, this analysis shows that the genetic capacity to produce RNPs is common in most microbial phyla and that RNPs represent one of the most underappreciated classes of bioactive molecules.

Given the sheer volume of predicted bacterial PNPs in publicly available genome strains, we set out to develop a method that takes advantage of recent technological advances in MS and genomics

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, California, USA. [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, California, USA. [3]Department of Bioengineering and Therapeutic Sciences and California Institute of Quantitative Biosciences, University of California, San Francisco, California, USA. [4]Department of Pharmacology and Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California, USA. *e-mail: pdorrestein@ucsd.edu or bsmoore@ucsd.edu.

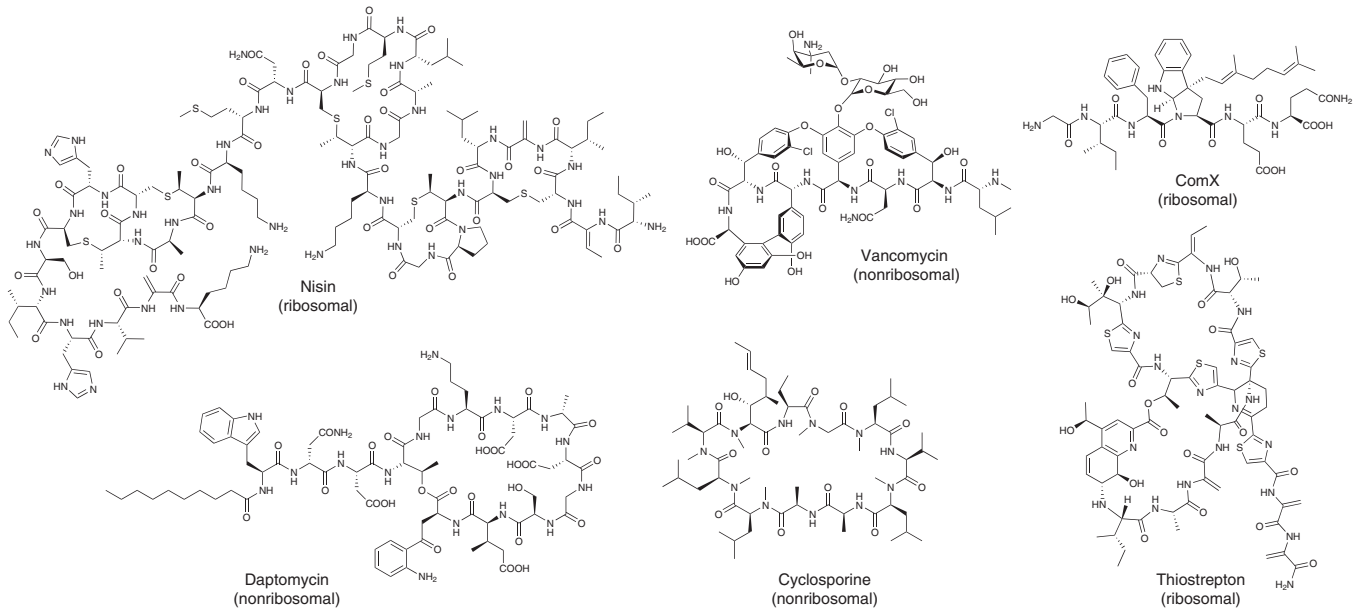


Figure 1 | Structural diversity of peptide natural products.

to streamline the discovery process. The recent development of genome mining has transformed natural product discovery by allowing the targeting of new chemical entities predicted by bioinformatics[16]. In the case of RNPs, a produced peptide structure can be directly linked to the corresponding biosynthetic genes by identifying the core peptide sequence in the translated genome sequence[4]. Furthermore, large portions of NRPs often readily correlate to the predicted amino acid specificity found on their associated modular synthetases[17] (**Supplementary Fig. 1**). This connection of PNP chemotype to genotype has been accomplished in numerous genome-mining studies[8–10,17–20]. One of the major limitations with these approaches is that they only characterize one molecule at a time or require extensive genetic manipulations[21]. With an increasing number of available genome sequences, there is a growing need for new genome-mining methods that can readily connect expressed natural products (chemotype) with their gene clusters (genotype) and that have the potential for automation.

MS is an important technique in the analysis of peptide natural products because of its high sensitivity, its easy implementation into automated processes such as metabolomic or proteomic platforms and its capability for *de novo* peptide structure elucidation by tandem MS[22]. Peptides fragment in MS^n experiments, for example, collision-induced dissociation (CID), in a common way to yield fragment ions in the MS^n spectrum that differ in mass by the amino acid monomers of the corresponding peptide sequence and, thus, enable *de novo* peptide sequencing. MS^n is used in proteomic workflows to identify proteins by connecting peptide MS^n data to protein sequence databases. One approach to link a proteolytic peptide to its database gene uses short *de novo* sequence tags for the database search[23]. However, automated *de novo* sequencing makes errors in one in every four amino acids, and this error rate is enhanced when post-translational modifications (PTMs) are included. In addition, database proteomic tools still struggle to connect modified RNPs with their precursor genes in genomic databases because of scoring functions, which have difficulty in recognizing many PTMs per peptide[24]. The scoring allows for a specific percentage of false-positive rates (usually 1–5%) without any further confirmation of a spectrum-peptide match (**Supplementary Table 4**). Finally, there are no tools that connect MS^n data of nonribosomally synthesized peptides to the corresponding NRPS genes. Given the advantage of MS to automatically acquire data of partial peptide structures from small amounts of material, it could enable a more rapid connection of peptidic natural products with their biosynthetic genes if MS^n data processing is effectively combined with genome mining of RNP and NRP biosynthetic pathways.

In this study, we establish the concept of MS–guided genome mining for peptide natural products called natural product peptidogenomics (NPP). We first highlight proof-of-concept experiments in which NPP characterizes the ribosomal lantipeptide AmfS from *Streptomyces griseus* IFO 13350, the nonribosomal lipopeptide stendomycin I from *Streptomyces hygroscopicus* ATCC 53653 and their corresponding biosynthetic gene clusters. In all, we show that NPP can be applied to characterize many PNP chemotypes and genotypes by introducing 14 new streptomycete PNPs in a very effective genome mining approach.

RESULTS

The NPP concept

NPP is an easy to implement and unbiased, MS-based, chemotype-to-genotype genome mining approach to rapidly characterize ribosomal and nonribosomal peptide natural products and their biosynthetic gene clusters from sequenced organisms (**Fig. 2**). In short, NPP aims to match a series of mass shifts obtained from an MS^n spectrum of a putative PNP to the genes that are responsible for its production. The NPP genome-mining workflow has several iteration steps, which ensure that a match of peptide MS^n data to a genomics-derived peptide structure makes sense biosynthetically. In this way, NPP takes advantage of the enormous wealth of knowledge of PNP biosynthesis gained over the past decade[2].

In practice, the NPP workflow starts with a MALDI-TOF MS analysis of the organism or extract in order to detect unknown masses. We targeted the mass range of 1,500–5,000 Da, as most masses in this window are not described in microbes at the chemical level, and thus they provided an opportunity to apply the NPP approach. However, there is no inherent limitation in size in the NPP approach as long as the MS^n data becomes a unique identifier for a biosynthetic pathway. MALDI-TOF MS analysis of crude butanol extracts or MALDI imaging of agar cultures ensure that the compounds are actively expressed and are captured on semi-solid media. Though not necessary for the PNP discovery process, MALDI imaging links secreted metabolites directly to the morphology of microbial colonies[25] and, thus, decreases potential media or extraction artifacts. Putative peptides are subsequently enriched using a MS-guided isolation using

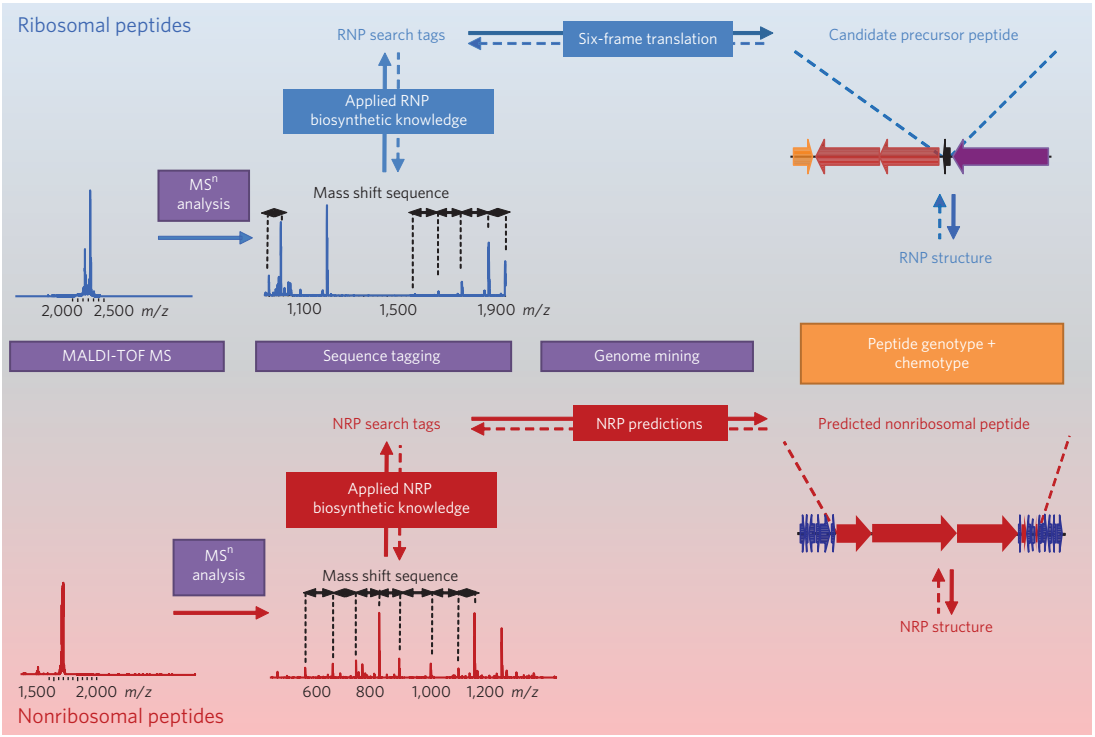


Figure 2 | General workflow of natural product peptidogenomics. NPP can be applied to characterize both ribosomal and nonribosomal peptide natural products in their genotype and chemotype from genome-sequenced organisms. Two proof-of-concept NPP experiments are outlined: ribosomal peptides (RNPs) or nonribosomal peptides (NRPs) and their respective biosynthetic gene cluster can be characterized from a *Streptomyces* extract by MALDI-TOF mass spectrometry (MS) detection, MS^n sequence tagging and PNP genome mining. The iterative approach in matching MS^n data to the genomics-derived peptide structures is shown with dashed arrows. See **Figures 3** and **4** for a detailed NPP analysis of ribosomal and nonribosomal peptides.

size-exclusion chromatography followed by enrichment and desalting steps. An enriched putative PNP is then analyzed in a sequence-tagging step by MS^n. In general, NPP sequence tagging is the formation of *de novo* sequence tags that are searchable in the genome-mining query space of PNPs (**Fig. 2**). This includes the generation of an amino acid sequence tag from a mass shift sequence in an MS^n spectrum and the subsequent processing of the MS^n sequence tag into search tags. The mass shifts define the candidate amino acid residues from all possible monomers that could be encoded in an RNP-based precursor gene or that could be loaded by a corresponding NRPS. This processing of MS^n mass shifts to genome-mining monomers considers PTMs, nonribosomal substrates, fragmentation gas-phase behavior and chemical modifications of amino acid residues during purification and MS analysis. NPP-based RNP genome mining interrogates the six-frame translation of the genome for candidate precursor peptides that comprise any of the search tags. As there may be multiple matches to a search tag that is 5–10 amino acids long, the correct RNP precursor gene is identified by applied biosynthetic knowledge in which the search tag should associate with the C-terminal half of a <100-amino-acid–long open reading frame (ORF) that clusters with RNP biosynthetic genes. NPP-based NRP genome mining, on the other hand, queries all predicted nonribosomal peptides of the target genome for the search tags.

The effectiveness of NPP in connecting PNP structures with biosynthetic genes lies in its iterative approach in matching MS^n-based structures to genomics-based candidate structures following biosynthetic logic, as each search tag match has to be confirmed in mass, sequence and biosynthetic signatures with the MS^n analysis (**Fig. 2**). We showed this effectiveness in a comparison of the NPP approach to current proteomic approaches in identifying precursor genes in RNP genome mining. None of the standard proteomic platforms such as Mascot[26] or InsPecT[23] could identify any of the NPP-characterized RNPs in a search with variable common RNP PTMs or in blind or unrestricted searches designed to find unknown PTMs (**Supplementary Table 4**). InsPecT was able to characterize two of the RNPs after predefining NPP-dissected PTMs in the analysis for each peptide.

NPP characterization of ribosomal peptide AmfS

As a proof of concept of the NPP workflow for RNPs, we targeted the known ribosomal peptide AmfS from *S. griseus* IFO 13350 because this is a well-characterized lantipeptide with four PTMs[27] (**Fig. 3**). MALDI imaging of *S. griseus* and MALDI-TOF MS analysis of an extract resulted in the detection of a secreted mass of 2,212 Da. We then subjected the peptide to CID fragmentation. In the MS^2 spectrum, we assigned the charge states of sequential fragment ions and identified the mass shift sequence 99-99-113-69-101 (**Fig. 3**). We matched the mass shifts to all likely candidate amino acids to yield sequence tags by first substituting with proteinogenic amino acids where possible (**Supplementary Table 5**). We then substituted nonproteinogenic masses with all possible RNP monomers arising from known PTMs. We substituted the shift of 69 Da to the nonproteinogenic amino acid dehydroalanine (Dha) (**Fig. 3** and **Supplementary Table 6**). Dha is a candidate amino acid for ribosomal peptides because dehydrated serine and threonine or dethiolated cysteine are commonly observed in PNP MS^n spectra either as a post-translational modification[28] or as an MS^n gas-phase rearrangement (**Supplementary Fig. 3**). From the resulting sequence tag VVI(L)S(C)T, we created a list of all possible search tags in both sequence directions to give eight putative PNP sequence tags for a search against the *S. griseus* genome sequence (**Fig. 3**)[29]. Of the millions of possible peptide sequences based on a six-frame translation, we identified just one candidate 43-amino-acid–long precursor by the search tag VVLCT (**Fig. 3**). This result fulfilled the

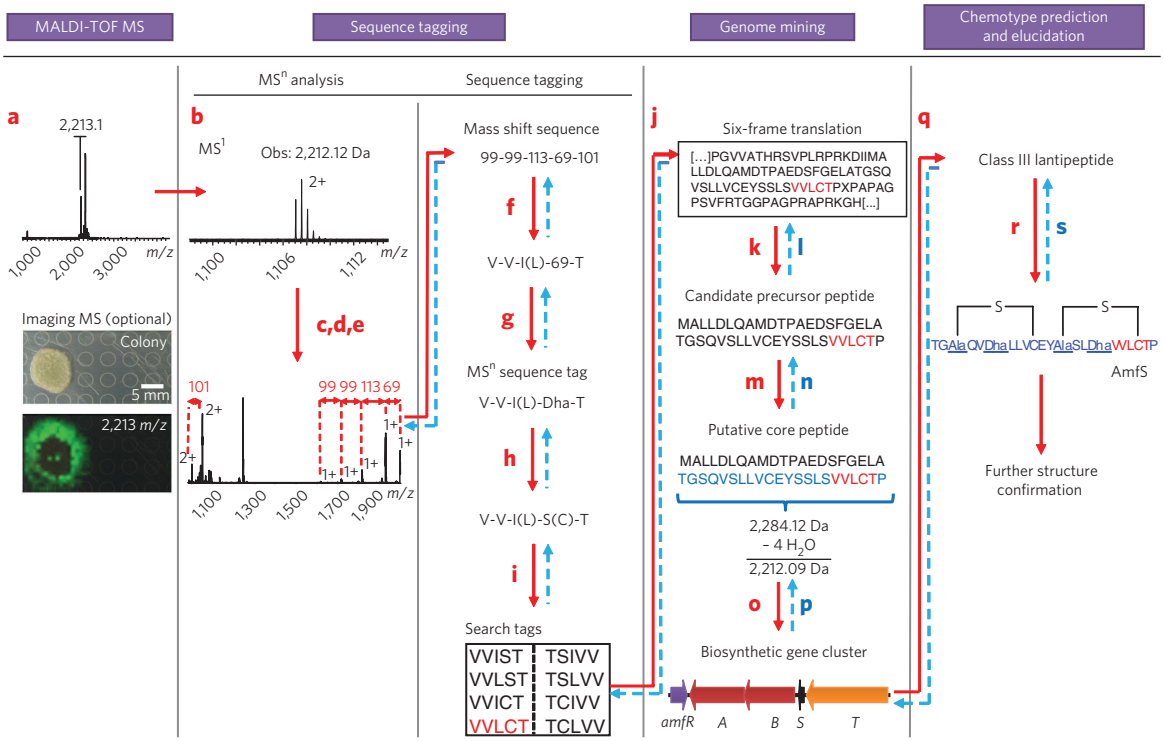


Figure 3 | Peptidogenomic connection of a RNP chemotype with its biosynthetic genes in the characterization of the class III lantipeptide AmfS from *S. griseus* IFO 13350. Analysis was carried out through sequence tagging and genome mining. Iterative aspects in connecting MS^n data of the peptide chemotype to the genotype are highlighted in blue and with the dashed arrows. The steps are as follows: (**a**) detection of putative peptide mass signals by MALDI-TOF MS or imaging MS, (**b**) determination of molecular weight, (**c**) MS^n fragmentation (CID), (**d**) assignment of charge states, (**e**) identification of mass shifts, (**f**) substitution of proteinogenic mass shifts (**Supplementary Table 5**), (**g**) substitution of nonproteinogenic mass shifts with putative RNP monomers (**Supplementary Table 6**), (**h**) MS^n sequence-tag processing of putative biosynthetic or MS gas-phase modifications, (**i**) MS^n sequence-tag processing of sequence tag direction, (**j**) search in six-frame translation of the target genome, (**k**) identification of candidate precursor peptide through RNP biosynthetic rationale, (**l**) verification of precursor peptide sequence, (**m**) prediction of core peptide sequence based on observed mass and putative PTM mass shifts, (**n**) verification of core peptide sequence and mass, (**o**) prediction of biosynthetic gene cluster, (**p**) verification of putative PTMs, (**q**) RNP classification, (**r**) structure prediction based on RNP class and MS^n data and (**s**) structure verification by MS^n data. Dha, dehydroalanine; Obs, observed.

RNP biosynthetic requirement of the search tag being located in the C-terminal half of a <100-amino-acid–long gene product. Next, we compared a predicted core peptide sequence in its calculated mass to the observed mass in the MS^1 spectrum considering putative PTMs such as, for example, dehydrations. The calculated mass of the 22-amino-acid–long core peptide ^{22}T-^{43}P differed by four putative PTM dehydrations from the observed mass (**Fig. 3**), which is in agreement with the formation of two Dha and two lanthionine bridges in AmfS[30]. In addition, the predicted core peptide sequence could be further verified at this step by comparison to the MS^n data. Subsequent BLAST analysis of the neighboring genes identified the remainder of the AmfS biosynthetic gene cluster[27] and, therefore, further verified the connection of RNP chemotype and genotype. Based on the gene cluster components, in particular the AmfS core peptide and the PTM-introducing enzymes AmfA and AmfB, we could characterize the analyzed peptide as a class III lantipeptide from known RNP biosynthetic gene clusters (**Supplementary Table 1**). Finally, we could verify an AmfS structure based on the given core peptide sequence, the MS^n data and the knowledge about AmfS-like lantipeptide PTMs[30] (**Fig. 3**, **Supplementary Fig. 2** and **Supplementary Table 1**). The proof-of-concept characterization of AmfS and its gene cluster highlights the effectiveness of the NPP workflow by its iterative utilization of MS^n data and genetic data to enable a peptidogenomic connection of a PNP chemotype with its genotype.

NPP characterization of nonribosomal lipopeptides

With a minor adjustment to the NPP workflow, we can also discover NRPs (**Fig. 4**). We exemplified this approach with a set of lipopeptides detected by MALDI imaging from a colony of *S. hygroscopicus* ATCC 53653 (**Fig. 4**). MS^n analysis of SHY-1628 yielded the sequence fragment 99-99-83-83-71-113-99-57-115 (**Fig. 4** and **Supplementary Fig. 4b**), which we first processed into the RNP workflow because most of its mass shifts corresponded to proteinogenic amino acids through the sequence tag V-V-83-83-A-I(L)-V-G. We substituted the 83-Da masses with Dhb, whose biosynthetic precursor is threonine (**Supplementary Table 6**). Although we queried the six-frame translation of *S. hygroscopicus* with the search tags VVTTAI(L)VG, we detected no precursor peptides based on the described biosynthetic requirements of RNPs. The inability to identify a precursor peptide from a long sequence tag suggested that the SHY-1628–based peptides could instead be a set of nonribosomal peptides. To explore this scenario, we revised the original sequence tag to include NRP-specific, nonproteinogenic residues for NRP genome mining (**Supplementary Table 7**). Hence, the 83-Da mass shifts could correspond to Dhb, *N*Me-Dha or homoserine lactone (HseL). We excluded HseL because of its common C-terminal NRP location. Dhb and *N*Me-Dha most likely derived biosynthetically from threonine and serine, respectively, during or after NRP assembly because of the enamine instability of putative Dha and Dhb monomers[31]. In addition, a Dhb mass shift could also derive from a MS gas-phase–induced ring-opening elimination of a threonine-macrolactone bond with the NRP C terminus[32]. We evaluated the VVT(S)T(S)AI(L)VG sequence tags (**Fig. 4**) against NRP sequences predicted by NP.searcher[33] and antiSMASH[34] algorithms that predict NRPS gene clusters and their NRP products from the genome supercontig. This analysis matched the reduced and full eight-amino-acid

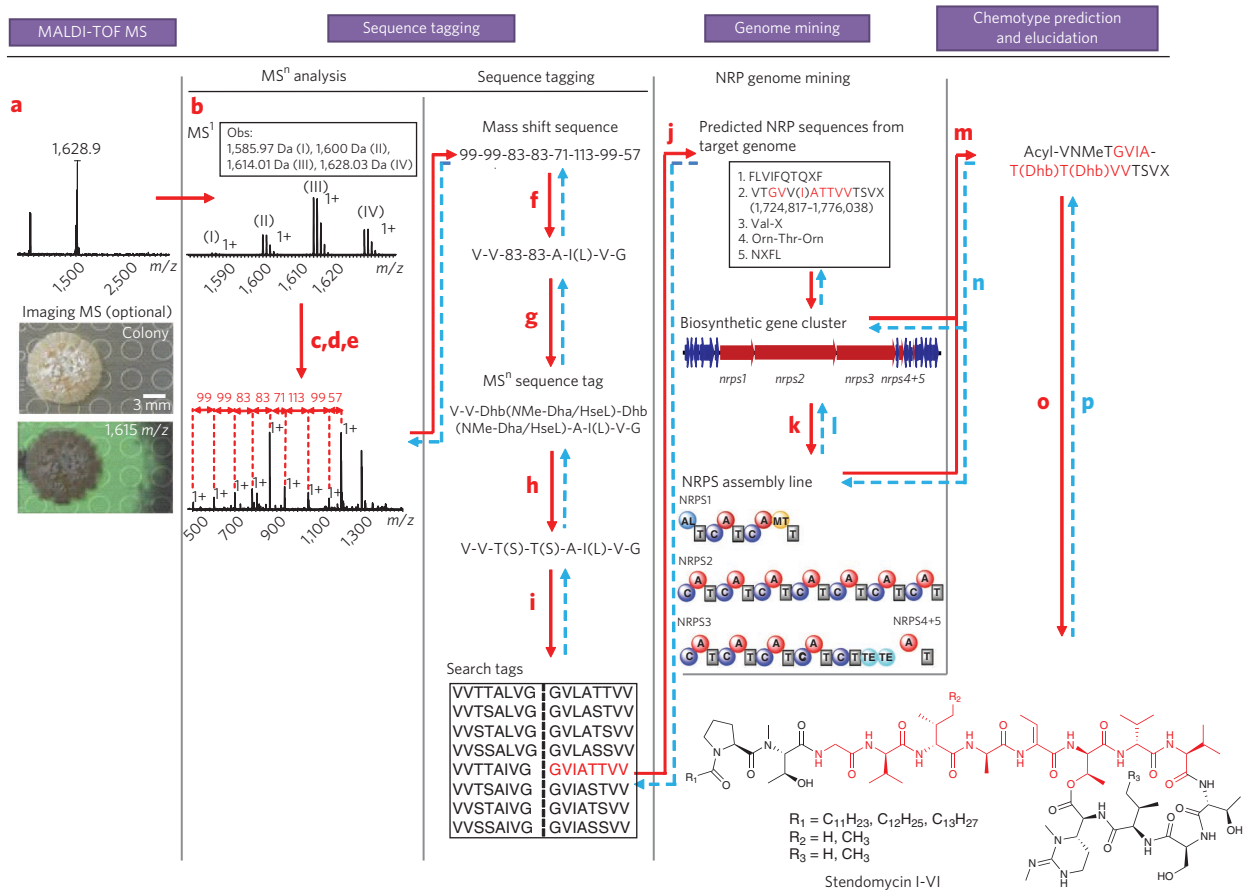


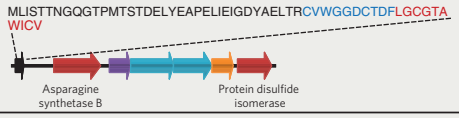
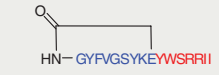
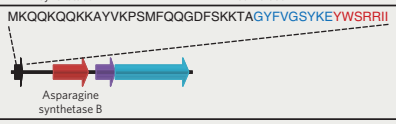
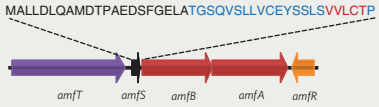
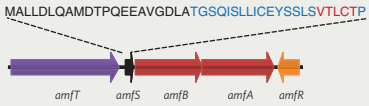
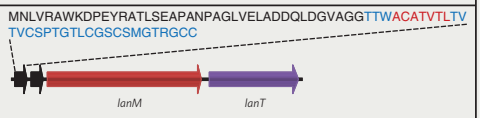
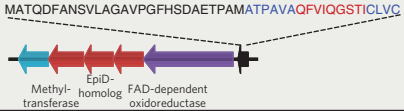
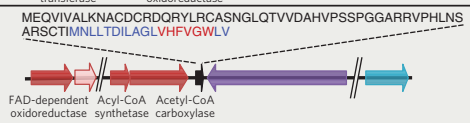
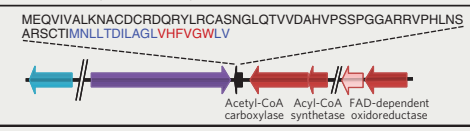
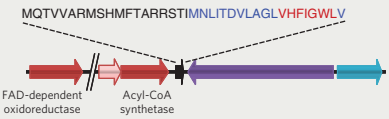
Figure 4 | Peptidogenomic connection of a NRP chemotype with its biosynthetic genes in the characterization of the lipopeptide stendomycin complex from *S. hygroscopicus* ATCC 53653. Analysis was carried out through sequence tagging and genome mining. Iterative aspects in connecting MS^n data of the peptide chemotype to the genotype are highlighted in blue and with the dashed arrows. The steps are as follows: (**a**) detection of putative peptide mass signals by MALDI-TOF MS or imaging MS, (**b**) determination of molecular weight, (**c**) MS^n fragmentation (CID), (**d**) assignment of charge states, (**e**) identification of mass shifts, (**f**) substitution of proteinogenic mass shifts (**Supplementary Table 5**), (**g**) substitution of nonproteinogenic mass shifts with putative NRP monomers (**Supplementary Table 7**), (**h**) MS^n sequence-tag processing of putative biosynthetic and MS gas-phase modifications, (**i**) MS^n sequence-tag processing of sequence tag direction, (**j**) search of predicted NRP sequences from the target genome with all search tags (NP.searcher or antiSMASH), (**k**) biosynthetic gene cluster analysis, (**l**) verification of predicted NRPS assembly line, (**m**) NRP structure prediction, (**n**) verification of predicted NRP structure, (**o**) full structure elucidation based on MS^n and NMR data and (**p**) verification of NRP structure. Dha, dehydroalanine; dhb, dehydrobutyrine; hseL, homoserine lactone; orn, ornithine; Obs, observed.

sequence tag to one candidate NRP sequence out of the five predicted NRPS sequences in the *S. hygroscopicus* genome (**Fig. 4** and **Supplementary Fig. 5**).

Again, through an iterative process, we inspected the corresponding gene cluster and found it to contain an N-terminal acyl ligase domain associated with lipopeptide biosynthesis, which is in full agreement with the observed 14-Da separation of the parent ions characteristic of lipopeptides. Further MS^n (**Supplementary Fig. 7b** and **Supplementary Tables 8–12**) and nuclear magnetic resonance (NMR) analysis (**Supplementary Fig. 6** and **Supplementary Table 13**) identified the lipopeptides as members of the stendomycin antibiotic family of lipotetradecapeptides that contain a seven-membered macrolactone and a total of seven modifications[35]. Aside from stendomycin I, which was originally characterized in *Streptomycetes endus*[35], we characterized five new stendomycin analogs (II–VI) that differed in the acyl chain and in valine or isoleucine substitutions at positions 5 and 13 for the first time in *S. hygroscopicus* ATCC 53653. The biosynthetic features of the identified gene cluster matched the structure of stendomycin I in the NRPS substrates and modifications (**Supplementary Figs. 5,8**). Thus, as we found for RNPs, the iteration between the MS^n analysis and genome mining enabled the fast and reliable connection of an NRP chemotype and genotype (**Fig. 4**). For example, we detected a low-resolution mass shift of 115 Da in the MS^2 spectrum of stendomycin I (**Supplementary Fig. 4b**) that was first assigned to aspartic acid. However, the corresponding module of the putative stendomycin NRPS instead predicts *N*Me-threonine (also a 115-Da shift) at this position and, thus, the mass shift in the MS^n spectrum could be explained. This example illustrates that in NRP sequence tagging, modifications such as N-methylations of proteinogenic masses and even nonproteinogenic masses should be considered if the first iterative round of NRP genome mining misses the assignment of the tag.

We also successfully applied the NPP method to other NRPS-derived molecules such as the structurally diverse calcium-dependent antibiotic[36], surfactin[37], plipastatin[37], pyoverdine[38] and daptomycin[36], and in each case, we identified the correct gene cluster (data not shown). Recently, the NPP workflow enabled the discovery of the arylomycin gene cluster with a sequence tag of just two amino acids[39]. This highlights the point that with NRPS-derived molecules, minimal sequence information can be sufficient to find a match in an NRP database of less than ten predicted NRP sequences per genome despite there being >526 known NRP monomers[40] because of the iterative nature of using biosynthetic knowledge in the workflow. To complete the structure analysis, additional analytical methods such

Table 1 | NPP characterization of nine new RNPs and their associated gene clusters from seven genome-sequenced *Streptomyces* strains

Observed PNP	Class	Chemotype	Genotype
SSV-2083	Class I lasso peptide		
SRO15-2005	Class II lasso peptide		
SRO15-2212	Class III lantipeptide		
SAL-2242	Class III lantipeptide		
SRO15-3108	Class II lantipeptide	TTWACATVTLTVTVCSP TGTLCGSCSMGTRGCC (Core peptide – $9H_2O$)	
SGR-1832	Linaridin		
SLI-2138	N-formylated peptide		
SCO-2138	N-formylated peptide		
SWA-2138	N-formylated peptide		

Precursor peptide | Post-translationally modifying enzyme | Transporter | Protease | MreB | Regulator

Shown is a summary of the diverse RNP chemotypes and genotypes characterized by NPP in this study. Detailed analyses are described in **Supplementary Results.**

as NMR and Marfey's analysis are needed to complement the wealth of tandem MS and biosynthetic information, as has been done with stendomycins. The characterization of five stendomycin derivatives and their biosynthetic gene cluster in *S. hygroscopicus* shows that the NPP workflow can be readily accommodated to additionally discover modified NRPs.

NPP characterization of new RNP chemotypes and genotypes

Next, we interrogated several sequenced *Streptomycetes* to explore the practicality of NPP in the identification of other uncharacterized RNPs. From seven *Streptomyces* strains, we identified multiple previously uncharacterized RNPs and their gene clusters using the NPP approach (**Table 1**). The first unknown RNP and its gene cluster that we characterized by NPP was a class I lasso peptide, SSV-2083, from *Streptomyces sviceus* ATCC 20983 (**Table 1** and **Supplementary Fig. 9**). The discovery and isolation of secreted SSV-2083 from sporulating colonies was guided by MALDI imaging and MALDI-TOF MS of the ion at 2,084 m/z. An MS^n analysis of the unmodified compound provided no sequence information (**Supplementary Fig. 9b**). One of the main experimental challenges in the generation of the sequence tag is that many of these molecules are constrained by disulfide or thioether linkages, and therefore they provide poor to no fragmentation data (**Supplementary Fig. 10**). In such cases, samples are reductively dethiolated with $NaBH_4$ and $NiCl_2$ treatment[41] and resubjected to tandem MS to reveal longer sequence tags for PNP genome mining. Deconstrained SSV-2083 yielded a ten-amino-acid MS^n sequence tag that we identified in the six-frame translation of the *S. sviceus* genome in a 56-amino-acid candidate precursor peptide. This observation enabled the identification of the SSV-2083 biosynthetic gene cluster containing conserved lasso peptide biosynthetic genes as well as a new protein disulfide-isomerase–encoding gene (**Supplementary Fig. 9c**). Alignment with known class I lasso peptides in combination with tandem MS data (**Supplementary Fig. 9d**) enabled the prediction of the candidate SSV-2083 structure (**Table 1**), and these results represent the first class I lasso peptide gene cluster[42].

NPP characterization of new RNP classes from *Streptomyces*

Our NPP analysis also resulted in the discovery of two new RNP classes and their genetic origins from well-scrutinized

Streptomycetes, namely SGR-1832 (**Table 1** and **Supplementary Fig. 11**) from *S. griseus* IFO 13350 and SCO-2138 (**Table 1** and **Supplementary Fig. 12**) from *Streptomyces coelicolor* A3(2)[43]. Based on the gene cluster and the MS fragmentation data, we determined SGR-1832 to be a linear 19-residue peptide with an N-terminal *N*,*N*-dimethylalanine, two dehydrobutyrines and a rare C-terminal aminovinylcysteine (AviCys) residue. These unusual post-translational modifications are reminiscent of those seen in cypemycin, a related AviCys-containing linaridin from *Streptomyces* sp. OH-4156, whose biosynthesis was recently revealed by genome mining[18]. Peptide SCO-2138, detected only in organic extracts, is also a previously unidentified 19-amino-acid RNP from *S. coelicolor* A3(2) that produces a number of other peptide natural products[44]. The corresponding gene neighborhood containing a conserved unknown protein, a protease and a rod-shape–determining protein[45] is also found in other *Streptomyces* genomes (**Supplementary Fig. 12c**,**d**). Accordingly, we isolated and characterized two SCO-2138 homologs using NPP from *Streptomyces lividans* TK24 (SLI-2138, which is identical to SCO-2138; **Table 1** and **Supplementary Fig. 12**) and *S.* sp. E14 (SWA-2138, which is isomeric to SCO-2138; **Table 1** and **Supplementary Fig. 12**). These RNPs have a 28-Da N-terminal modification, which we confirmed by Fourier transform MS ($FTMS^n$) to be an *N*-formyl unit (**Supplementary Fig. 12e**). The SGR-1832 and SCO-2138 peptides represented undiscovered classes of RNPs at the time of this analysis and showcase that new RNP classes can be discovered by the NPP method.

Characterization of multiple PNPs in one NPP experiment

NPP analysis of the daptomycin-producing bacterium *Streptomyces roseosporus* NRRL 15998 (ref. 46) enabled the identification of three new RNPs and their gene clusters in a single NPP experiment (**Supplementary Fig. 13**). SRO15-2005 (**Table 1** and **Supplementary Fig. 14**) is a class II lasso peptide; SRO15-2212 (**Table 1** and **Supplementary Fig. 15**) is identical to the class III lantipeptide AmfS, which was previously uncharacterized in this strain; and SRO15-3108 (**Table 1** and **Supplementary Fig. 16**) is a class II lantipeptide that undergoes nine dehydrations during maturation. The detection of these three RNPs and their corresponding gene clusters in one NPP experiment shows the potential of NPP as a high-throughput discovery methodology.

DISCUSSION

In this work, we introduce NPP as a chemotype-to-genotype genome-mining approach for the characterization of ribosomal and nonribosomal peptide natural products and their respective biosynthetic gene clusters by identifying 14 peptides from well-known genome-sequenced *Streptomycetes*. In contrast to global metabolomic[47] and peptidomic[24] strategies, NPP is a targeted approach in which MALDI imaging or MALDI-TOF MS analysis of organic extracts is defined by a preselection of ions that are putative peptide natural products of expressed biosynthetic pathways. The innovation of NPP in efficiently linking these putative peptides to their gene clusters is firmly grounded in the connection of *de novo* MS^n peptide sequence tags of modified peptides to precursor peptides or to predicted NRPS products by applying biosynthetic knowledge and iterative steps between MS^n analysis and PNP genome mining for confirmation of putative chemotype-genotype matches. Because peptides are often structurally constrained, the generation of an MS^n sequence tag is facilitated by structural deconstraining the peptide before MS^n analysis. This yields simpler peptide structures and, thus, higher quality sequence tags, as in the case of the class I lasso peptide SSV-2083 (**Supplementary Fig. 9b**). Deconstraining also aids in the elucidation of post-translational modifications, such as the AviCys group of linaridin SGR-1832. In MS^n sequence-tag processing, the approach takes advantage of the degeneration of residues in the MS^n sequence tag by mass, reactions in the mass spectrometer, biosynthesis or sequence directionality to ensure that the resulting search tags can be found in genomics-derived peptide sequences (**Figs. 3** and **4**). In PNP genome mining, the sequence tags are searched against a query space that is different for RNPs and NRPs. In RNP genome mining, the query space is the six-frame translation of the target genome and, thus, is large. The sequence tag for effective genome mining of a precursor peptide in this large query space should be at least five amino acids; otherwise too many candidate precursor peptides will be obtained to be further differentiated based on RNP biosynthetic requirements. Several characterized precursor peptides that we identified in this study were not previously annotated in the NCBI database[48] (peptides SCO-2138 and SGR-1832). We found these peptides only in the six-frame translations of the *S. coelicolor* and *S. griseus* genome supercontigs. Although the drawback of an extended database providing more candidate precursor peptides for a certain sequence tag is a potential concern, this larger protein inference problem (as it is known in global proteomics[24]) is effectively solved in NPP by the iterative matching of the candidate precursor peptides in mass, sequence and biosynthetic signatures to the MS^n data.

MS^n sequence tag processing and the iterative MS^n and genomics analysis make the NPP *de novo* sequencing approach more effective in identifying precursor genes in RNP genome mining than current proteomic approaches. Neither Mascot[26] nor InsPecT[22] could identify any of the NPP-characterized RNPs in searches for unknown PTMs (**Supplementary Table 4**). InsPecT, which also relies on *de novo* sequence tagging, was able to characterize just two of the RNPs (SCO-2138 and SLI-2138), but only after we predefined NPP-characterized PTMs in the analysis. This is about what one would expect, as proteomic tools typically annotate 5–15% of the collected data, although in rare cases this percentage can be higher. The main reason that these programs do not work for these peptides is because their scoring functions have been designed to work for protease-cleaved, water-soluble peptides. Proteomic programs require specific scoring functions for specific PTMs (for example, specific for trypsin-cleaved ubiquitination tags or specific for phosphorylation) and simply have not been developed for RNP-based PTMs.

We further showed that NRPs are readily incorporated in the NPP workflow, as in the case of stendomycins (**Fig. 4**). Even though >50% of all amino acids in NRPs are L- or D-proteinogenic amino acids[40], mass shift sequences obtained from an MS^n spectrum define the candidate monomers to be used for the generation of all possible sequences that are to be compared to the predicted sequences based on the amino acid specificity of the adenylation domains using programs such as NRPSpredictor2 (ref. 49). In NRP genome mining, the query space consists of NRP megasynthetases predicted from the target genome by NP.searcher or antiSMASH and, thus, is relatively small, as most microbial genomes contain less than ten NRPS gene clusters. Consequently, short sequence tags of just two amino acids can be sufficient to correlate the NRP to its cognate NRPS gene cluster[39]. In the case of stendomycin, even though we ultimately applied the 8-amino-acid tag GVIATTVV, we could have functionally operated with and would have obtained the similar results with just a two-amino-acid tag such as VV, VI, TT, IA, AT or GV, as only one of the five *S. hygroscopicus* NRPS gene sets was appropriate in size and sequence. NRP sequences often contain modified and/or nonproteinogenic amino acid residues that can be addressed by including all appropriate nonproteinogenic monomers to a mass shift sequence and by considering their corresponding biosynthetic machineries during genome mining (**Supplementary Table 7**).

Because NPP is a MS-guided approach, it is ultimately dependent on generating quality sequence tags. The challenge in NPP characterization of peptides <500-Da or four-amino-acids long or less is in applying a limited sequence tag for genome mining rather than for dealing with matrix background in the low m/z region during peptide detection by MALDI-TOF MS. The analysis of putative peptides in the mass range <1,500 m/z will also increase the discovery

of PNPs, and in particular, of NRPs. NRPs with curated gene clusters in the NORINE database (which contains nonribosomal peptides) have an average mass of ~950 Da and eight monomers (**Supplementary Fig. 1**), whereas RNPs usually have a higher molecular weight, and NPP is appropriate for all such peptides. NPP, however, in its current implementation, is challenged by NRPs with multiple heterocycles, such as thiopeptides[8], and hybrid NRPS-PKS products with major polyketide portions. This will remain a challenge until the fragmentation rules are established. Another NPP restraint is the bioinformatics predictability of PNP sequences from inadequate genomic data in which poor sequence or annotation quality result in misassigned precursor and NRPS genes. Better genome assembly, improved gene annotation (especially of small ORFs), increased understanding of gas-phase fragmentation behaviors and deeper knowledge of NRPS substrate specificity codes will further empower the tools described in this work.

In conclusion, NPP is a new, MS-based genome-mining platform to guide the discovery of new ribosomal and nonribosomal peptides. This approach enables streamlined screening of peptide chemotypes from multiple organisms and facilitates expanded studies on their isolation, complete structure elucidation, biological evaluation and pathway engineering that leads to an increased appreciation for the understanding of the biological roles and therapeutic potential of peptide natural products. With further automatization of the NPP workflow such as training for offset functions of complex peptides, better understanding of MS fragmentation behaviors and the expansion to smaller masses and additional organisms, NPP has the potential to open up new research directions in the (bio)chemistry of peptide natural products.

METHODS

MALDI imaging of *Streptomyces* colonies. *Streptomyces* strains were grown on solid ISP2 medium (1 l of medium contained 4 g yeast, 10 g malt extract, 4 g dextrose and 20 g agar at pH 7) for 4–10 d at 28 °C until sporulation. *Streptomyces* spores from one plate were suspended in 1 ml sterile water and glycerol (3:1) and stored at –80 °C after inoculation. Thin-layer ISP2 agar plates of sporulating *Streptomyces* colonies were prepared as described elsewhere[25]. The applied matrix was a universal MALDI matrix (Sigma-Aldrich). MALDI imaging of *Streptomyces* samples on a Bruker MSP 96 anchor plate was performed on a Microflex Bruker Daltonics mass spectrometer outfitted with Compass 1.2 software suite (which consists of flexImaging 2.0, flexControl 3.0 and flexAnalysis 3.0). Target plate calibration was done as described elsewhere[25]. The sample was run in positive reflectron mode, with 800-µm laser intervals in XY. After the target-plate calibration was complete, the AutoXecute command was used to analyze the samples. The flexControl method we used had settings as previously described[25] with detection parameters adjusted as follows: mass range of 800–4,200 m/z and detector gain, reflector of 3.7–8.1. Mass calibration was accomplished using a peptide standard mix (Bruker Daltonics) as an external standard. After data acquisition, the data were analyzed using the flexImaging software. The resulting mass spectrum was analyzed manually for mass signals >1,500 m/z. Putative peptide mass signals >1,500 m/z were assigned with individual colors for a display of the distribution of the mass signal in the image.

Mass spectrometry analysis and sequence tagging. Peptide extraction, enrichment and preparation for MS analysis are described in the **Supplementary Methods**. Prepared peptide samples were injected for MS analysis by a nanomate-electrospray ionization robot (Advion) for consecutive electrospray into the MS inlet of a LTQ 6.4T Fourier transform ion cyclon resonance (FT-ICR) mass spectrometer (Thermo Finnigan). MS and MS^n data were acquired in the positive ion mode. FTMS data were acquired in 400–2,000 m/z scans. Selected peptide mass signals were manually isolated and fragmented by CID. MS^n data was collected either in ion trap or FT detection mode. All data were analyzed using QualBrowser, which is part of the Xcalibur LTQ-FT software package (Thermo Fisher). FTMS masses were analyzed using Extract software (Thermo Electron Bremen). Peptide MS^n sequence tags were assigned from MS^n data by manual *de novo* sequencing within the mass accuracy of the mass spectrometer using a mass shift list of proteinogenic amino acid monomers (**Supplementary Table 5**) and nonproteinogenic monomers (**Supplementary Tables 6,7**). Sequence tagging emphasized a correct assignment of 5–10-amino-acid MS^n sequence tags rather than longer, incorrect assignments for reliable genome mining. The MS^n sequence tag was further manually processed into a set of search tags depending on the degree of degeneration of the MS^n sequence tag. The MS^n sequence tag processing included differentiation of positions with identical masses (for example, isoleucine and leucine), positions with biosynthetic modifications (for example, Dha derived from serine or cysteine in RNPs; **Supplementary Tables 6,7**) and positions modified by MS analysis (for example, Dha derived from the cysteine of a lanthionine PTM or Dhb derived from the threonine of a macrolactone linkage). In $NaBH_4^-$ and $NiCl_2$-treated samples, positions were differentiated that might be chemically altered (for example, alanine derived from cysteine or alanine). The MS^n sequence tag was also differentiated in its reversed direction.

Genome mining of ribosomal peptides. A six-frame translated supercontig was searched with all possible RNP search tags from a given MS^n sequence tag in a standard text processing program. A candidate precursor peptide was defined in its N terminus by a pBLAST search of its C-terminal partial sequence to find homologs or was defined by reanalysis of the region in the supercontig in order to find missed alternative start codons that were not translated as methionine in the six-frame translation. A candidate precursor peptide was confirmed by (i) mass matching of putative core peptide sequence to the observed peptide mass by considering possible PTMs, (ii) sequence matching of the putative core peptide to the MS^n data and (iii) pBLAST analysis of the neighboring ORFs (gene cluster analysis). Based on the gene cluster components and the observed PTMs, an RNP class could usually be characterized (**Supplementary Table 1**). In cases of unusual gene cluster components during the RNP gene cluster analysis, a putative new RNP gene cluster could be defined by a search of homologous gene clusters (**Supplementary Figs. 11c,12c**). Finally, a structure of the RNP could be predicted based on the characterized core peptide sequence and PTMs that were characterized or predicted from the MS and bioinformatic analysis of the target peptide and its gene cluster.

Genome mining of nonribosomal peptides. A search tag that did not yield a candidate precursor peptide by six-frame translation-based genome mining was subjected to genome mining of NRP gene clusters. The mass shift sequence was reanalyzed by applying NRP monomer mass shifts (**Supplementary Table 7**) to characterize all possible NRP search tags. The supercontig of the target organism (for example, *S. hygroscopicus* ATCC 53653; **Supplementary Fig. 5**) was analyzed by NP.searcher[33] and by antiSMASH[34], and NRP search tags were compared to the predicted NRP sequences in monomers and in length. In case of a putative match, the corresponding NRP gene cluster was analyzed in its assembly-line organization in the corresponding antiSMASH output and by InterPro[50]. The accessibility of NRP families to genome mining by the NPP approach was assessed by an NP.searcher- and antiSMASH-based analysis of the GenBank files of characterized NRPS gene cluster families as described in the **Supplementary Methods**.

Additional methods. Bioinformatic prediction of PNP pathways, proteomic analysis of characterized RNPs and isolation and structure elucidation of Q027-1628 (stendomycin I) from the marine *Streptomyces* strain CNQ-027 are described in the **Supplementary Methods**.

Received 12 July 2011; accepted 27 August 2011; published online 9 October 2011

References

1. Daffre, S. *et al.* Bioactive natural peptides. in *Studies in Natural Products Chemistry*, 1st edn., Vol. 35 (ed. Rahman, A.U.) 597–691 (Elsevier, 2008).
2. Nolan, E.M. & Walsh, C.T. How nature morphs peptide scaffolds into antibiotics. *ChemBioChem* **10**, 34–53 (2009).
3. Donadio, S., Monciardini, P. & Sosio, M. Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat. Prod. Rep.* **24**, 1073–1109 (2007).
4. Velásquez, J.E. & van der Donk, W.A. Genome mining for ribosomally synthesized natural products. *Curr. Opin. Chem. Biol.* **15**, 11–21 (2011).
5. Ganz, T. Defensins and host defense. *Science* **286**, 420–421 (1999).
6. Moore, B.S. Extending the biosynthetic repertoire in ribosomal peptide assembly. *Angew. Chem. Int. Edn Engl.* **47**, 9386–9388 (2008).
7. Willey, J.M. & van der Donk, W.A. Lantibiotics: peptides of diverse structure and function. *Annu. Rev. Microbiol.* **61**, 477–501 (2007).
8. Li, C. & Kelly, W.L. Recent advances in thiopeptide antibiotic biosynthesis. *Nat. Prod. Rep.* **27**, 153–164 (2010).
9. Donia, M.S., Ravel, J. & Schmidt, E.W. A global assembly line for cyanobactins. *Nat. Chem. Biol.* **4**, 341–343 (2008).
10. Duquesne, S. *et al.* Two enzymes catalyze the maturation of a lasso peptide in *Escherichia coli*. *Chem. Biol.* **14**, 793–803 (2007).
11. Duquesne, S., Petit, V., Peduzzi, J. & Rebuffat, S. Structural and functional diversity of microcins, gene-encoded antibacterial peptides from enterobacteria. *J. Mol. Microbiol. Biotechnol.* **13**, 200–209 (2007).
12. Cotter, P.D., Hill, C. & Ross, R.P. Bacteriocins: developing innate immunity for food. *Nat. Rev. Microbiol.* **3**, 777–788 (2005).
13. Oman, T.J. & van der Donk, W.A. Follow the leader: the use of leader peptides to guide natural product biosynthesis. *Nat. Chem. Biol.* **6**, 9–18 (2010).
14. Challis, G.L., Ravel, J. & Townsend, C.A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).

15. Finn, R.D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
16. Winter, J.M., Behnken, S. & Hertweck, C. Genomics-inspired discovery of natural products. *Curr. Opin. Chem. Biol.* **15**, 22–31 (2011).
17. Lautru, S., Deeth, R.J., Bailey, L.M. & Challis, G.L. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* **1**, 265–269 (2005).
18. Claesen, J. & Bibb, M. Genome mining and genetic analysis of cypemycin biosynthesis reveal an unusual class of posttranslationally modified peptides. *Proc. Natl. Acad. Sci. USA* **107**, 16297–16302 (2010).
19. Li, B. *et al.* Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. USA* **107**, 10430–10435 (2010).
20. Kodani, S. *et al.* The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene *ramS* in *Streptomyces coelicolor*. *Proc. Natl. Acad. Sci. USA* **101**, 11448–11453 (2004).
21. Gressler, M., Zaehle, C., Scherlach, K., Hertweck, C. & Brock, M. Multifactorial induction of an orphan *PKS-NRPS* gene cluster in *Aspergillus terreus*. *Chem. Biol.* **18**, 198–209 (2011).
22. Ng, J. *et al.* Dereplication and *de novo* sequencing of nonribosomal peptides. *Nat. Methods* **6**, 596–599 (2009).
23. Tsur, D., Tanner, S., Zandi, E., Bafna, V. & Pevzner, P.A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23**, 1562–1567 (2005).
24. Duncan, M.W., Aebersold, R. & Caprioli, R.M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28**, 659–664 (2010).
25. Yang, Y.L., Xu, Y., Straight, P. & Dorrestein, P.C. Translating metabolic exchange with imaging mass spectrometry. *Nat. Chem. Biol.* **5**, 885–887 (2009).
26. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
27. Ueda, K. *et al.* AmfS, an extracellular peptidic morphogen in *Streptomyces griseus*. *J. Bacteriol.* **184**, 1488–1492 (2002).
28. McIntosh, J.A., Donia, M.S. & Schmidt, E.W. Ribosomal peptide natural products: bridging the ribosomal and nonribosomal worlds. *Nat. Prod. Rep.* **26**, 537–559 (2009).
29. Ohnishi, Y. *et al.* Genome sequence of the streptomycin-producing micro-organism *Streptomyces griseus* IFO 13350. *J. Bacteriol.* **190**, 4050–4060 (2008).
30. Willey, J.M., Willems, A., Kodani, S. & Nodwell, J.R. Morphogenetic surfactants and their role in the formation of aerial hyphae in *Streptomyces coelicolor*. *Mol. Microbiol.* **59**, 731–742 (2006).
31. Wilkinson, B. & Micklefield, J. Biosynthesis of nonribosomal peptide precursors. *Methods Enzymol.* **458**, 353–378 (2009).
32. Romano, A., Vitullo, D., Di Pietro, A., Lima, G. & Lanzotti, V. Antifungal lipopeptides from *Bacillus amyloliquefaciens* strain BO7. *J. Nat. Prod.* **74**, 145–151 (2011).
33. Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S. & Sherman, D.H. Automated genome mining for natural products. *BMC Bioinformatics* **10**, 185 (2009).
34. Medema, M.H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters. *Nucleic Acids Res.* **39**, W339–W346 (2011).
35. Bodanszky, M., Izdebski, J. & Muramatsu, I. Structure of the peptide antibiotic stendomycin. *J. Am. Chem. Soc.* **91**, 2351–2358 (1969).
36. Strieker, M. & Marahiel, M.A. The structural diversity of acidic lipopeptide antibiotics. *ChemBioChem* **10**, 607–616 (2009).
37. Roongsawang, N., Washio, K. & Morikawa, M. Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. *Int. J. Mol. Sci.* **12**, 141–172 (2010).
38. Visca, P., Imperi, F. & Lamont, I.L. Pyoverdine siderophores: from biogenesis to biosignificance. *Trends Microbiol.* **15**, 22–30 (2007).
39. Liu, W.T., Kersten, R.D., Yang, Y.L., Moore, B.S. & Dorrestein, P.C. Imaging mass spectrometry and genome mining via short sequence tagging identified the anti-infective agent arylomycin in *Streptomyces roseosporus*. *J. Am. Chem. Soc.* (in the press).
40. Caboche, S., Leclere, V., Pupin, M., Kucherov, G. & Jacques, P. Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J. Bacteriol.* **192**, 5143–5150 (2010).
41. Kawulka, K.E. *et al.* Structure of subtilosin A, a cyclic antimicrobial peptide from *Bacillus subtilis* with unusual sulfur to α-carbon cross-links: formation and reduction of α-thio-α-amino acid derivatives. *Biochemistry* **43**, 3385–3395 (2004).
42. Knappe, T.A., Linne, U., Xie, X. & Marahiel, M.A. The glucagon receptor antagonist BI-32169 constitutes a new class of lasso peptides. *FEBS Lett.* **584**, 785–789 (2010).
43. Bentley, S.D. *et al.* Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**, 141–147 (2002).
44. Nett, M., Ikeda, H. & Moore, B.S. Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Nat. Prod. Rep.* **26**, 1362–1384 (2009).
45. Vats, P. & Rothfield, L. Duplication and segregation of the actin (MreB) cytoskeleton during the prokaryotic cell cycle. *Proc. Natl. Acad. Sci. USA* **104**, 17795–17800 (2007).
46. Miao, V. *et al.* Daptomycin biosynthesis in *Streptomyces roseosporus*: cloning and analysis of the gene cluster and revision of peptide stereochemistry. *Microbiology* **151**, 1507–1523 (2005).
47. Koal, T. & Deigner, H.P. Challenges in mass spectrometry based targeted metabolomics. *Curr. Mol. Med.* **10**, 216–226 (2010).
48. Warren, A.S., Archuleta, J., Feng, W.C. & Setubal, J.C. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* **11**, 131 (2010).
49. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
50. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).

Acknowledgments

We thank N. Castellana and V. Bafna for providing the algorithm to enable the six-frame translations of supercontigs. We also thank M. Meehan for FTMS training. Financial support was provided by the US National Institutes of Health (GM085770 to B.S.M. and GM086283 to P.C.D.) and the Beckman Foundation.

Author contributions

R.D.K. designed and carried out experiments, analyzed data and wrote the paper. Y.-L.Y., Y.X. and S.-J.N. carried out experiments and analyzed data. P.C. and M.A.F. carried out the bioinformatic analysis and analyzed data. W.F. analyzed data. B.S.M. and P.C.D. designed experiments, analyzed data and wrote the paper.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available online at http://www.nature.com/naturechemicalbiology/. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to P.C.D. or B.S.M.